



COURSE MANUAL FISHERIES STATISTICS

R.S. BIRADAR



**CENTRAL INSTITUTE OF
FISHERIES EDUCATION**

(INDIAN COUNCIL OF AGRICULTURAL RESEARCH)

VERSOVA, BOMBAY - 400 061.



COURSE MANUAL FISHERIES STATISTICS

R.S. BIRADAR

y



**CENTRAL INSTITUTE OF
FISHERIES EDUCATION**

(INDIAN COUNCIL OF AGRICULTURAL RESEARCH)

VERSOVA, BOMBAY - 400 061.

CIFE-173



FOREWORD

PREFACE

1.	DEFINITION AND SCOPE OF FISHERIES STATISTICS	
1.1	Definition of statistics and fisheries statistics	.. 1
1.2	Scope of fisheries statistics	.. 1
1.3	The need for fisheries statistics	.. 3
1.4	Sources of fisheries data	.. 3
2.	SOME BASIC CONCEPTS AND COLLECTION OF DATA	
2.1	Population	.. 5
2.2	Sample	.. 5
2.3	Sampling frame	.. 6
2.4	Random sample	.. 6
2.5	Sampling with and without replacement	.. 7
2.6	Qualitative & quantitative characters	.. 8
2.7	Constant, variable and attribute	.. 8
2.8	Continuous and discrete variables	.. 8
2.9	Collection of data	.. 8
2.10	Advantages and disadvantages of census enumeration	..10
2.11	Advantages and disadvantages of sample surveys	..10
3.	PRESENTATION OF DATA	
3.1	Introduction	..12
3.2	Classification and tabular presentation of data	..12
3.3	Relative frequency distribution	..16
3.4	Diagrams and graphs	..16

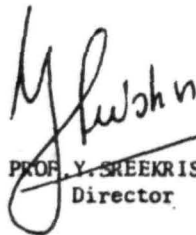
4	MEASURES OF CENTRAL TENDENCY, DISPERSION, SKEWNESS AND KURTOSIS	
4.1	Introduction	..26
4.2	Measures of central tendency	..26
4.3	Measures of dispersion	..36
4.4	Measures of skewness	..46
4.5	Measures of Kurtosis	..49
5.	ELEMENTARY PROBABILITY THEORY	
5.1	Introduction	..52
5.2	Terminology	..52
5.3	Definition of probability	..53
5.4	Mutually exclusive events	..58
5.5	Addition theorem	..58
5.6	Independent events	..59
5.7	Conditional probability	..60
5.8	Multiplication theorem	..60
6.	PROBABILITY DISTRIBUTIONS	
6.1	Introduction	..63
6.2	Random variable	..63
6.3	Probability distributions	..63
6.3.1	Binomial distribution	..64
6.3.2	Poisson distribution	..69
6.3.3	Normal distribution	..71
7.	SAMPLING DISTRIBUTIONS	
7.1	Sampling distribution	..80
7.2	Standard error	..81
7.3	Important uses of standard error	..85
7.4	Central limit theorem	..85

8.	ESTIMATION	
8.1	Introduction	.. 87
8.2	Types of estimators	.. 87
8.3	Properties of a good estimator	.. 87
8.4	Point estimation	.. 88
8.5	Interval estimation and confidence limits	.. 89
• 9.	TESTING OF HYPOTHESES	•
9.1	Introduction	.. 92
9.2	Terminology	.. 92
9.3	Tests of hypothesis for large samples	.. 95
9.4	Tests of hypothesis for small samples	..101
9.5	The Chi-square distribution	..110
9.6	The F distribution	..121
10.	CORRELATION AND REGRESSION	•
10.1	Introduction	..127
10.2	Scatter diagram	..127
10.3	Simple correlation	..129
10.4	Simple linear regression	..133
11.	SAMPLING METHODS	
11.1	Introduction	..151
11.2	Simple random sampling	..151
11.3	Stratified random sampling	..154
11.4	Systematic sampling	..159
11.5	Cluster sampling	..162
11.6	Subsampling or two stage sampling	..165
11.7	Sampling design to estimate total marine fish landings	..169
11.8	Estimation of inland fish catch	..170

12.	BASIC EXPERIMENTAL DESIGNS	
12.1	Introduction	..171
12.2	Terminology	..171
12.3	Basic principles of experimental designs	..172
12.4	Experimental designs	..173
	12.4.1 Completely randomized design	..173
	12.4.2 Randomised complete block design	..180
	12.4.3 Latin square design	..196
12.5	Advanced designs	..196
13.	TIME SERIES	
13.1	Introduction	..197
13.2	Components of time series	..197
13.3	Analysis of time series	..199
14.	INDEX NUMBERS	
14.1	Introduction	..211
14.2	Types of index numbers	..211
14.3	Base period and current period	..211
14.4	Price relatives	..212
14.5	Quantity relatives	..212
14.6	Necessity of single index number	..213
14.7	Construction of price index numbers	..214
14.8	Quantity index numbers	..218
14.9	Value index number	..223
14.10	Tests for consistency of an index number	..223
14.11	Cost of living index number	..224
14.12	Basic requirements in the construction of index numbers	..226
14.13	Uses of index numbers	..228
	REFERENCES AND READING LIST	..229

FOREWORD

Fisheries Statistics forms a part of the curricula of the different courses conducted by the Central Institute of Fisheries Education, Bombay. The students enrolling for these courses are generally graduates in biology/zoology/fisheries with mathematics upto matric level. Because of weak mathematical base they find it difficult to comprehend the statistics books which generally cater to a mixed target group. Moreover, the examples and applications contained in these books do not usually fulfil the needs of fisheries students. To overcome these lacunae, Prof. R.S.Biradar, Scientist of this Institute has written a manual on 'Fisheries Statistics' in simple & systematic manner to meet the requirements and expectations of the students of fishery science. I am sure the manual would also serve as an useful handbook to research workers.



PROF. Y. SREEKRISHNA
Director

PREFACE

This manual on 'Fisheries Statistics' is an outgrowth of my class room lectures delivered to the postgraduate students of fisheries sciences at CIFE, Bombay during the last five years. I may add that there is no intention or will to call this manual a text book. The primary objective of writing this manual is to provide a companion to the students of two year postgraduate Diploma in Fisheries Science and M.Sc.(Fisheries Management) courses of the Institute matching the syllabus relating to statistical techniques as applied in fisheries.

In this manual, the emphasis is on applications of statistical methods through examples drawn from the field of fisheries and as far as possible mathematical derivations have been avoided.

In the preparation of this manual many standard texts and journals have been consulted. However, in the bibliography selected references only have been included to enable the students to refer to them wherever detailed explanations are felt necessary.

It is a pleasure to mention that the idea of preparing this manual was mooted by Dr.S.N.Dwivedi, former Director of CIFE, Bombay, whose guidance and encouragement

made the manual possible. It was during his tenure, a preliminary version of the manual was brought out. I express my sincere gratitude to Prof.Y.Sreekrishna, Director,CIFE, Bombay for his encouragement and keen interest in bringing out this manual in a printed form. I am grateful to Dr.S.S.Pillai, Joint Director, IASRI and Dr.A.Dey, Senior Scientist, IASRI, New Delhi and Prof. K.K.Ghosh, Senior Scientist of the Institute for their critical scrutiny & suggesting improvements to the manual. I am thankful to Dr.M.Devaraj, Senior Scientist of the Institute for his valuable suggestions. I thank my colleague Mrs.R.Tewari, for her untiring efforts in bringing out this manual in the present form. Mrs.S.S.Gajbhiye and Mr.M.J.Shahakar have rendered valuable help in secretarial work and Mr.A.Sadanandan and Mr.D.R.Khogare in preparation of graphs, figures and art work for the cover page which I sincerely acknowledge.

While every effort has been made to make the manual useful and error free, any omissions or mistakes that may have crept into the manual if pointed out, would be welcome and appreciated.

R.S.BIRADAR

Chapter 1

DEFINITION AND SCOPE OF FISHERIES STATISTICS

Definition of statistics and fisheries statistics

The term 'statistics' has different meanings depending upon its usage as a plural or a singular noun. When used in the plural context it stands for numerical facts and figures. For instance consider the following statements :

The marine and inland fish production in the country during 1979 was 1.5 and 0.85 million tonnes respectively. Kerala topped the marine fish production with 3.30 lakh tonnes, while West Bengal topped the inland fish production with 2.32 lakh tonnes. These are 'Statistics' relating to fish production in the country. The use of the word statistics to signify numerical facts and figures is not very exact. The proper word to indicate numerical facts and figures is 'data'.

When the word statistics is used as a singular noun it stands for the 'Science of Statistics' or for 'Statistical methods'. There are many definitions for the science of statistics. One of these definitions is 'Statistics is the science of collection, presentation, analysis and interpretation of numerical facts'.

Biometry is the science of statistics as applied to quantitative study of the biological phenomena. Fisheries statistics relates to the branch of biometry applied to the study of fish and fisheries as well as to the study of socio-economic aspects of fisheries as a resource wealth utilised by man for avocation and food. It also encompasses fisheries data.

Scope of fisheries statistics or *Role of Statistics in fisheries or aquaculture.*

The scope of fisheries statistics can broadly be classified in to the following areas :

Inventory of potential resources

Estimation of total water area available for exploitation, area actually exploited, manpower employed in fishing and allied activities, the type of craft and gear employed for fishing etc.

1.2.2 Production

Estimation of inland and marine fish landings disaggregated according to mechanised and non-mechanised crafts, species, sizes etc. It also encompasses all forms of biotic productions from aquatic resources such as sea weeds, frogs etc.

1.2.3 Fish stock assessment

Growth of fish populations, their size (length or weight) and age structure, natality, recruitment and mortality, estimation of stock, optimum yield etc.

1.2.4 Morphometric and meristic analysis

Measurement of various body proportions such as total length, standard length, fork length, head length etc., for the purpose of statistical comparison with similar measurements for a sub species or a closely related species and establishing the levels of significance at which differences occur. In other words it serves the purpose of establishing the variations and relationships between different quantitative morphological characters of two or more closely related species.

Counts of spines and rays of fins, scales, vertebrae etc., constitute meristic characteristics of fish. These characteristics form one of the important taxonomic tools for differentiating closely related species.

1.2.5 Designing experiments for quantitative inferences

Designing field and laboratory experiments to quantify various biotic and abiotic phenomena in aquatic environment and interpret casual relationships in quantitative terms of these variables on fish behaviour, growth, production, survival, spawning, etc.

1.2.6 Quality control

Checking the quality of frozen fish and fish products using statistical quality control techniques.

1.2.7 Market research

Estimation of cost of production, price and price spread, estimation of supply and demand, consumption and distribution pattern, income, its distribution, investment and returns etc.

1.2.8 Genetic studies

Study of the various fish characters as regards to their heritable and non-heritable properties and patterns in different generations, efficiency of different selection procedures for improving fish stocks etc.

1.3 The need for fisheries statistics

1.3.1 Statistical methods are needed for collection, compilation and interpretation of data required for planning, development and management of the fishery sector.

1.3.2 Statistical methods are required to analyse and interpret the biological phenomena characterising fishery science.

1.3.3 Statistical methods are helpful in estimating the size of available fishery resources and the proper level at which to maintain stocks in order to obtain optimum yields.

1.4 Sources of fisheries data

1.4.1 Handbook on fisheries statistics

National fisheries data are periodically released by Government of India, Fisheries Division, Ministry of Agriculture and Cooperation as an official document for Central Board of Fisheries meeting. The latest is handbook on 'Fisheries statistics'. It contains information on production, exports, fishing harbours, training in fisheries, outlays and expenditure, prices, fishing resources and other aspects of fisheries.

1.4.2 Statistics of marine products exports

It is published annually by the Marine Products Export Development Authority, Cochin. It contains information on country-wise exports, region-wise exports, item-wise exports, average unit value, world markets, prices, marine fish landings etc.

**1.4.3 Marine fisheries information service
(Technical and Extension series)**

It is published by the Central Marine Fisheries Research Institute, Cochin. It contains information on marine and brackish water fishery resources and allied data.

1.4.4 Yearbook of fishery statistics

It is published annually by the Food and Agricultural Organisation (FAO) of the United Nations. It contains data on world catches, production of preserved and processed fishery commodities, estimated total international trade, imports, exports etc.

Chapter 2

SOME BASIC CONCEPTS AND COLLECTION OF DATA

2.1 Population or Universe

It is defined as the collection or an aggregate of all possible values (measurements or counts) of a particular characteristic for a specified group of individuals or the individuals themselves from which these values are obtained. For example,

- i) Population of fish weights of all fishes in a pond.
- ii) Population of income of fishermen families in a state.
- iii) Population of fish lengths in a sea.

A population can be finite or infinite. It is said to be finite if it contains finite number of individuals or units, Examples (i) and (ii) given above refer to finite populations.

A population of unlimited or very large measurable number of individuals is called infinite population. Example (iii) given above refers to infinite population.

The number of individuals or observations in a population is called population size and is usually denoted by 'N'.

2.2 Sample

A group of individuals or units that is chosen from a population is called a sample.

A group of (say 60) fishes selected from a pond to study their lengths is an example of a sample.

The number of individuals or observations in a sample is called the sample size and is generally denoted by 'n'.

2.3 Sampling frame

It is a list, map or other specification of units which constitute available information regarding the population. It forms the basis for drawing samples.

2.4 Random sample

The type of sample of importance and meaning in statistics is the 'random sample'. A random sampling is a method of sampling in which each individual in a population has a pre-assigned chance of being included in a sample. Generally, units are drawn one by one from the population. If the chance of selecting any unit at any draw is the same then the sampling is called simple random sampling. When the sample is so selected every possible sample has the same chance of being drawn. Simple random sample can be obtained either by using the 'lottery method' or by the 'use of random number tables'.

2.4.1 Lottery method

In this method, first number the individuals (units) of the population. Then write these numbers on identical chits and fold them so that the numbers are not visible. Then place these chits in a box. Shake the box thoroughly and draw chits one by one till the number of chits drawn equals the sample size. Note down the numbers of these chits. The individuals with these numbers form a sample.

2.4.2 Use of random numbers

Prepared tables of random numbers (Table XXX III, statistical Tables by Fisher and Yates, 1963) are available for drawing a simple random sample. These tables consist of series of digits from 1 to 9 which appear independent of each other and appear approximately equal number of times.

As a first step, units of the population are numbered from say 1 to N . From random number tables, select a number between 1 to N and include the unit bearing this number in the sample. Continue this procedure till the number of units included in the sample equals the sample size. In this procedure numbers larger than N are not considered. To avoid rejection of such numbers, modified procedures are adopted. One such procedure called 'remainder approach' is described below :

2.4.2.1 Remainder approach

If N is a 'd' digit number, determine first the highest 'd' digit multiple of N , let this be N' . Then a random number r is selected from 1 to N' . Divide this selected number r by N and find out the remainder. A unit with serial number equal to this remainder is selected. If the remainder is zero, the last unit (N) is selected. The procedure can best be illustrated with the following example :

If $N = 20$, the highest 2 digit multiple of 20 is 80. Then select a random number from 1 to 80. Let this number be 72. Division of this number by 20 gives a remainder of 12. Hence, the unit with serial number 12 is included in the sample. Select another number from 1 to 80 and repeat the procedure, till the number of units selected equals the sample size.

2.5 Sampling with and without replacement

There are two types of random sampling procedures viz : sampling 'with replacement' and sampling 'without replacement'. In sampling with replacement, the units drawn are replaced back before the next draw is made. If a table of random numbers is employed, the number drawn previously is considered in the subsequent draws also. In this procedure, the same unit can enter the sample more than once. In sampling without replacement, the unit once selected at any draw is not replaced back, so that the same unit cannot enter the sample more than once. If a table of random numbers is employed for drawing a random sample, the number that has been drawn previously, is ignored in the subsequent draws. If a sample of size n has to be drawn from a population of size N , then there will be N^n possible samples in the case of sampling

with replacement and N_C samples in the case of sampling without replacement.

2.6 Qualitative and Quantitative characters

The individuals or units of a population have some characteristics. These characteristics may or may not be numerically measurable. A characteristic which is numerically measurable is called 'quantitative' character, whereas, the characteristic which is not numerically measurable but is distinguished based on some quality or, attribute is known as 'qualitative' character. Length of fish, weight of fish, number of fish discarded are some of the examples of quantitative character. Sex, type of fishing boat, etc., are instances of qualitative character.

2.7 Constant, variable and attribute

If a characteristic remains the same for all individuals in a population it is called a constant.

A quantitative characteristic which varies from individual to individual is called a 'variable', where as, a qualitative characteristic which varies from individual to individual is called an 'attribute'.

2.8 Continuous and discrete variables

A variable which can take any value in a certain range including fractions is called 'continuous variable'. Length of fish, weight of fish etc., are examples of continuous variable.

A variable which takes only specific values in a certain range is called 'discrete variable'. Number of fish discarded, number of fin rays, number of vertebrae etc., are examples of discrete variable.

2.9 ✓ Collection of data

There are two sources of collecting data—primary source and secondary source. The data collected through direct personal investigation are said to have been collected from a primary source, and such data

are referred to as 'primary data'. The data collected from published or unpublished source are said to have been collected from a secondary source, and such data are called 'secondary data'. There are two methods of collecting data.

- i) Census or complete enumeration and
- ii) sample survey.

2.9.1 Census or complete enumeration method

In this method the required data are collected on all the individuals or units of the population. In order to study certain characteristics of a population, it is always advisable to measure all the individuals of the population, in which case, census enumeration is adopted. For example, in 1980 the Central Marine Fisheries Research Institute, Cochin, conducted an all India census of marine fishermen, craft and gear to bring out an inventory of fishing resources available in the country.

2.9.2 Sample survey method

In this method, the required data are collected on some individuals or units selected from the population. When the population is large, most of the investigations are carried out by this method and the results are generalised for the whole population.

Sample survey is the only logical alternative to census enumeration in the following situations :

- i) While dealing with infinite populations. For example, to study the characteristics of a fish population in a sea, it is not possible to collect and measure all the fish from sea and have to be satisfied with a sample.
- ii) If study requires destroying of units of individuals of the population.

For instance, in estimation of 'fecundity', finding 'biochemical composition' of fish etc., sampling is the only alternative.

2.10 Advantages and disadvantages of complete enumeration

2.10.1 Advantages

- i) It provides a statistical frame to other census and surveys.
- ii) Complete enumeration is better when information required is greater from smaller areas.
- iii) Census enumerations are quite often used as the basis for improving current statistics.

2.10.2 Disadvantages

In the case of infinite populations, census enumeration is not practically possible. Census enumeration is costly and requires more time and labour. Large scale census enumerations lead to non sampling errors, which are difficult to detect.

2.11 Advantages and disadvantages of sample surveys

2.11.1 Advantages

- i) **Reduced cost :** Sample surveys are less expensive and hence more studies can be carried out with fixed amount of resources like money and labour.
- ii) **Greater speed :** Sample surveys supply results quickly.
- iii) **Greater scope :** It is possible to have an intensive study in sample surveys. This is because a small sample may be thoroughly investigated whereas, for a large population, this may be impossible or too costly.
- iv) **Greater operational facility :** In sample surveys there is greater operational facility as compared to complete enumeration

- v) **The only alternative** : It obtains data that is not possible otherwise, for instance, in the case of infinite population and also where the study requires destruction of units of the population.
- vi) **Scientific** : It provides an estimate of sampling error which is useful in ascertaining the reliability of results.

2.11.2 Disadvantages

Breakdown of information for smaller areas or at sub-stratum level may not be possible in the case of sample surveys.

Chapter 3

PRESENTATION OF DATA

3.1 Introduction

Once data are collected, the first objective should be to summarise and present them in a form which highlights their main characteristics. (Tables, diagrams and graphs are important forms of presentation of statistical data.)

3.2 Classification and tabular presentation of data

(In tabular presentation, observations are classified systematically in to different groups or classes on the basis of some common characteristic and are arranged in rows and columns of a table. Characteristic used as a basis for classification can be qualitative or quantitative.) For example, in qualitative classification, fish may be classified as male or female. Fishing units may be classified as trawlers, gill netters, dol-netters etc. Fisherman population may be classified based on geographical considerations say according to the districts or states to which they belong. The quantitative classification is based on some measurable characteristic. For example, fish may be classified according to their length, weight etc. (An important form of quantitative classification is 'frequency distribution' table.)

3.2.1 Frequency distribution table

It consists of dividing the range of observations in to intervals (usually of equal size) and noting of the number of observations falling into each interval (called frequency) and presenting it in a tabular form. The data presented in the form of a frequency distribution table is called 'grouped data'.

3.2.2 Terminology

The terms which are commonly used in formation of a frequency distribution table are explained below :

i) **Class interval and class limits**

A symbol defining a class 5-9 is called a class interval. The end numbers 5 and 9 are called class limits, the smaller number 5 is called the lower class limit and the larger number 9 is called the upper class limit. The terms class and class interval are often used interchangeably.

ii) **Inclusive and exclusive class intervals**

The class interval for instance 5-9 will be called inclusive class interval if it includes all the values from 5 to 9 including the lower limit 5 and the upper limit 9. On the other hand, it will be called exclusive class interval if it does not include the upper limit 9. The given data can be classified either using inclusive or exclusive class intervals.

iii) **Class boundaries or true class limits**

If the lengths of fish are recorded to the nearest-centimetre, the class interval 5-9 theoretically includes all measurements from 4.5 to 9.5 cm. These numbers indicated by exact numbers 4.5 and 9.5 are called 'class boundaries' or 'true class limits'. The smaller number 4.5 is the lower class boundary and the larger number 9.5 is the upper class boundary. In the case of inclusive class intervals when the interval is in integers, class boundaries are obtained by subtracting 0.5 from the lower class limit and adding 0.5 to the upper class limit. In general, the class boundaries are obtained by adding the upper limit of one class interval to the lower limit of the next higher class interval and dividing by 2. In the case of exclusive class intervals class limits coincide with class boundaries.

iv) **The size or width of class interval**

It is the difference between the lower and upper class boundaries. In the case of equal class intervals it is the difference between the two successive lower class limits or two successive upper class limits.

v) **Class mark or class mid point**

The class mark is the mid point of the class interval and is obtained by adding the lower and upper class limits and dividing by 2. Thus the class mark of the interval 5-9 is $\frac{5+9}{2} = 7$. The class mark is also called class mid point.

General rules to be followed while forming a frequency distribution table

1. The number of class intervals should ordinarily be between 5 and 20 though there is no hard and fast rule in this respect*. If it exceeds 20 the computation becomes tedious and if it is less than 5, a great amount of accuracy is lost.
2. Class intervals should be clearly defined.
3. Open end class intervals should not be used.
4. As far as possible class intervals of equal size (width) should be used for the sake of ease of computations.
5. It is convenient to make the mid point of a class interval a whole number, i.e. an integer.

Steps in the formation of frequency distribution table.

The following steps have to be followed while forming a frequency distribution table :

*Sturges empirical formula for determining number of classes is

$$k = 1 + 3.322 \log N$$

Where k is the number of classes and N is the total number of observations. This formula takes in to account only the number of observations and not the spread of the variable under study. Hence this formula is not very satisfactory.

1. Determine the largest and the smallest numbers in the given data and then find the range by subtracting the smallest number from the largest number.
2. Divide the range into a convenient number of class intervals having the same size.
3. Determine the number of observations falling into each class interval, i.e., find the frequency. This is done by using tally marks.

Example 1

The lengths (in cm.) of 30 randomly selected fish are given below :

14, 25, 17, 20, 35, 38, 40, 25, 32, 31
 21, 27, 16, 11, 19, 22, 49, 30, 24, 29
 27, 34, 37, 35, 40, 26, 26, 34, 19, 20

Arrange the observations into a frequency distribution table.

Answer

Here the smallest observation is 11 and the largest is 49. Hence, the range is 38. If we take class intervals of width 5, there will be 8 class intervals which is in accordance with the rules to be followed in the formation of frequency tables. We have to take class intervals in such a way that the first class interval includes the smallest observation and the last class interval includes the largest observation.

Class interval (length in cm)	Tally marks	Frequency (number of fish)
10-14		2
15-19		4
20-24		4
25-29		8

Table contd.

Class interval (length in cm)	Tally marks	Frequency (number of fish)
30-34		5
35-39		4
40-44		2
45-49		1
TOTAL		30

3.3 Relative frequency distribution

The relative frequency distribution is formed by dividing the frequencies in each class of a frequency distribution table by the total number of observations. If the relative frequencies of each class are expressed in percentages by multiplying by 100, the resulting distribution is called percentage frequency distribution. The relative frequency distribution or percentage frequency distribution is especially useful in comparing two or more sets of data having different number of observations (total frequency).

3.4 Diagrams and graphs

In addition to frequency distribution tables, diagrams and graphs are also commonly used in the presentation of data. Well designed diagrams and graphs make the unweildy data readily intelligible and bring to light the outstanding features of data at a glance. Diagrams and graphs are easily understood by a common man and the impression created by them is long lasting. They also make comparison of trends, values and relationships very easy. Graphs are also useful in locating some statistical measures like median, quartiles and other partition values. In spite of these advantages, it must be noted that diagrams and graphs are merely visual aids and can only supplement and not replace tables or the original data. Most commonly used diagrams and graphs are discussed below :



3.A.1 Simple bar diagram

It consists of lines or bars of equal width with length proportional to the value of the variable or character under study. Bars may be drawn either horizontally or vertically. These bars are simple to draw and are very effective for comparing the magnitude of different values.

Example 2

The following data represent the total fish production in the country from 1975 to 1979. Represent the data by bar diagram.

Year	1975	1976	1977	1978	1979
Fish production (lakh tonnes)	22.66	21.74	23.12	23.06	23.43

Answer

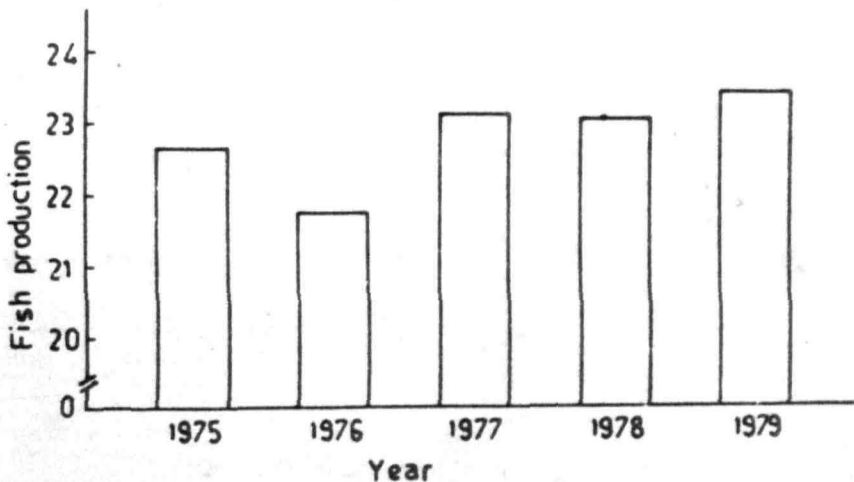


Fig. 1. Bar diagram.

3.4.2 Component or subdivided bar diagram

If the variable or character under study has two or more component parts, then the component bar diagram is used. As in the case of simple bar diagram, the bars of equal width with length proportional to the total value of the variable are drawn with a convenient scale. Then these bars are subdivided into component parts and are marked in different colours to distinguish the components from one another. This diagram is more useful when one wants to compare the size and also the relation between each component and the total.

3.4.3 Multiple bar diagram

When there are two or more different comparable sets of closely related variables or characters, multiple bar diagram is used. The diagram consists of multiple bars drawn contiguous to one another representing the component parts and then marked in different colours to distinguish the components.

Example 3

The data on marine and inland fish catch (in lakh tons) from 1976 to 1979 are given below. Draw (i) component bar diagram and (ii) multiple bar diagram.

Year	1976	1977	1978	1979
Marine	13.75	14.48	14.90	14.95
Inland	7.99	8.64	8.16	8.48

Answer

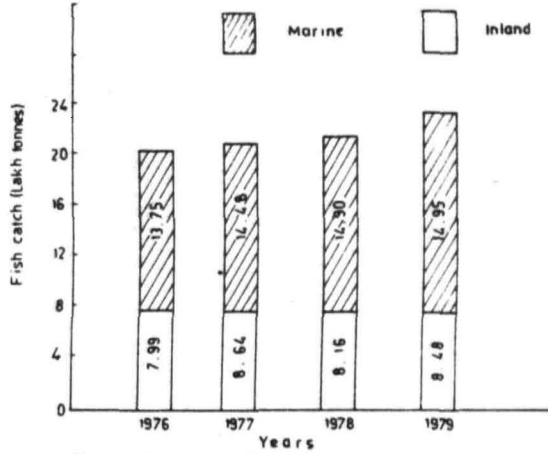


Fig. 2a. Component bar diagram

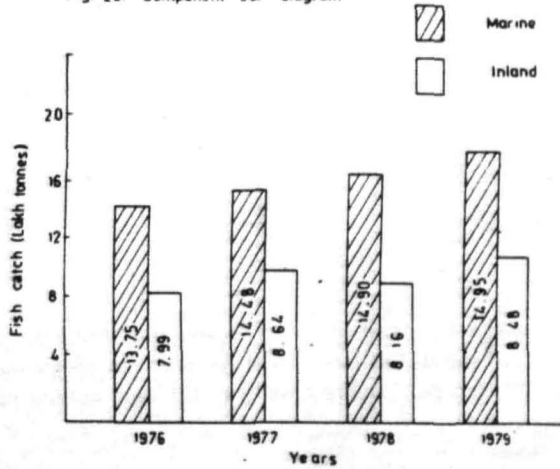


Fig. 2b. Multiple bar diagram

3.4.4 Pie diagram (Pie chart)

This diagram is used to portray relations among component parts of the total. It is drawn by dividing a circle into different sectors with areas proportional to the magnitude (frequency) of the respective components. Since the complete angle at the centre of the circle is 360° , the angle of degrees required for different components is worked out by the following formula :

$$\frac{360 \times \text{magnitude of the component}}{\text{Total}}$$

The circle is divided into sectors based on angles of degrees of respective components, using the protractor.

Example 4

The following data show the areas in million square miles of the oceans of the world. Represent the data by Pie diagram.

Ocean	Pacific	Atlantic	Indian	Antarctic	Arctic	Total
Area (Million sq.miles)	70.8	41.2	28.5	7.6	4.8	152.9

Answer

To construct a Pie diagram we use the fact that the total 152.9 million sq.miles corresponds to the total number of degrees, i.e., 360° . Thus the angle of degrees required for the different oceans are,

	Ocean			Angles
1)	Pacific	:	$\frac{360}{152.9} \times 70.8$	= 166.7
2)	Atlantic	:	$\frac{360}{152.9} \times 41.2$	= 97.0

	Ocean					Angles
3)	Indian	:	$\frac{360}{152.9}$	x	28.5	= = 67.1
4)	Antarctic	:	$\frac{360}{152.9}$	x	7.6	= 17.9
5)	Arctic	:	$\frac{360}{152.9}$	x	4.8	= 11.3
						<hr/>
						360.0
						<hr style="border-top: 1px dashed black;"/>

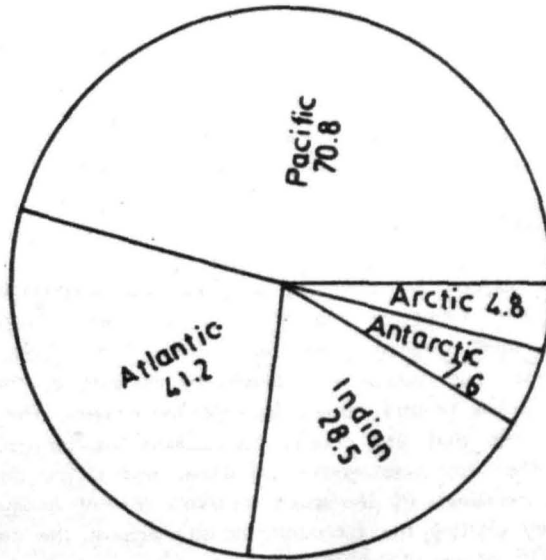


Fig. 3. Pie diagram.

The following types of graphs are used in representing frequency distributions.

3.4.5 Histogram

The histogram consists of vertical rectangles adjacent to each other. The area of each rectangle represents the frequencies of the corresponding class interval. In drawing the histogram first mark off along the x - axis all the class intervals and then taking these class intervals as the bases draw rectangles with their heights proportional to the frequency of the respective class interval. The resulting figure is called a "histogram" relating to the data of class intervals of equal width.

If the class intervals are of unequal width, the rectangles are drawn with heights proportional to the frequency density (ratio of class frequency to class width) so that the area is proportional to the class frequency.

If the frequency distribution has inclusive class intervals, first convert them into exclusive type and then draw the histogram.

3.4.6 Frequency polygon

Frequency polygon is drawn by plotting class frequencies against mid values of the respective class intervals and then joining these plotted points by small straight lines. Some times the polygon is left open at each end but usually it is closed by drawing a straight line from each end down to the horizontal axis (X - axis). The points on the horizontal axis that are chosen for closing the polygon are the mid points of the first class interval on either end of the distribution which has zero frequency. If the class intervals are of unequal width, it is obtained by plotting the frequency density against the class mid values.

Alternatively, frequency polygon can be obtained by joining the mid points of the upper sides of the adjacent rectangles in the histogram, by small straight lines.

3.4.7 Frequency curve

Frequency curve is drawn by plotting class frequency against the mid values of the respective class intervals and then joining these points by a smooth curve. When the class intervals are of unequal width, it is obtained by plotting the frequency density against the class mid values. Frequency polygon approaches the frequency curve when the number of observation (total frequency) is large and smaller class intervals are used.

Alternatively, frequency curve can also be formed by drawing a smooth curve through the mid points of the upper sides of the adjacent rectangles of the histogram.

Frequency curves are widely used for comparison purposes, for analysing different statistical theories etc.

Frequency polygon as well as frequency curves can also be constructed using percentages of frequencies of each class to the total frequency instead of the frequencies in each class. These percentages of frequencies are plotted against the mid points of the class intervals taking the former on the Y - axis and the latter on the X - axis. This type of presentation is more useful in comparing two or more sets of data having different total frequencies.

Example 5

Draw (i) Histogram, (ii) Frequency polygon and (iii) Frequency curve for the following data.

	Length of fish (mm)				
	5-15	15-25	25-35	35-45	45-55
Numbers	9	21	40	22	8

Answer

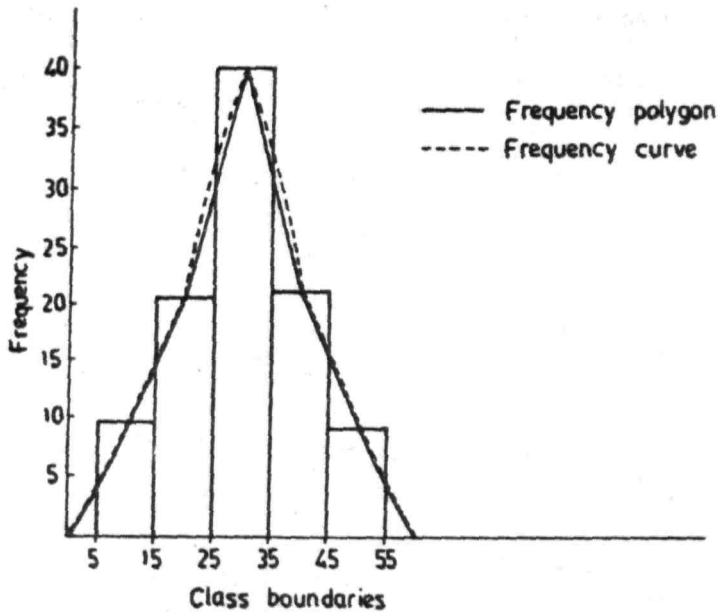


Fig. 4. Histogram, Frequency polygon and Frequency curve.

3.4.8 Ogive curves

The graph of cumulative frequency distribution is known as ogive. Class boundaries are taken on X - axis. At each class boundary, the corresponding cumulative frequency is marked and these points are joined by a smooth curve. There are two types ogive curves for each frequency distribution : (a) less than type ogive and (b) more than type ogive. The 'less than' type is obtained by plotting the less than cumulative frequencies against the upper class boundaries while the other type is obtained by plotting the more than cumulative frequencies against the lower class boundaries. The curve helps to find out how many items lie below a certain value of the variable or above it. Median, quartile and other partition values of data can also be determined using ogive curve.

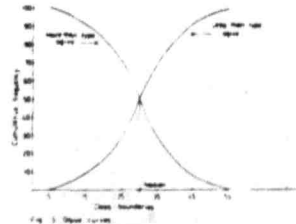
Example 6

Draw ogives of the less than type and the more than type for the data on fish length given in example 5 above.

Answer

First prepare a cumulative frequency table of less than type and more than type. For the given data the following cumulative frequency table is obtained :

Class interval (length cm)	Frequency	Cumulative frequency	
		less than	more than
5-15	9	9	100
15-25	21	30	91
25-35	40	70	70
35-45	22	92	30
45-55	8	100	8
Total	100		



Now to get an ogive of less than type, plot the less than cumulative frequencies against the upper class boundaries, taking the class boundaries on X - axis and the cumulative frequencies on the Y - axis and join these points by a smooth curve. To get an ogive of more than type, plot the more than cumulative frequencies against the lower class boundaries, taking the class boundaries on the X - axis and the cumulative frequencies on the Y - axis and join these points by a smooth curve. From the point of intersection of the two ogives, a perpendicular is drawn to the X - axis. The point where the perpendicular meets the X - axis is the median.

Chapter 4

MEASURES OF CENTRAL TENDENCY, DISPERSION,
SKEWNESS AND KURTOSIS

4.1 Introduction

Tabular presentation of data is useful in condensing large number of observations in to a few classes or groups. Diagramatic and graphical representation facilitate comparison of trends and relationships. However, more exact description of important characteristics of a data set is provided by single numbers called 'measures of data' or 'summary measures'. These measures describe data set in a simple and concise manner and enable us to gain more precise understanding of data. There are 4 such measures which describe the important characteristics of a data set. They are,

- (a) Measures of central tendency
 - (b) Measures of dispersion
 - (c) Measures of skewness
 - and (d) Measures of kurtosis
- } fixed for 'N' distribution

For a majority of biological characteristics the frequency distributions approximate to a symmetrical bell-shaped curve known as the 'normal curve'. For such frequency distributions only the first two measures, ~~namely~~ the measures of central tendency and dispersion are important, the third and fourth being fixed.

4.2 Measures of central tendency

In biological characteristics, generally, the observations taken on a group of individuals will not be equal, but show a general tendency to cluster around a particular value. This value around which the observations tend to cluster is called the 'central tendency' or 'central position' of that group.

There are 3 commonly used measures of central tendency. They are :

- (1) the arithmetic mean
- (2) the median
- and (3) the mode

4.2.1 The arithmetic mean or average

(a) Calculation from ungrouped data

In an ungrouped data the arithmetic mean is simply the total of all the values divided by the number of observations. For example, if 5 observations of fish length (in cm) are 22, 38, 29, 33, 28, then the arithmetic mean will be,

$$\frac{22 + 38 + 29 + 33 + 28}{5} = \frac{150}{5} = 30 \text{ cm}$$

Thus if there are n observations, x_1, x_2, \dots, x_n , then the arithmetic mean which is usually denoted by \bar{x} will be

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + \dots + x_n}{n} \\ &= \frac{\sum x}{n} \end{aligned}$$

where \sum stands for the sum of all observations and is read as summation.

(b) Calculation from grouped data

(i) Direct method

Suppose that the data are given in the form of a frequency distribution in k classes. Let x_1, x_2, \dots, x_k represent the mid points of 1st, 2nd \dots k th class and f_1, f_2, \dots, f_k their frequencies respectively. Then the arithmetic mean is calculated by the formula.

$$\begin{aligned} \bar{x} &= \frac{f_1 x_1 + f_2 x_2 + \dots + f_k x_k}{f_1 + f_2 + \dots + f_k} \\ &= \frac{\sum fx}{f} \\ &= \frac{\sum fx}{n} \end{aligned}$$

where $n = \sum f$

(ii) **Short method for the computation of arithmetic mean**

The process of computation of the arithmetic mean from frequency table can be shortened by changing the mid points of the class intervals to a coded value, when the class intervals are of equal size. This is done by assigning the value zero to any mid point (class mark) preferably one corresponding to the middle of the frequency distribution. The mid points of successive class intervals above and below this mid point are coded as minus or plus deviations, d of 1, 2, 3 or more intervals. The mean is calculated by summing the products obtained by multiplying the frequency in each class by the corresponding d value of the class interval and dividing this sum by n , the number of observations. Since this average is in coded values, it must be converted to the original units by multiplying by the width of the class interval ' h ' and added to the mid point of the interval that was selected as zero. If A is used to represent the mid point to which the value of zero was assigned, the formula for the arithmetic mean, calculated by this method is,

$$\begin{aligned} \bar{x} &= A + h\bar{d} \\ &= A + h \frac{\sum fd}{n} \end{aligned}$$

Example 1

Compute the arithmetic mean for the following data on fish length by (i) Direct method and (ii) Short method.

Class interval (length in cm)	5-15	15-25	25-35	35-45	45-55
Frequency (No. of fish)	9	21	40	22	8

Answer(i) **Direct Method**

Class interval	Frequency (f)	Mid point (x)	fx
5-15	9	10	90
15-25	21	20	420
25-35	40	30	1200
35-45	22	40	880
45-55	8	50	400
	$\Sigma f = 100 = n$		$2990 = \Sigma fx$

$$\begin{aligned}
 \text{Arithmetic mean, } \bar{x} &= \frac{\Sigma fx}{n} \\
 &= \frac{2990}{100} \\
 &= 29.90
 \end{aligned}$$

Short method

Class interval	Mid point (x)	d	Frequency f	fd
5-15	10	-2	9	-18
15-25	20	-1	21	-21
25-35	A 30	0	40	0
35-45	40	1	22	22
45-55	50	2	8	16
<hr/>			<hr/>	<hr/>
Total			100	-1

$$\begin{aligned}
 \text{Arithmetic mean, } \bar{x} &= A + h \frac{\sum fd}{n} \\
 &= 30 + 10 \frac{(-1)}{100} \\
 &= 30 - 0.10 \\
 &= 29.90
 \end{aligned}$$

The median

The median is defined as the middle value when the values (observations) are arranged in the ascending or descending order of magnitude. Median divides the data into two equal halves, half the number of observations lying below it and half above it.

(a) Calculation from ungrouped data

To find out the median from ungrouped data it is first necessary to arrange the values in the ascending or descending order of magnitude, with serial numbers 1, 2 n, where n stands for the total number of observations in the given data. If n is an odd number, then the median is the value corresponding to the serial numbers $\frac{n+1}{2}$. If n is an even number then median is taken as the arithmetic mean of the two middle values, i.e., the arithmetic mean of the values corresponding the serial number $\frac{n}{2}$ and $\frac{n}{2} + 1$.

Example 2

Find the median length for the following data on fish length (cm) :
22, 12, 16, 32, 18, 26, 20

Answer

Arrange the observations, say, in the ascending order.

Sl. No.	1	2	3	4	5	6	7
Observation	12	16	18	20	22	26	32

Here the number of observations, $n = 7$

Middle observation is $\frac{n+1}{2} = \frac{8}{2} = 4$ th item

Hence, the median is the value corresponding to serial No. 4. The value corresponding to serial number 4 is 20. Hence, the median is 20.

Example 3

Find the median for the following data
22, 12, 16, 32, 18, 16, 20, 9

Answer

Arrange the observation in, say, the ascending order.

Sl. No.	1	2	3	4	5	6	7	8
Observation	9	12	16	16	18	18	22	32

Here the number of observations n is 8

$$\frac{n}{2} = 4 \text{ and } \frac{n}{2} + 1 = 5$$

Therefore, the median is the arithmetic mean of the 4th and the 5th observations.

$$\text{i.e., } \frac{16 + 18}{2} = 17$$

(b) Calculation from grouped data

In the case of grouped data the formula for computing the median is given by,

$$\text{Median} = L_1 + \frac{\left(\frac{n}{2} - f^c\right)h}{f}$$

where L_1 = lower class boundary of the median class (median class is the class whose cumulative frequency equals or exceeds $\frac{n}{2}$ items)

n = Total frequency

f = Cumulative frequency before entering the median class

f^c = The frequency of the median class

h = Width of the class interval

Example 4

The size frequency of a fish sample is given below. Find the median.

Size group (mm)	200-220	220-240	240-260	260-280
Frequency	10	40	120	150
	280-300	300-320	320-340	
	40	20	20	

Answer

First find out the cumulative frequency and then locate the median class.

Size class (mm)	Frequency	Cumulative (frequency)
200-220	10	10
220-240	40	50
240-260	120	170
260-280	150	320

Table contd.

280-300	40	360
300-320	20	380
320-340	20	400
	<u>100</u>	

As $\frac{n}{2} = 200$ lies between the cumulative frequencies of 240-260 and 260-280 class intervals, the median class is 260-280.

$$\begin{aligned}
 L_1 &= 260, h = 20, f = 150, f_c = 170 \\
 \text{Median} &= L_1 + \frac{\left(\frac{n}{2} - f_c\right) h}{f} \\
 &= 260 + \frac{(200 - 170) 20}{150} \\
 &= 260 + \frac{30 \times 20}{150} \\
 &= 264
 \end{aligned}$$

4.2.3 The mode

The mode is defined as the most frequently occurring value. The mode may not exist and even if it exists it may not be unique as there may be more than one mode. A distribution having only one mode is called unimodal, whereas, the distribution having two modes is called bimodal and the distribution having more than two modes is called multimodal. We will be concerned mostly with unimodal distributions.

(a) Calculation from ungrouped data

When the data are not grouped the mode can be located by arranging the observations either in the ascending or the descending order. This arrangement helps to find out the value which repeats large number of times.

Example 5

Find the modal value for the following data on fish length :

12, 22, 17, 9, 22, 28, 17, 30, 20, 22

Answer

Arranging the data in, say, ascending order the following array is obtained:

9, 12, 17, 17, 20, 22, 22, 22, 28, 30

Here 22 occurs 3 times and the remaining observations occur less than 3 times. Hence, 22 is the mode.

(b) Calculation from grouped data

The class interval which has the maximum frequency is known as the modal class. The modal value lies in this class interval. The following formula is used for calculating the mode of the frequency distribution.

$$\text{Mode} = L_1 + \frac{(f_m - f_1)}{2f_m - f_1 - f_2} h$$

where, L_1 = Lower class boundary of the modal class

f_1 = Frequency of the class previous to the modal class

f_2 = Frequency of the class next to the modal class

f_m = Frequency of the modal class

h = Width of the class interval

If the mean and the median have already been calculated then the following empirical relationship can be used to calculate the mode of moderately asymmetrical distribution.

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

Example 6

Find the mode for the data on length of fish given in example 4.

Answer

Size group (mm)	Frequency
200-220	10
220-240	40
240-260	120
260-280	150 ✓
280-300	40
300-320	20
320-340	20
	400

As the class 260-280 contains the maximum frequency, it is the modal class

$$L_1 = 260, h = 20, f_m = 150, f_1 = 120, f_2 = 40$$

$$\begin{aligned}
 \text{Mode} &= L_1 + \frac{(f_m - f_1) h}{2f_m - f_1 - f_2} \\
 &= 260 + \frac{(150 - 120)}{(2 \times 150 - 120 - 40)} \times 20 \\
 &= 260 + \frac{30}{140} \times 20 \\
 &= 260 + 4.28 \\
 &= \underline{264.28}
 \end{aligned}$$

4.2.4 Some characteristics and uses of different measures of central tendency

4.2.4.1 The arithmetic mean

The arithmetic mean is the most important and commonly used measure of central tendency for statistical work in biometry. It is a complete and adequate measure, as it takes into account both the total number of observations and their size. The arithmetic mean is more useful when observations of the data are distributed symmetrically and when other statistics are to be computed later. Its greatest weakness is that it is affected by extreme values.

4.2.4.2 The Median

The median is not affected by extreme values and hence more representative than the arithmetic mean in extremely skewed distributions. It takes into account only the number of observations and not their size. Hence, the median is less reliable than the arithmetic mean and so it is less commonly used. However, occasions often arise where median is the most appropriate. For instance, in the study of fishermen income, investment in fisheries etc.

4.2.4.3 The mode

The mode is particularly useful in the study of typical size. For instance, boat yard is not interested in finding out the mean size or the median size of boat. Rather, it wants to find out the boat size most in demand so that it may build a larger quantity of that size. It is not affected by extreme values. In some distributions there may be two or more modes of (equal concentration) which make the mode largely useless in such situations. It is not amenable for further statistical analysis. It deliberately excludes arithmetical precision as its aim is to present a really typical measure. It is, therefore not based on all observations of the data. The arithmetic mean is designed to be numerically accurate, and may have to sacrifice its typical feature some times for numerical accuracy. Hence, it is very often necessary to calculate both the mean and the mode. Mode is particularly useful in age and growth studies of fishes. Mode of length frequency data plotted over successive time intervals (e.g. month) aid in the determination of age and growth of fishes.

4.3 Measures of dispersion or variation

A series of observations can be described by a measure of central tendency. Usually all observations will not be equal to the central tendency value but they vary. Measure of the degree to which the observations vary about the central tendency value is called the 'measure of dispersion'. None of the measures of central tendency indicate how the observations are scattered around the measure. Two sets of data may have the same mean but the observations in one may scatter widely around the mean and can be highly congregative in another. For example, consider two sets of data on length of fish in centimetres.

Set 1 : 15 15 14 15 16 15
 Set 2 : 22 14 12 17 10 15

which have the same mean value 15, although the pattern of individual observations is different in both the cases. In set 1, the observations congregate around the mean, whereas, in the second none of the observations have the value of the mean and are highly scattered. Thus in order to get the true picture of the data a measure of the central tendency has to be complemented by a suitable measure of dispersion. Three important measures of dispersion viz. range, mean deviation and standard deviation, are discussed here.

4.3.1 Range ✓

The range is defined as the difference between the highest and lowest values in a given data.

Example 7

Compute the range for the following data on the length of fish (In cm):

16, 12, 26, 17, 22, 25, 18, 14, 19, 24

Answer

In the given data, the highest value is 26, and the lowest is 12. Therefore,

$$\text{Range} = 26 - 12 = 14$$

In a grouped data, the range is taken as the difference between the class mark (mid point) of the highest class and the lowest class.

Example 8

Calculate the range for the following data.

Class interval	Frequency
5-9	2
10-14	4
15-19	7
20-24	5
25-29	2

Answer

In the given data, the highest class is 25-29 and its class mark will be $\frac{25+29}{2} = 27$. The lowest class is 5-9 and its class mark is $\frac{5+9}{2} = 7$.

Range = Class mark of the highest class - Class mark of the lowest class
 = $27 - 7 = 20$

4.3.2 The mean deviation (average deviation)

The arithmetic mean of the difference of each individual observation from the mean ignoring sign is called the mean deviation.

(a) Calculation from ungrouped data

Let x_1, x_2, \dots, x_n denote the values of the characteristic under study. Then mean deviation is given by

$$M. D. = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|}{n}$$

$$M.D. = \frac{\sum |x - \bar{x}|}{n}$$

Example 9

Trawl catches (kg) of 10 trips at a landing centre are given below :

35, 38, 42, 40, 52, 38, 45, 32, 35, 43

Compute the mean deviation.



Answer

First compute the arithmetic mean, \bar{x} and then subtract \bar{x} from each observation. Add these differences irrespective of the sign and then divide by the number of observations to get the mean deviation.

$$\text{Arithmetic mean} = \frac{400}{10} = 40$$

S.No.	1	2	3	4	5	6	7	8	9	10	Total
x	35	38	42	40	52	38	45	32	35	43	
x-40	5	2	2	0	12	2	5	8	5	3	44

$$\text{Mean deviation} = \frac{\sum |x - \bar{x}|}{n} = \frac{44}{10} = 4.4$$

(b) **Calculation from grouped data**

Suppose that the data are given in the form of a frequency distribution in k classes. Let x_1, x_2, \dots, x_k represent mid points of 1st, 2nd, \dots , k th class intervals and f_1, f_2, \dots, f_k their respective frequencies. Then the mean deviation (M. D.) is given by

$$\begin{aligned} \text{M. D.} &= \frac{|x_1 - \bar{x}| f_1 + |x_2 - \bar{x}| f_2 + \dots + |x_k - \bar{x}| f_k}{n} \\ &= \frac{\sum |x - \bar{x}| f}{n} \end{aligned}$$

where \bar{x} is the arithmetic mean

Example 10

Compute the mean deviation for the following data on net profit from adoption of composite fish culture in 100 farms.

Net profit - Rs. in 1000/ha

	2-4	4-6	6-8	8-10	10-12	12-14
No. of farms	5	10	40	25	15	5
Answer						
Class interval	x	f	fx	x-8	f x-8	
2-4	3	5	15	5	25	
4-6	5	10	50	3	30	
6-8	7	40	280	1	40	
8-10	9	25	225	1	25	
10-12	11	15	165	3	45	
12-14	13	5	65	5	25	
Total		100	800		190	

$$\bar{x} = \frac{800}{100} = 8$$

$$\text{M. D.} = \frac{190}{100} = 1.9$$

Note : The mean deviation was computed taking deviations from arithmetic mean. It can also be worked out taking deviations from the median. It is to be noted that the mean deviation will be least when it is taken from the median.

4.3.3 The variance and standard deviation

The variance of a population is calculated as the average of squares of deviations of observations from the arithmetic mean. While dealing with the variance of a sample, the sum of squares of deviations of observations from the arithmetic mean is usually divided by one less than the total number of observations. Variance of a population is usually denoted by σ^2 , where as variance of a sample is denoted by S^2 . Formula for computing σ^2 and S^2 are given below :

$$\sigma^2 = \frac{\sum (x - m)^2}{N}$$

$$= \frac{1}{N} \left[\sum x^2 - \frac{(\sum x)^2}{N} \right] \text{ for ungrouped data}$$

In the above expression m is the population mean given by,

$$m = \frac{\sum x}{N}$$

$$\sigma^2 = \frac{\sum f(x-m)^2}{N} \text{ for grouped data}$$

$$= \frac{1}{N} \left(\sum fx^2 - \frac{(\sum fx)^2}{N} \right)$$

If the estimate of population variance σ^2 is obtained from sample measurements using \bar{x} for m , the average of the squared deviations for a sample of size n tends to under estimate σ^2 . However, the quantity,

$$S^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$$

has the average value σ^2 in repeated sampling. Hence, the formula to compute sample variance for grouped and ungrouped data are :

$$\begin{aligned} \text{Ungrouped : } S^2 &= \frac{1}{n-1} \sum (x - \bar{x})^2 \\ &= \frac{1}{n-1} \left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \end{aligned}$$

$$\begin{aligned} \text{Grouped : } S^2 &= \frac{1}{n-1} \sum f (x - \bar{x})^2 \\ &= \frac{1}{n-1} \left[\sum fx^2 - \frac{(\sum fx)^2}{n} \right] \end{aligned}$$

The positive square root of variance is called the standard deviation. Population standard deviation is usually denoted by σ where as, sample standard deviation is denoted by S .

(a) Calculation from ungrouped data

Example 11

Calculate the variance and standard deviation for the following sample data on length (cm) of fingerlings: 9, 5, 8, 9, 7, 4, 10, 8



D7064

Answer

Serial No.	1	2	3	4	5	6	7	8	Total
x	9	5	8	9	7	4	10	8	60
x ²	81	25	16	81	49	16	100	16	480

Number of observations, $n = 8$

$$\begin{aligned} \text{Variance } S^2 &= \frac{1}{n-1} \left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \\ &= \frac{1}{7} \left[480 - \frac{(60)^2}{8} \right] \\ &= \frac{30}{7} = 4.2854 \end{aligned}$$

Standard deviation, $S = \sqrt{4.2854} = 2.07$ **(b) Calculation from grouped data**

Suppose that the data are given in the form of a frequency distribution in k classes. Let x_1, x_2, \dots, x_k represent mid points (class marks) of 1st, 2nd, ..., k th and f_1, f_2, \dots, f_k the respective frequencies. Let 'h' denote the width of the class interval and $n = \sum f_i$, the total number of observations in the given data.

The process of computation of standard deviation from the frequency distribution table, when class intervals are of equal size, may be shortened by changing the mid points of the class intervals to a coded value. This is done by assigning the value zero to any mid point preferably to the one corresponding to the middle of the frequency distribution table. The mid points of successive intervals above and below these mid points are coded as minus or plus deviations, d , of 1, 2, 3 or more intervals. Then the formula for working out the variance for a given data based on coded values is given by

$$\text{Variance} = \frac{1}{n-1} \left[\sum fd^2 - \frac{(\sum fd)^2}{n} \right] \times h^2$$

$$\text{Standard deviation} = +\sqrt{\text{Variance}}$$

The above method of computing variance and standard deviation is usually referred to as 'short method'.

Example 12

Calculate the variance and the standard deviation for the following data on fish length using (i) direct method and (ii) Short method.

Class interval (Length in cm)	5-15	15-25	25-35	35-45	45-55
Frequencies (no. of fish)	9	21	40	22	8

Answer

(i) Direct method

Class interval	Frequency f	Mid point x	x^2	fx	fx^2
5-15	9	10	100	90	900
15-25	21	20	400	420	8400
25-35	40	30	900	1200	36000
35-45	22	40	1600	880	35200
45-55	8	50	2500	400	20000
Total	$\sum f = 100$			$\sum fx = 2990$	$\sum fx^2 = 100500$

$$\sum f = 100, \quad \sum fx = 2990, \quad \sum fx^2 = 100500$$

$$\text{Variance} = \frac{1}{n-1} \left[\sum fx^2 - \frac{(\sum fx)^2}{n} \right]$$

$$= \frac{1}{99} (100500 - 89401)$$

$$= \frac{1}{99} (11099)$$

$$= 112.111$$

$$\begin{aligned} \text{Standard deviation} &= \sqrt{\text{variance}} \\ &= \sqrt{112.111} \\ &= 10.588 \end{aligned}$$

(ii) Short method

Class interval	Mid point	d	d^2	f	fd	fd^2
5-15	10	-2	4	9	-18	36
15-25	20	-1	1	21	-21	21
25-35	30	0	0	40	0	0
35-45	40	1	1	22	22	22
45-55	50	2	4	8	16	32
Total				100	-1	111

$$h = 10, n = 100 \quad \sum fd = -1 \quad \sum fd^2 = 111$$

$$\text{Variance} = \frac{1}{n-1} \left[\sum fd^2 - \frac{(\sum fd)^2}{n} \right] \times h^2$$

$$= \frac{1}{99} \left[111 - \frac{(-1)^2}{100} \right] \times (10)^2$$

$$= \frac{1}{99} (111 - 0.01) \times 100$$

$$= \frac{1}{99} (110.99) \times 100 = 112.111$$

$$\begin{aligned} \text{Standard deviation} &= \sqrt{\text{Variance}} = \sqrt{112.111} \\ &= 10.588 \end{aligned}$$

4.3.4

Comparison of different measures of dispersion.

Of the 3 measures of dispersion discussed so far, range is the easiest to compute and provides some indication of the amount of variability

present in the data. However, it is not a satisfactory measure as it is based on only two out of the whole bulk of observations. For this reason it does not adequately reflect the information regarding the variability present in data, unless we are dealing with a small set of data.

Mean deviation is easy to understand and simple to compute. It takes into account deviations of all values from a measure of central tendency (say mean or median) and hence superior to range as a measure of dispersion. However, it has the disadvantage of being mathematically unsound as algebraic signs are totally disregarded in its computation. This mathematical flaw is one of the main reasons for the mean deviation not being used commonly.

The standard deviation resembles the mean deviation in that it is also based on the deviation of every value from the arithmetic mean. It has a further advantage of being algebraically sound and hence can be used satisfactorily in further statistical analysis. Hence, standard deviation is the most commonly used measure of dispersion.

4.3.5 Relative measure of dispersion - Coefficient of variation

Coefficient of variation :

All the above measures of dispersion have a unit attached to them. For example, the standard deviation calculated from a data recorded in centimeters will have units as centimeters and so on. Therefore, the variability of fish length recorded in centimeters and fish weight in grams cannot be compared using standard deviation, as they are in different units. In such situations a relative measure of dispersion called the coefficient of variation which is independent of the units of measurements is used. It is calculated as the ratio of standard deviation to arithmetic mean and is expressed in percentage, i.e.,

$$\text{Coefficient of variation} = \frac{\text{Standard deviation}}{\text{arithmetic mean}} \times 100$$

Example 13

If the arithmetic mean and standard deviation of length of fish are 32 and 8 respectively, find the coefficient of variation.

Answer

$$\begin{aligned}\text{Coefficient of variation} &= \frac{\text{Standard deviation}}{\text{arithmetic mean}} \times 100 \\ &= \frac{8}{32} \times 100 = 25\%\end{aligned}$$

Example 14

Two characters, length and weight were recorded on a random sample of 100 fishes of a particular species. It was found that the mean length was 50 centimeters with standard deviation of 20 centimeters, whereas, the mean weight was 200 grams with standard deviation of 65 grams. Find out which of the two characters is more variable.

Answer

$$\begin{aligned}\text{Coefficient of variation for length} &= \frac{20}{50} \times 100 \\ &= 40\%\end{aligned}$$

$$\begin{aligned}\text{Coefficient of variation for weight} &= \frac{65}{200} \times 100 \\ &= 32.5\%\end{aligned}$$

Since the coefficient of variation of length is more than the coefficient of variation of weight, the length is more variable than weight.

4.4

Measures of Skewness

Even if two frequency distributions have similar means and variances, their frequency curves can not be expected to be similar. One may be vastly different than the other. A frequency distribution is said to be symmetrical when the observations equidistant from the central maximum & have the same frequencies. The frequency curve of such distribution will be bellshaped as shown in Fig. 1.

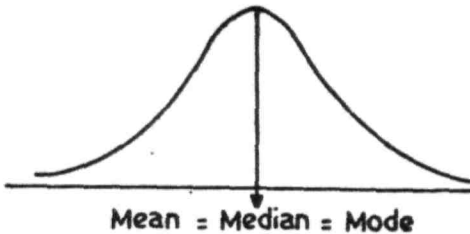


Fig. 1 Frequency curve of symmetrical distribution.

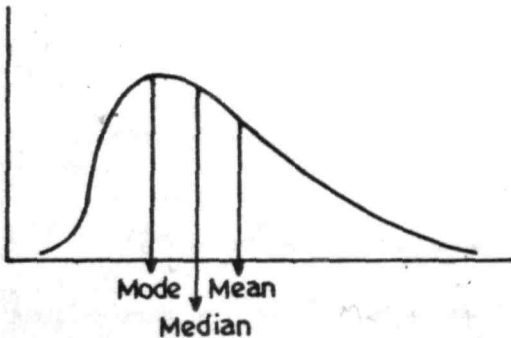


Fig. 2 Positively Skewed curve

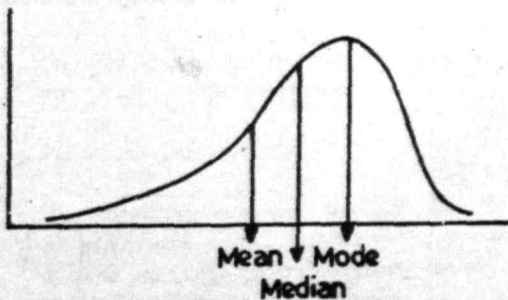


Fig. 3 Negatively skewed curve

An important example of the symmetrical distribution is the normal distribution, whose frequency curve is bell-shaped. For symmetrical distribution mean, median and mode coincide.

Skewness means departure from symmetry. If the frequency curve of distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or to have positive skewness (Fig. 2). In such frequency distributions, mean is greater than the mode.

If the curve has a longer tail to the left of the central maximum than to the right then the distribution is said to be skewed to the left or to have negative skewness (Fig. 3). In such distributions mean is less than the mode.

The degree of skewness can be measured by the Karl Pearson's coefficient of skewness which has the following formula :

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} \dots \dots (1)$$

In some frequency distributions it may be difficult to calculate mode accurately. In such cases the coefficient of skewness which is based on median can be used. It is given by,

$$\text{Coefficient of skewness} = \frac{3 (\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

This coefficient is zero for symmetrical distributions and is positive for positively skewed distribution and negative for negatively skewed distributions.

An important and commonly used measure of skewness, which is based on the (third moment) about the mean (see annexure 1 for moments) is given by,

$$\beta_1 = \frac{m_3^2}{m_2^3}$$

Some times the measure of skewness is given by $\gamma_1 = \sqrt{\beta_1}$

In the above formula m_2 and m_3 are 2nd and 3rd moments about the mean. The values of β_1 and γ_1 are zero for a symmetrical distribution. These measures of skewness are free from units of measurements and are therefore useful in comparing skewness of distributions recorded in different units.

4.5 Measures of kurtosis

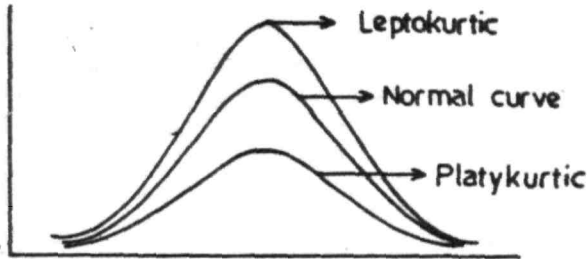


Fig. 4. Kurtosis

of relatively higher peakedness than the normal curve is called leptokurtic, whereas the curve having a flat top is called platykurtic (Fig. 4). The normal curve which is neither very peaked nor flat-topped is called mesokurtic.

A measure of kurtosis which is based on the 4th moment about the mean is given by,

$$\beta_2 = \frac{m_4}{m_2^2}$$

For normal distribution, $\beta_2 = 3$. Because of this reason sometimes measure of kurtosis is given by $\gamma_2 = \beta_2 - 3$. In the above formula m_4 is the 4th moment about the mean. For normal distribution γ_2 is equal to zero. It will have positive value for leptokurtic distribution and negative value for platykurtic distribution. The degree of kurtosis is indicated by the magnitude of this coefficient. /

4.5 Parameter and statistic

Any measure computed from all values of units in the population such as population mean, variance, etc., is called population 'parameter', whereas, the measure computed from values of the units in the sample such as sample mean, sample variance, etc., is called sample 'statistic'.

Frequency curve of a distribution may be symmetrical it can still be different in structure from the normal curve. For instance it may show more peakedness or less peakedness than the normal curve. Kurtosis is the degree of peakedness of a frequency curve as compared to the normal curve. A curve

In biological investigations, it is not practically possible to collect data on the whole population. Hence, the values of parameters are seldom known. They are estimated by statistics.

ANNEXURE 1

Moments :

The r th moment about the mean (m_r) is defined as the mean of the r^{th} power of the deviations of the values from the arithmetic mean.

$$\text{Thus, } m_r = \frac{\sum (x_i - \bar{x})^r}{n} \dots (I) \quad \text{for ungrouped data}$$

$$m_r = \frac{\sum f (x_i - \bar{x})^r}{n} \dots (II) \quad \text{for grouped data}$$

When $r = 1$, the formula (I) gives the first moment,

$$m_1 = \frac{\sum (x_i - \bar{x})}{n} = 0$$

Thus the first moment about the arithmetic mean is zero. When $r = 2$, the formula (I), gives $m_2 = \frac{\sum (x_i - \bar{x})^2}{n}$ which is the variance of population.

When $r = 3$, the formula (I) gives the 3rd moment m_3 and soon. The moments about the mean are usually called the central moments.

Chapter 5

ELEMENTARY PROBABILITY THEORY

5.1 Introduction

The basic foundation of statistics is the probability theory which aims to systematise the laws of chance to discover the regularities in the pattern in which the events depending on chance repeat themselves. Probability had its beginning with the games of chance such as the tossing of a coin, throwing of a dice drawing a card, etc., in the 17th century. It was only in the 19th century Gregor Mendel while studying the genetic laws in peas showed that it can be applied to biological investigations also. Since then, it is being applied very successfully to various problems in biology.

5.2 Terminology

Some terms which are frequently used in probability theory are explained below :

(i) **Deterministic experiment**

If the same results are obtained when an experiment is repeated under the same conditions, such an experiment is called deterministic experiment. For example, for a perfect gas, $PV = \text{constant}$, provided temperature is constant. The same result will be obtained whenever the experiment is repeated. Thus the results of a deterministic experiment can be predicted with certainty.

(ii) **Random experiment**

The experiment which do not yield the same result when repeated under the same conditions is called random experiment. In such an experiment it is not possible to predict the result in advance with certainty. For example in tossing of a coin experiment one toss may yield head and other toss may yield tail. It is not possible in advance to predict the outcome with certainty.

(i) (4)
= Head

(iii) **Simple event or event**

Every distinct outcome of a random experiment is called simple event or an outcome. For example, in tossing of a coin experiment, head is one outcome and tail is another outcome. Hence, head and tail are the two events in tossing of a coin experiment. In throwing of a dice experiment, getting number 1 on top is one event, similarly getting 2, 3, 4, 5, 6 are other events.

(iv) **Sample space**

The totality or collection of all possible outcomes of a random experiment is called sample space. It is denoted by S . For example in tossing of a coin experiment sample space consists of two events Head (H) and Tail (T). It is usually written as

$$S = (H, T)$$

(v) **Compound event**

A compound event is the one which consists of two or more simple events. For example, getting an even number in tossing of a dice experiment is a compound event.

(vi) **Equally likely events or outcomes (equal chances)**

The outcomes are said to be equally likely when there is no reason to expect any one rather than the other. For example, in tossing of a coin experiment, either head or tail may appear, so that both the outcomes are equally likely.

5.3 Definition of probability

5.3.1 Definition I : Classical or mathematical or a priori definition

Suppose an event E can happen in ' m ' different ways (outcomes) out of a total of ' n ' different equally likely ways (outcomes), then the

probability of occurrence of an event denoted by p or $P(E)$ is given by

$$p = P(E) = \frac{\text{No. of favourable ways to } E}{\text{Total No. of equally likely ways}}$$

$$= \frac{m}{n}$$

Note 1 : Probability of an event is a non-negative number which lies between 0 and 1. Symbolically $0 \leq p \leq 1$.

Note 2 : If the event E can happen in 'm' ways out of total of 'n' ways, then the number of ways in which the event E will not happen is $n - m$. Hence the probability that an event E will not happen (denoted by q) is given by,

$$q = \frac{n - m}{n} = 1 - \frac{m}{n} = 1 - p, \text{ where } p = m/n$$

So that, $p + q = 1$

i.e. the sum of the probabilities of occurrence and non-occurrence of an event is equal to 1.

Example 1

What is the probability of getting head when an unbiased coin is tossed?

Answer

Total number of equally likely outcomes = 2

No. of favourable outcome = 1

Therefore, the probability of getting head denoted by $P(H)$ is, $P(H) = 1/2$

Example 2

What is the probability of getting an even number when an unbiased dice is thrown ?

Answer

Total Number of equally likely outcomes = 6
 Number of favourable outcomes = 3
 $P(\text{Even number}) = \frac{3}{6} = \frac{1}{2}$

Example 3

In a pond containing 100 fishes, 20 are marked. If one fish is subsequently caught what is the probability of it being (i) marked (ii) unmarked?

Answer

(i) Total number of fishes = 100
 Number of marked fishes = 20

Hence, the number of favourable chances for marked fish are 20.

$$\therefore P(\text{marked fish being caught}) = \frac{20}{100} = 0.2$$

(ii) $P(\text{marked fish being caught})$
 + $P(\text{unmarked fish being caught}) = 1$
 $P(\text{unmarked fish being caught}) = 1 - P(\text{marked fish being caught})$
 $= 1 - 0.2$
 $= 0.8$

Example 4

In a composite fish culture experiment, fingerlings of 6 species of fish namely, rohu, catla, mrigal, common carp, silver carp and grass carp, were stocked in the ratio of 1:1:2.5:3:1.5 respectively. A fingerling is subsequently drawn, what is the probability that it is of catla?

Answer

Fingerlings of rohu, catla, mrigal, common carp, silver carp and grass carp are stocked in the ratio of 1:1:2.5:3:1.5 respectively. Thus out of 10 fingerlings we have 1 fingerling of catla. Hence, the probability that the fingerling drawn is of catla, $= \frac{1}{10} = 0.10$

5.3.2 Definition II : Relative frequency or a posteriori or empirical definition of probability

Classical definition of probability defined earlier assumes that outcomes are equally likely, the total number of outcomes are known and finite. When these assumptions are not met with, it is not possible to compute the probability of an event using the classical definition. In order to overcome the above limitations, a new approach called relative frequency concept of probability is adopted. According to this concept the probability of occurrence of an event E is the limiting value of ratio of frequency of occurrence of the event to the total number of outcomes. For instance if an experiment was repeated n times under the same conditions and if an event E has occurred f times, then the estimate of the probability of an event E as the number of trials n increase indefinitely is given by,

$$P(E) = \lim_{n \rightarrow \infty} \frac{f}{n}$$

It is to be noted that as the number of trials (frequency) increases the estimate of probability of an event stabilizes around a particular value.

Example 5

The frequency distribution of lengths in 1000 randomly selected fishes of a particular species are given below. What is the probability that a fish chosen at random will have length between 35-45 cm?

Length :	5-15	15-25	25-35	35-45	45-55	
Numbers :	88	210	400	220	82	= 1000
(Frequency)						

Answer

Frequency of the class interval 35-45 is 220. Therefore, the relative frequency of this class to the total frequency is $\frac{220}{1000} = 0.22$.

Hence, the probability that the fish chosen at random will have length between 35-45 cm is 0.22.

Example 6

One thousand fertilized eggs of a major carp were kept under observation to find out the number of individuals reaching different stages in the life history. Observed data are given below :

State :	Fertilized egg	Hatchling	Fry	Fingerling	Adult
Number :	1000	700	210	200	196

Find the probability that,

- (i) Fertilized egg reaches fingerling stage
- (ii) hatchling reaches fry stage
- (iii) fry reaches adult stage

Answer

- (i) Out of 1000 fertilized eggs, only 200 reached the fingerlings stage. Therefore, the probability of fertilized egg reaching the fingerling stage is $\frac{200}{1000} = 0.2$
- (ii) Out of 700 hatchlings only 210 reached the fry stage, therefore, the probability of hatchling reaching the fry stage is $\frac{210}{700} = 0.30$
- (iii) Out of 210 fry 196 reached the adult stage, therefore, the probability of the fry reaching the adult stage is $\frac{196}{210} = 0.928$

5.4

Mutually exclusive events

Two events A and B are said to be mutually exclusive if the occurrence of one event precludes the occurrence of another, i.e., both the events cannot happen simultaneously. In other words A and B have no common outcomes. In the tossing of a coin experiment, head and tail are mutually exclusive events, as they cannot happen simultaneously. Similarly in an experiment on rolling a dice events of getting number 1 on top and the events of getting number 2 on top simultaneously are mutually exclusive as numbers 1 and 2 cannot appear on the top of the dice simultaneously.

5.5

Addition theorem

Let A and B be two events with respective probabilities P (A) and P (B). The probability of occurrence of at least one of these two events denoted by P (A+B) is given by $P (A+B) = P (A) + P (B) - P (AB)$ where P (AB) is the probability of simultaneous occurrence of A and B.

Corollary

If events A and B are mutually exclusive, then $P (A+B) = P (A) + P (B)$

It is because $P (AB) = 0$, when A and B are mutually exclusive.

Example 7

In a certain district 25% of the fish farmers practice composite fish culture of rohu, catla and mrigal, 15% fish farmers follow monoculture of rohu only and 10% farmers follow composite fish culture as well as monoculture of rohu in their farm. Find the probability that a randomly selected fish farmer follows at least one of the practices.

Answer

Let events A and B be,

A : The farmer follows composite fish culture

B : The farmer follows monoculture of rohu

Then, $P (A) = 0.25$ $P (B) = 0.15$ $P (AB) = 0.10$

The probability that the farmer follows atleast one of the practices is denoted by $P(A+B)$ and is given by $P(A+B) = P(A) + P(B) - P(AB)$

$$\begin{aligned} &= 0.25 + 0.15 - 0.10 \\ &= 0.30 \end{aligned}$$

Example 8

A pond contains 150 fishes of rohu, 225 fishes of catla and 125 fishes of mrigal. Find the probability that a fish randomly selected is rohu or a catla.

Answer

Let events A and B be,

A : Selected fish is rohu

B : Selected fish is catla

Events A and B are mutually exclusive as a fish selected cannot be both rohu and catla. Hence,

$$P(A+B) = P(A) + P(B)$$

$$\text{We have } P(A) = \frac{150}{500} = \frac{3}{10} = 0.30$$

$$P(B) = \frac{225}{500} = 0.45$$

$$\begin{aligned} \therefore P(A+B) &= 0.30 + 0.45 \\ &= 0.75 \end{aligned}$$

Independent events

The events A and B are said to be independent if the occurrence of one does not depend on the occurrence or non-occurrence of the other. For instance when a coin is tossed two times, the result of the second throw would in no way be affected by the result of the first throw.

5.7 Conditional probability

Let A and B be any two events. The conditional probability of event A, given that event B has happened, is denoted by $P(A/B)$. Similarly, the conditional probability of B, given that event A has happened is given by $P(B/A)$.

5.8 Multiplication theorem

Let A and B be two events with probability $P(A)$ and $P(B)$ respectively. Let $P(B/A)$ denote the conditional probability of event B, given that event A has happened, $P(A/B)$ the conditional probability of event A, given that event B has happened. Then the probability of occurrence of both the events A and B denoted by $P(AB)$ is given by,

$$\begin{aligned} P(AB) &= P(A) \cdot P(B/A) \\ &= P(B) \cdot P(A/B) \end{aligned}$$

If events A and B are independent, then,

$$\begin{aligned} P(AB) &= P(A) \cdot P(B) \\ &= P(B) \cdot P(A) \end{aligned}$$

Example 9

In a pond containing 100 fishes, 25 are marked. If two are caught one after another and without replacement, what is the probability that both the fishes caught are marked?

Answer

Let A denote the event of catching marked fish in the first attempt and B denote the event of catching marked fish in the 2nd attempt.

$$\text{then } P(A) = \frac{25}{100} = 0.25$$

The probability of drawing marked fish in the 2nd catch, given that the first fish caught was marked is,

$$P(B/A) = \frac{24}{99}$$

Hence, P (both the fish caught are marked) = $P(AB)$

$$\begin{aligned} &= P(A) \cdot P(B/A) \\ &= \frac{25}{100} \cdot \frac{24}{99} \\ &= \frac{6}{99} \\ &= 0.06 \end{aligned}$$

Example 10

A pond contains 200 fishes of which 40 are marked. A second pond contains 300 fishes of which 50 are marked. One fish is drawn from each of the ponds. What is the probability that the fishes drawn are both marked?

Answer

Let A denote the event of catching marked fish from 1st pond, B denote the event of catching marked fish from 2nd pond.

$$\text{Hence, } P(A) = \frac{40}{200} = \frac{1}{5}; P(B) = \frac{50}{300} = \frac{1}{6}$$

As the events A and B are independent,

$$\begin{aligned} P(AB) &= P(A) \cdot P(B) = \frac{1}{5} \cdot \frac{1}{6} \\ &= \frac{1}{30} = 0.033 \end{aligned}$$

Example 11

An urn contains 5 white and 7 black pomfrets. A second urn contains 7 white and 8 black pomfrets. One pomfret is taken out at random and put into the second urn without noticing its colour. A fish is then drawn at random from the second urn. What is the probability that it is a white pomfret?

Answer

Two cases arise here

Case (i) Pomfret taken from 1st urn is white

Let A denote the event of drawing white pomfret from 1st urn and let B denote the event of drawing white pomfret from 2nd urn.

$$P(A) = \frac{5}{12}, P(B/A) = \frac{8}{16}$$

$$\text{Hence, } P(AB) = P(A) \cdot P(B/A) = \frac{5}{12} \cdot \frac{8}{16} = 0.208$$

Case (ii) Pomfret taken out from 1st urn is black

Let A denote drawing black pomfret from 1st urn and let B denote drawing white pomfret from 2nd urn

$$P(A) = \frac{7}{12}, P(B/A) = \frac{7}{16}$$

$$\begin{aligned} \text{Hence, } P(AB) &= P(A) \cdot P(B/A) = \frac{7}{12} \cdot \frac{7}{16} \\ &= 0.255 \end{aligned}$$

$$\begin{aligned} \text{Therefore, required probability} &= P(\text{Case i}) + P(\text{Case ii}) \\ &= 0.208 + 0.255 \\ &= 0.463 \end{aligned}$$

Chapter 6

PROBABILITY DISTRIBUTIONS

6.1 Introduction

Probability distribution is analogous to a relative frequency distribution with probabilities replacing relative frequencies. Thus, probability distributions can be regarded as theoretical or limiting forms of relative frequency distributions, when the number of observations made is very large. Hence, probability distributions can be considered as distributions of populations, whereas relative frequency distributions are distributions of samples drawn from these populations. Frequency distributions which arise in practice can be approximated by well known theoretical probability distributions which serve as useful tools in making inferences and decisions under conditions of uncertainty on the basis of limited data or theoretical considerations.

6.2 Random variable

It is a numerically valued function defined over a simple space.

6.3 Probability distribution

There are two types of probability distributions, discrete and continuous. Probability distribution is said to be discrete when it is based on a discrete random variable and continuous when it is based on a continuous random variable.

A probability distribution for discrete random variable is a listing of all possible values with respective probabilities of occurrence. As discrete random variable can take only a finite number of values or a countable infinite number of values, it is possible to list all the values with the corresponding probabilities. In case of continuous random variable, it is no longer meaningful and hence the probability of a random variable falling in a given interval is listed.

Example 1

Find the probability distribution of an outcome in throwing of a dice experiment.

Answer

Let X denote the outcome of the experiment. Then the probability distribution of X is given by,

X	1	2	3	4	5	6
$P(X)$	1/6	1/6	1/6	1/6	1/6	1/6

6.3.1 Binomial distribution or Bernulli distribution

Binomial distribution is a discrete distribution. It has great practical applications in research and industrial inspection problems. It arises whenever there is dichotomous classification, i.e. when an event (character) can occur in one of the two possible ways. For example, male or female, with scales or scaleless, dead or alive, tall or dwarf, it responds or does not respond to a given stimuli and so on. The mathematical description of the Binomial distribution is as follows: Suppose that the individuals examined possess certain character with probability p and, does not possess it with probability $1-p = q$. Then, the probability of X individuals out of a sample of n possessing the character is given by,

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x} \dots \dots (1)$$

$$\text{for } x = 0, 1, 2, \dots, n.$$

The equation (1) is referred to as the Binomial distribution. If p and n are known, this distribution can be completely determined. Hence, p and n are called parameters of the Binomial distribution. In this distribution p is assumed to be constant from observation to observation and outcomes of observations are independent.

6.3.1.1 Important properties of the binomial distribution

- (i) Mean of the binomial distribution = np
- (ii) Variance of the binomial distribution = npq
Standard deviation = \sqrt{npq}
- (iii) For $p = q = 1/2$ it is symmetrical;

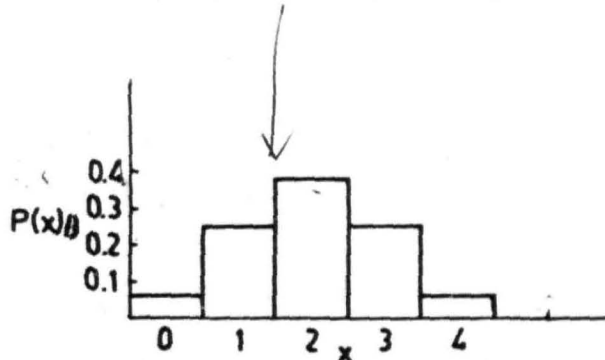


Fig. 1. Symmetric distribution

For $p < 1/2$ it is positively skewed,
i.e. $\mu < \text{mean}$ then 0.5.

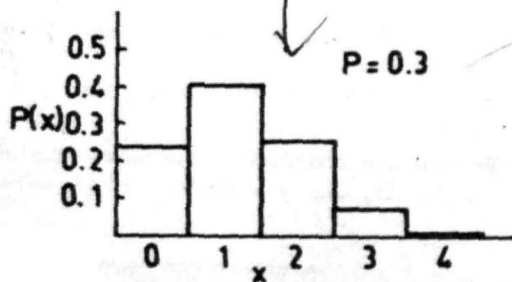


Fig. 2. Positively skewed distribution

For $p > 1/2$ it is negatively skewed.

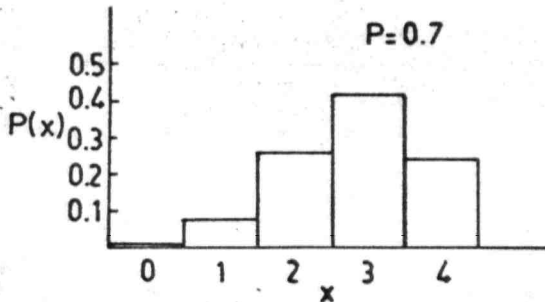


Fig. 3. Negatively skewed distribution.

- (iv) As n increases the binomial distribution approaches (tends towards) the normal distribution (to be discussed in 6.2.3)

Example 2

Find the probability of finding only 1 catla in a sample of 10 fishes drawn one by one, if the probability of a catla being drawn in any draw is 0.2.

Answer

If we have a sample of size n , the probability of getting x catla is given by $P(x) = {}^n C_x p^x q^{n-x}$ for $x = 0, 1, 2, \dots, 10$

In this example p is the probability of a catla being drawn in any draw which is given to be 0.2, i.e., $p = 0.2$

$$\therefore q = 1 - p = 1 - 0.2 = 0.8$$

sample size $n = 10$, $x = 1$, i.e. getting one catla

\therefore Probability of getting one catla is

$$\begin{aligned} P(1) &= {}^{10}C_1 (0.2)^1 (0.8)^{10-1} \\ &= 10 (0.2) (0.8)^9 \\ &= 2(0.8)^9 = 2 \times 0.134 = 0.268 \end{aligned}$$

Example 3

What is the probability of finding 2 males in a sample of 5 fishes drawn one by one? (Assume probability of finding male fish = 0.5).

Answer

The probability of finding male = $1/2 = p$ (say). The probability of not finding male (i.e. finding female) = $1 - 1/2 = 1/2 = q$ (say).

Applying binomial distribution,

$$P(x) = nC_x p^x q^{n-x}$$

with $n = 5$, $x = 2$, $p = 1/2$, $q = 1/2$

we have

$$\begin{aligned} P(2) &= {}^5C_2 (1/2)^2 (1/2)^3 \\ &= \frac{5!}{2!3!} (1/2)^5 \\ &= \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} (1/2)^5 \\ &= \frac{5}{16} \\ &= 0.312 \end{aligned}$$

6.3.1.2 Fitting of Binomial distribution

Binomial distribution is fitted by estimating p from the observed data. As the mean of the Binomial distribution is np , dividing mean by n will give p . Once p and n are known, Binomial probabilities and the expected frequencies can be computed.

Example 4

The number of sets of catla which responded in induced breeding out of 10 sets tried per experiment were noted. A total of 100 such experiments were conducted in a centre. The results are summarised in the following frequency distribution table :

Number responded (x)	0	1	2	3	4	5	6	7	8	9	10
Frequency (f)	1	1	1	2	4	12	22	27	19	9	2

Fit the Binomial distribution.

Answer

Here $n = 10, \sum fi = 100 = N$

$$np = \text{Mean} = \frac{\sum fix_i}{\sum fi} = \frac{659}{100} = 6.59$$

$$\text{Therefore, } p = \frac{6.59}{n} = 0.659 \approx 0.66$$

$$q = 1 - p \\ = 0.34$$

The Binomial probabilities for different values of x are computed using the Binomial distribution,

$$P(x) = {}_n C_x p^x q^{n-x}, \quad x = 0, 1, 2, \dots, 10$$

Computations are summarised below.

X	0	1	2	3	4	5	6	7	8
P (X)	0.0000	0.0004	0.0035	0.0181	0.0615	0.1434	0.2320	0.2573	0.1873
Expected* frequency	0	0.04	0.35	1.81	6.15	14.34	23.20	25.73	18.73

9 10

0.0807 0.0157

8.07 1.57

* Expected frequency = $N \times P(x)$

6.3.2 Poisson distribution

Poisson distribution is another discrete probability distribution which has frequent applications in faunal sampling operations where the character or variate under study is the number of animals or species per unit of observation. In practice, if the count data represent the number of rare events occurring within a given unit of time or space, the distribution of these counts can be described by the Poisson distribution. If 'p' the probability of occurrence of an event is very small and 'n' the number of trials is very large such that np is constant, then Binomial distribution tends to a Poisson distribution. Formula for the Poisson distribution is

$$P(X) = \frac{e^{-m} m^x}{x!}, \quad x = 0, 1, 2, \dots$$

where e is the base of natural logarithm having a value of 2.7183, m is the mean of the distribution. If m is known we can completely determine this distribution. An important characteristic of the Poisson distribution is that its variance is equal to the mean of the distribution. The Poisson distribution is positively skewed.

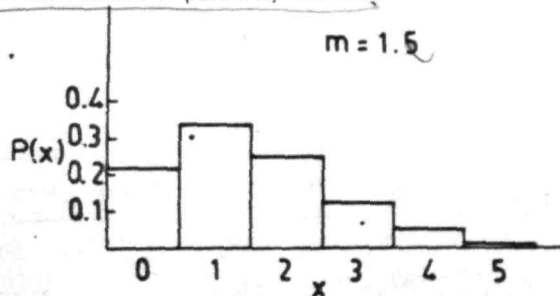


Fig. 4. Poisson distribution

However, as $m (= np, \text{ when } n \text{ is large})$ increases, it will tend to normal distribution. In Poisson distribution, it is assumed that rare events occur randomly and independently. Some examples of Poisson variable are number of ships arriving in a harbour per hour, number of animals per square of plankton species, number of car arrivals per minute at a toll bridge etc.

Example 5

The data given below refer to the number of animals per square of a particular species of plankton counted in a plankton counting cell.

Compute the Poisson probabilities and the expected frequencies

Number of animals per square (x)	0	1	2	3	4
Number of squares (f)	30	42	18	8	2

Answer

To compute Poisson probabilities, arithmetic mean 'm' of the distribution is required.

In the given example,

$$m = \frac{\sum fx}{n} = \frac{110}{100} = 1.1$$

The Poisson probabilities for different values of x are computed using the poisson distribution,

$$p(x) = \frac{e^{-m} m^x}{x!}, \quad x = 0, 1, 2, 3, 4$$

Computations are summarised below

X	Probability (P)	Expected frequency*
(1) X=0	$P(0) = e^{-1.1} = 0.3329$	33.29
(2) X=1	$P(1) = e^{-1.1} \cdot 1.1 = 0.3662$	36.62
(3) X=2	$P(2) = \frac{e^{-1.1} (1.1)^2}{2!} = 0.2014$	20.14
(4) X=3	$P(3) = \frac{e^{-1.1} (1.1)^3}{3!} = 0.0738$	7.38

Table contd..

X	Probability (P)	Expected frequency*
(5) X=4	$P(4) = \frac{e^{-1.1} (1.4)^4}{4!} = 0.0203$	2.03

* Expected frequencies are obtained by multiplying the respective probabilities by 'n', the total frequency.

6.3.3 Normal distribution

Normal distribution is one of the most important distributions in statistics. Its equation was first given by De Moivre in 1733. Later it was rediscovered and developed by Gauss in 1809 and by Laplace in 1812. Therefore, this distribution is sometimes referred to as Gaussian and Laplace distribution. The curve representing the normal distribution is called the normal curve which has the following equation,

$$y = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

where m and σ are respectively the mean and the standard deviation of the normal distribution.

π and e are constants whose values are equal to 3.1416 and 2.7183 respectively.

Normal distribution can be completely identified if mean (m) and standard deviation (σ) are known. The distribution will vary depending upon the values of m and σ (Fig. 5). It is continuous distribution and can theoretically assume any value from $-\infty$ to $+\infty$. However, for all practical purposes the values lie in the range of plus or minus three standard deviations from the mean.

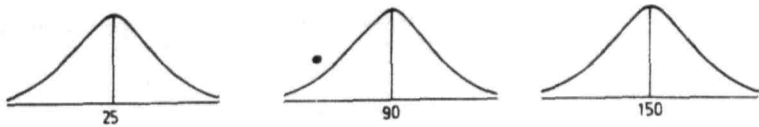


Fig 5a. Distributions with same standard deviation but different means

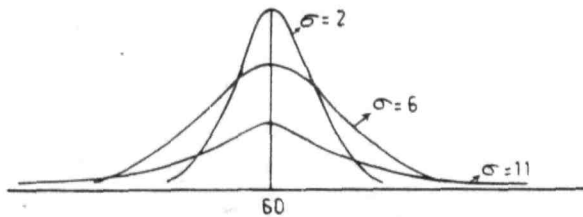


Fig 5b. Distributions with same mean but different standard deviations

6.3.3.1 Properties of normal curve

- (a) It is continuous, symmetrical and bell-shaped curve.
- (b) It is asymptotic. Both the tails extend to infinity, i.e., the tail approaches the base but never touches it.
- (c) The arithmetic mean, median and mode coincide.
- (d) The central position of the curve will be described by the mean and the spread of the curve by the standard deviation.
- (e) The coefficient of skewness is zero and the coefficient of kurtosis is 3.
- (f) (i) Mean plus or minus one standard deviation ($m \pm \sigma$) includes 68.00 per cent (68.27% to be more precise) of the total frequency or total area of the curve.

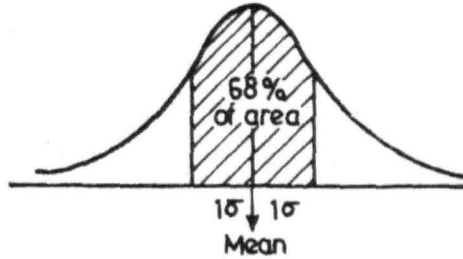


Fig. 6a. Area between $m \pm 1\sigma$

- (ii) Mean plus or minus 1.96 standard deviation ($m \pm 1.96\sigma$) includes 95% of the total frequency.

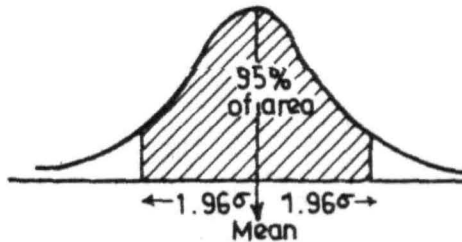


Fig. 6b. Area between $m \pm 1.96\sigma$

- (iii) Mean plus or minus 2.58 standard deviation ($m \pm 2.58\sigma$) includes 99% of the total frequency.

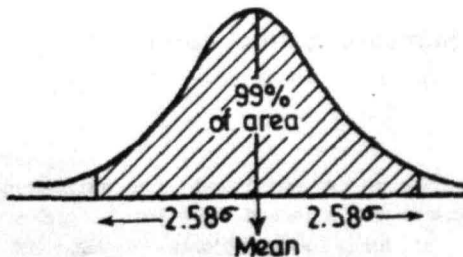


Fig. 6c. Area between $m \pm 2.58\sigma$

6.3.3.2 Area under the normal curve

The area bound by the normal curve and the x - axis is 1. Quite frequently the area under this curve that falls between two points on the x - axis, say, $x = a$ and $x = b$ is required. This area can be worked out using integral calculus. However, it is not necessary to work out the area by this method as tables giving the areas under the normal curve are available for ready use. These tables give the area under the normal curve which has mean zero and standard deviation one (called standard normal curve). Hence to make use of this table, we have to transform the normal variable X to a standard normal variable Z by the following relation,

$$Z = \frac{X - m}{\sigma}$$

As the standard normal curve is symmetric (Fig. 7) about $Z = 0$, the area between $Z = 0$ and any negative Z value, say, $Z = -a$, is equivalent to the area between $Z = 0$ and $Z = +a$.

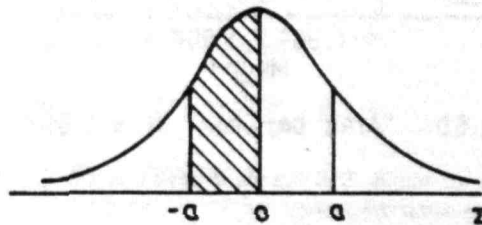


Fig.7. Standard normal curve.

Example 6

Weight of a particular species of fish was found to be distributed normally with the mean at 400 grams and standard deviation 50 grams. Find the standard normal variate of fishes with weights (i) 300, (ii) 450 and (iii) 430.

Answer

- (i) Weight $x = 300$, mean $m = 400$ and $\sigma = 50$
 Therefore, $Z = \frac{X-m}{\sigma} = \frac{300-400}{50} = \frac{-100}{50} = -2$
- (ii) When weight $X = 450$, standard normal variate
 $Z = \frac{X-m}{\sigma} = \frac{450-400}{50} = \frac{50}{50} = 1$
- (iii) When weight $X = 430$, standard normal variate
 $Z = \frac{X-m}{\sigma} = \frac{430-400}{50} = \frac{30}{50} = 0.6$

6.3.3.3 Different forms of area tables

Area under the normal curve is available in tables in different forms. For instance :

- (i) In Fisher and Yates (1963) the area of normal curve is tabulated from Z to ∞

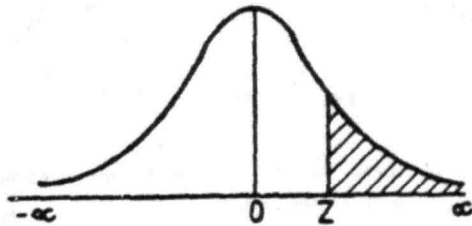


Fig. 7a. Area from Z to ∞

- (ii) In Spiegel (1981) area of the normal curve is tabulated from 0 to any positive value of Z

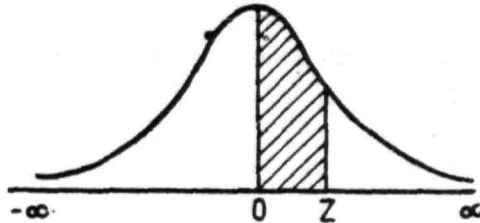


Fig. 7b. Area from 0 to Z

- (iii) In Woolf (1968) area of the normal curve is tabulated from $-\infty$ to Z .

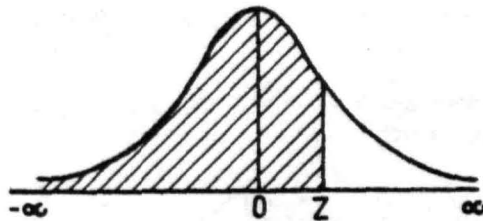


Fig. 7c. Area from $-\infty$ to Z

Before referring to these tables, it is therefore necessary to know the manner in which areas are presented.

in the present manual, area tables as presented in Spiegel (1981) are referred to.



Example 7

The mean length of a one year old brood of catla is 30 cm and standard deviation 2 cm. A fish is caught at random, find the probability that its length is,

- (i) (a) between 30 and 32 cm
(b) between 28 and 33 cm

(ii) Suppose it was decided to transfer all those having length greater than 31 cm, what per cent of fish is required to be transferred? Assume lengths are normally distributed.

Answer

- (i) (a) Compute the standard normal variate

Z , when $X_1 = 30$ and $X_2 = 32$

They are,

$$Z_1 = \frac{X_1 - m}{\sigma} = \frac{30 - 30}{2} = 0, \quad Z_2 = \frac{X_2 - m}{\sigma} = \frac{32 - 30}{2} = 1$$

The probability that the length of the fish caught is between 30 and 32 cm in terms of Z will be,

$$P(0 \leq Z \leq 1) = \text{Area between } (Z=0 \text{ and } Z=1)$$

$$= 0.3413$$

We get the area by referring to the area table of normal distribution.

- (b) Length between 28 and 33 cm i.e. $X_1 = 28$, $X_2 = 33$, corresponding standard normal variates are,

$$Z_1 = \frac{X_1 - m}{\sigma} = \frac{28 - 30}{2} = -1$$

$$Z_2 = \frac{X_2 - m}{\sigma} = \frac{33 - 30}{2} = \frac{3}{2} = 1.5$$

The probability that the length of the fish caught is between 28 and 33 cm in terms of Z will be

$$\begin{aligned}
 P(-1 \leq Z \leq 1.5) &= \text{Area between } Z=-1 \text{ and } Z=1.5 \\
 &= (\text{Area between } Z=-1 \text{ and } Z=0) + \\
 &\quad (\text{Area between } Z=0 \text{ and } Z=1.5) \\
 &= (\text{Area between } Z=0 \text{ and } Z=1) + \\
 &\quad (\text{Area between } Z=0 \text{ and } Z=1.5) \\
 &= 0.3413 + 0.4332 \\
 &= 0.7745
 \end{aligned}$$

(ii) Here $X_1 = 31$ cm hence, $Z = \frac{X-m}{\sigma} = \frac{31-30}{2} = 0.5$

$$\begin{aligned}
 P(\text{fish is having length greater than } 31 \text{ cm}) \\
 &= P(Z > 0.5) \\
 &= (\text{Area to the right of } Z=0) - (\text{Area between } Z=0 \text{ and } 0.5) \\
 &= 0.5 - 0.1915 \\
 &= 0.3085
 \end{aligned}$$

Therefore, 30.85% of the fishes require to be transferred.

6.3.3.4 Importance of normal distribution

Normal distribution plays an important role in statistics because of the following reasons :

- (1) Numerous continuous phenomena or characters such as fish length, weight, body depth etc., are approximately normally distributed.
- (2) Many of the discrete distributions occurring in practice such as binomial, poisson, etc., can be approximated by a normal distribution.
- (3) Even when the variable is not normally distributed, it is possible to bring it to approximately normal, by simple transformations such as square root or logarithmic or arc sign, etc.

- (4) Normal distribution has simple and interesting mathematical properties.
- (5) Normal distribution provides the basis for statistical inference (discussed in chapter 8 and 9).
- (6) Many of the distributions of sample statistics such as sample mean tend to normality for large 'n' and hence they can be studied with the help of normal distribution.

Chapter 7

SAMPLING DISTRIBUTIONS

7.1 Sampling distributions

From a population all possible samples of a given size can be drawn and for each sample a 'statistic' such as mean, standard deviation, etc., can be calculated. For example consider an artificial population of length (in cm) of 5 fingerlings 1, 2, 3, 4, 5 and with respective length values 3, 4, 7, 5 and 8. It is decided to estimate the mean length of fingerlings from a sample of 2. There will be 10 different samples of 2 fingerlings each, if sampling is without replacement and 25 samples with replacement, that can be drawn from a population of size 5. Possible samples of size 2 when the sampling is without replacement are listed in table 1. Similarly 25 possible samples can be listed when the sampling is with replacement.

Table 1 : Different possible samples of size 2, when sampling is without replacement.

S.No.	Sample consisting individuals	Sample values	Mean
1	1 and 2	3, 4	3.5
2	1 and 3	3, 7	5.0
3	1 and 4	3, 5	4.0
4	1 and 5	3, 8	5.5
5	2 and 3	4, 7	5.5
6	2 and 4	4, 5	4.5
7	2 and 5	4, 8	6.0
8	3 and 4	7, 5	6.0
9	3 and 5	7, 8	7.5
10	4 and 5	5, 8	6.5

In sampling without replacement, if a sample is randomly selected any one of the 10 samples listed in Table 1 is equally likely to be drawn. As can be seen from Table 1, means computed for different samples are not the same but they vary. Thus there is a distribution of the means, which is called the 'sampling distribution' of means. If variance

is calculated for different samples, it is likely to vary from sample to sample and sampling distribution of variance is obtained. Similarly, sampling distributions of other statistics such as standard deviation, median etc., can be obtained. Schematic representation of the concept of 'sampling distribution' is presented below :

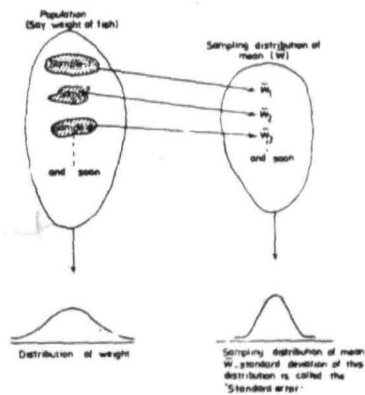


Fig 1: Concept of sampling distribution

7.2 Standard error

The Mean and standard deviation of sampling distributions of statistics can be computed as in the case of probability distributions based on individual observations. The standard deviation of sampling distribution of a statistic is called the 'standard error'. Generally, standard error decreases as the sample size increases.

7.2.1 Standard error of sampling distribution of mean

The standard error (SE) of mean of a sample is given by

$$\text{SE of mean} = \frac{\sigma}{\sqrt{n}} \left(\sqrt{\frac{N-n}{N-1}} \right)$$

where σ is the standard deviation of population. If the population is large (i.e., 'N' is large) or if sampling is with replacement then,

$$\text{SE of mean} = \frac{\sigma}{\sqrt{n}}$$

If the population standard deviation is not known then sample standard deviation 'S' can be used in its place, when n is large.

$$\text{Then, SE} = \frac{S}{\sqrt{n}}$$

$$\text{Where } S = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

7.2.2 Standard error of sum and difference of means

If \bar{X}_1 denotes the mean of a sample of size n_1 drawn from a population with variance σ_1^2 and \bar{X}_2 denotes the mean of a sample of size n_2 drawn from a population with variance σ_2^2 , then variance of the distribution of sum or difference of sample means is given by

$$\text{Variance } (\bar{X}_1 \pm \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\text{Standard error } (\bar{X}_1 \pm \bar{X}_2) = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)}$$

Example 1

Give the standard error of mean length for a sample of 25 fishes from a population with variance 4 cm.

Answer

Given

$$\text{Variance} = 4 \text{ cm} \quad \text{Hence, } \sigma = 2$$

Sample size, $n = 25$,

$$\text{Hence, standard error} = \frac{\sigma}{\sqrt{n}} = \frac{2}{5} = 0.4$$

Example 2

A sample of 100 fishes gave the estimates of mean and variance of a weight distribution of fishes as 800 and 100 respectively. What is the estimate of standard error of mean ?

Answer

Here the population standard deviation is not known. Hence, we have to take the sample standard deviation as its estimate.

Variance = 100, therefore standard deviation $S = 10$

Sample size, $n = 100$, hence $\sqrt{n} = 10$

therefore, standard error = $\frac{S}{\sqrt{n}} = \frac{10}{10} = 1$

Example 3

The standard deviation of the weight distribution of a certain species of fish is known to be 110 grms. An investigator wants to find out the mean weight of fish using a sample from this species. Determine the size of the sample required if it was decided that standard error of the mean should not exceed 5 grams.

Answer

Standard error of the mean = $\frac{\sigma}{\sqrt{n}}$

It is given that maximum standard error allowed is 5 grms. Therefore,

$$\frac{\sigma}{\sqrt{n}} = 5 \quad (1)$$

But it is given $\sigma = 110$

Hence, (1) becomes

$$\frac{(110)}{\sqrt{n}} = 5$$

$$\text{i.e. } n = \frac{(110)^2}{25} = \frac{12100}{25} = 484$$

Therefore, sample size required is 484 fish.

Example 4

The standard deviations of length in 2 fish populations are known to be in the ratio of 1:2. If altogether 100 fish are to be observed from the 2 populations, how many of each group should be observed in order to have the same precision on the estimate of the mean for both.

Answer

Let standard deviation of the 1st population be σ . Then standard deviation of the 2nd population will be 2σ . Let n_1 denote the number of fishes to be observed in 1st group. Then, standard error of 1st population = $\frac{\sigma}{\sqrt{n_1}}$

Let n_2 denote the number of fishes to be observed in 2nd population then, standard error of 2nd population = $\frac{2\sigma}{\sqrt{n_2}}$

It is given that the precision of estimates of mean of the first and second population to be the same. In other words, standard error of the first and second population to be the same.

$$\text{i.e. } \frac{\sigma}{\sqrt{n_1}} = \frac{2\sigma}{\sqrt{n_2}}$$

Squaring both sides gives

$$\frac{\sigma^2}{n_1} = \frac{4\sigma^2}{n_2}$$

Multiply both sides by $n_1 n_2$ to get $n_2 \sigma^2 = 4n_1 \sigma^2$

Divide both sides by σ^2 to get $n_2 = 4n_1 \dots \dots (I)$

It is given that altogether 100 fish are to be observed, in other words, $n_1 + n_2 = 100 \dots \dots (II)$

But it is known from (I) that $n_2 = 4n_1$

Substitute this in (II) to get

$$n_1 + 4n_1 = 100$$

$$\text{i.e. } 5n_1 = 100$$

$$\text{i.e. } n_1 = 20$$

$$\begin{aligned} \text{But by (1) it is known that } n_2 &= 4n_1 \\ &= 4 \times 20 = 80 \end{aligned}$$

Hence, observe 20 fishes in the first and 80 in the second population.

7.3 Important uses of standard error

Standard error plays an important role in statistical theory. Following are its important uses :

- (1) To measure the precision of a statistic. Higher the standard error lower is the precision of a statistic.
- (2) To fix confidence limits for population parameters.
- (3) To determine the size of the sample required to achieve the desired precision.
- (4) To compute test statistic in tests of significance.

7.4 Central limit theorem

If a random sample of n observations is drawn from a population with mean m and standard deviation σ , then the distribution of sample mean \bar{X} approaches the normal distribution with mean m and standard deviation

$$\frac{\sigma}{\sqrt{n}}; \text{ as } n \text{ increases.}$$

It should be noted that the central limit theorem does not specify that the sample comes from a normal population. The population from which the sample is drawn could be non-normal, still the mean will have a normal distribution.

This theorem occupies a unique place in drawing of inferences about populations based on random samples, when the sample size is large.

Chapter 8

ESTIMATION

8.1 Introduction

Statistical inference, a branch of statistics is concerned with drawing inferences about a population based on the information contained in a sample. One of the important functions of statistical inference is estimation of population parameters from the corresponding sample statistics.

8.2 Types of estimators

Population parameters can be estimated by two types of estimators viz: point estimators and interval estimators, the former estimation procedure being called the, 'point estimation' and the latter 'interval estimation'. In point estimation, an estimate of population parameter is specified by a single number, where as, in interval estimation, an estimate of population parameter is specified by two numbers between which the parameter may be considered to lie. For instance, marine fish landings during this year will be 1.8 million tons is an example of point estimate, whereas, marine fish landings during this year will be between 1.6 and 2 million tons is an example of interval estimate.

8.3 Properties of a good estimator

For a parameter there may be more than one estimator. Some estimators are better than the others. An estimator having the following properties is considered to be a good estimator :

(i) **Unbiasedness :**

An estimator is said to be unbiased if on an average, the value of the estimator equals the population parameter being estimated.

For example, in random sampling from a normal population, sample mean \bar{X} is an unbiased estimate of the population mean

(ii) Efficiency :

If the sampling distribution of two statistics have the same mean, the statistic with the smaller standard error is called an efficient estimator of the mean. Thus efficiency refers to the magnitude of the standard error.

For example, sampling distributions of the mean and median both have the same mean equal to the population mean, but variance of the sample distribution of means is smaller than that of the median. Hence, sample mean is an efficient estimator of population mean.

(iii) Sufficiency :

An estimator is said to be sufficient if it takes in to consideration all the possible information available from the sample.

For example, in random sampling from a normal population the sample mean is sufficient estimator of population mean, when σ^2 is known.

(iv) Consistency :

An estimator is said to be consistent if it approaches the value of the population parameter as the sample size increases.

For example, in a random sampling from normal population, the sample mean is a consistent estimator for the population mean.

8.4 Point estimation

Estimation of population mean and variance through point estimation procedure is discussed here with the following example.

Example 1

The prices of Peneus monodon (Rs./Kg) on 10 randomly selected days during a particular month in a local market were found to be :

88, 85, 82, 86, 85, 89, 90, 79, 81, 85

Estimate (i) the mean
and (ii) variance of prices of Peneus monodon during the month.

Answer

- (i) Sample mean \bar{X} is an unbiased estimator of the population mean. Hence, sample mean is computed to estimate the mean price of Peneus monodon during the month

$$\begin{aligned}\bar{X} &= \frac{\sum x_i}{n} \\ &= \frac{850}{10} = 85\end{aligned}$$

- (ii) An unbiased estimate of population variance is given by

$$\begin{aligned}S^2 &= \frac{1}{n-1} \sum (x - \bar{x})^2 \\ &= \frac{1}{n-1} \left(\sum x^2 - \frac{(\sum x)^2}{n} \right) \\ &= \frac{1}{9} \left(72362 - \frac{(850)^2}{10} \right) \\ &= \frac{1}{9} (112) = 12.44\end{aligned}$$

Note : $S = \sqrt{S^2} = 3.5277$ is an estimate of the population standard deviation, but this estimate is biased. The estimate is satisfactory when $n > 30$.

8.5 Interval estimation and confidence limits

In interval estimation, an interval is specified by two numbers, within which the parameter is considered to lie with a specified probability. This interval is defined by confidence limits with a certain degree of

probability. Higher, the probability, more is the confidence to be placed on the interval, to include the population parameter.

The addition or subtraction of 1.96 standard error to a statistic gives the confidence limits for population parameter with a probability of 0.95. These confidence limits are called 95 per cent confidence limits. Subtraction or addition of 2.58 standard error to a statistic gives confidence limits with probability of 0.99. These limits are called 99 per cent confidence limits. Like-wise the confidence limits with any desired level of probability can be computed. However, in estimation, 95 per cent and 99 per cent confidence limits are most commonly used.

The 95 per cent and 99 per cent confidence limits for the population mean are computed using the following formulae :

The 95 per cent confidence limits :

$$L_1, \text{ lower limit} = \bar{X} - 1.96 (\text{SE of Mean})$$

$$L_2, \text{ upper limit} = \bar{X} + 1.96 (\text{SE of Mean})$$

The 99 per cent confidence limits :

$$L_1, \text{ lower limit} = \bar{X} - 2.58 (\text{SE of Mean})$$

$$L_2, \text{ upper limit} = \bar{X} + 2.58 (\text{SE of Mean})$$

In general the population standard deviation σ is unknown. Hence, the sample standard deviation S has to be used in its place. As mentioned earlier this estimate is satisfactory when $n > 30$. But when n is less than 30, confidence intervals are computed using the table of t distribution which will be discussed later.

Just as the magnitude of standard error serves as a measure of reliability for a statistic, range of confidence limits also serves the same purpose. Smaller the range of the confidence limits the more reliable is the statistic as an estimate of the parameter.

Example 5

A random sample of $n = 100$ was selected to estimate the mean weight of fishes of a particular species. The sample mean was found to be

630 grams and the standard deviation 60 grams. Find 95% and 99% confidence limits for the population mean.

Answer

Standard deviation is given to be 60 grams i.e., $S = 60$ and $n = 100$

$$\begin{aligned} \text{Hence, standard error of mean} &= \frac{S}{\sqrt{n}} \\ (\text{SE of mean}) &= \frac{60}{\sqrt{100}} = 6 \end{aligned}$$

The 95% confidence limits for the population mean are,

$$\begin{aligned} \text{Lower limit} &= \bar{X} - 1.96 (\text{SE of mean}) \\ &= 630 - (1.96) (6) \\ &= 630 - 11.76 \\ &= 618.24 \end{aligned}$$

$$\begin{aligned} \text{Upper limit} &= \bar{X} + 1.96 (\text{SE of mean}) \\ &= 630 + (1.96) (6) \\ &= 630 + 11.76 \\ &= 641.76 \end{aligned}$$

Hence, the 95% confidence interval is 618.24 grams to 641.76 grams.

The 99% confidence limits are,

$$\begin{aligned} \text{Lower limit} &= \bar{X} - 2.58 (\text{SE of mean}) \\ &= 630 - (2.58) (6) \\ &= 630 - 15.48 \\ &= 614.52 \end{aligned}$$

$$\begin{aligned} \text{Upper limit} &= \bar{X} + 2.58 (\text{SE of mean}) \\ &= 630 + (2.58) (6) \\ &= 630 + 15.48 \\ &= 645.48 \end{aligned}$$

Hence, the 99% confidence interval will be 614.52 grams to 645.48 grams.

* * * * *

Chapter 9

TESTING OF HYPOTHESES

9.1 Introduction

Testing of hypothesis begins with a statement called hypothesis about a population in terms of its parameter (s). A sample drawn from this population is observed to verify the statement. The hypothesis is rejected if the sample provides ample evidence to do so, otherwise it is not rejected.

9.2 Terminology

9.2.1 Statistical hypothesis

✓ Statistical hypothesis is a statement about the population under study. It is usually a statement about one or more parameters of the population. Such statement may or may not be true. Some examples of hypothesis are mean weight of one year old oil sardine is 80 grams. Feed A and B are equally effective in increasing the weight of fish, the probability of getting number 4 when a dice is tossed is $1/6$.

9.2.2 Null hypothesis

✓ The hypothesis to be tested is commonly designated as "Null hypothesis" and is denoted usually by H_0 . For example, to decide whether one fish processing procedure is better than the other in terms of shelf life, the null hypothesis is formulated as 'there is no difference in the shelf life of two procedures'.

9.2.3 Alternative hypothesis

f Any admissible hypothesis that differs from a null hypothesis is called an alternative hypothesis and is denoted by H_1 . For example, in an experiment to compare the efficiency of 4 feeds, the hypothesis that there is no difference, is null hypothesis, whereas the hypothesis that there is significant difference among feeds is an alternative hypothesis.

9.2.4 Test statistic

It is a function of sample values. It extracts the information about population parameter contained in the sample. The observed value of the test statistic serves as a guide in rejecting or not rejecting the null hypothesis.

9.2.5 Rejection region

After the test statistic to be used is selected, the set of possible values of a statistic are divided into two mutually exclusive regions viz: rejection region (critical region) and acceptance region (Region of non rejection). If the observed value of a test statistic falls in the rejection region, H_0 is rejected. If it falls in the acceptance region, it is not rejected. It is to be noted that if the observed value falls in the acceptance region, it does not prove the hypothesis, it simply fails to disprove it.

9.2.6 Type I and type II errors

In testing a hypothesis two kinds of errors are likely to be committed. They are Type I and Type II errors. If null hypothesis is rejected when it is actually true, then such error is called Type I error. On the other hand, if null hypothesis is accepted when it is false, then Type II error is committed. This is summarised in the following table :

Table 1 : Statistical decision table

Actual situation	Test decision	
	Accept H_0	Reject H_0
H_0 true	Correct decision probability = $1 - \alpha$	Type I error probability = α
H_0 false	Type II error probability = β	Correct decision probability = $1 - \beta$

In order that any test of hypothesis to be good, it must be so designed as to minimise both the errors i.e., minimise both α and β

For a fixed sample size it is difficult to minimise both α and β , as an attempt to decrease one may lead to an increase in the other. It is customary to fix α at a predetermined level and choose a test procedure that minimises β i.e., α is prefixed in a test and β is minimised. Thus, we run the risk of rejecting a true H_0 $100\alpha\%$ times but reduce β , the acceptance of false H_0 to minimum. Test criterion are developed on these principles.

9.2.7 Level of significance

In testing a given hypothesis, the maximum probability with which we would be willing to risk a type I error is called the level of significance of the tests. In other words, it is a way of quantifying the amount of risk one wants to take in rejecting a true hypothesis. Usually 5% or 1% levels of significance are chosen. These levels, however, depend on the gravity of the risk vis a vis costs of decision making. To illustrate, suppose 5% level of significance is chosen in designing a test of hypothesis, then there are about 5 chances in 100 that the hypothesis is rejected when it should be accepted, i.e. one is 95% confident about the right decision.

9.2.8 Degrees of freedom

The number of independent observations available from the data for estimation of a particular parameter or a quantity is called the 'degrees of freedom'.

It can be calculated by deducting from the number of observations, the number of constants that are calculated from the data. For instance, the estimate of population variance based on a sample of 'n' observations is given by

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

In this case the constant (parameter), population mean, is estimated by the sample mean \bar{X} . Hence, deduct 1 from the total number of observations, "n" to get the degrees of freedom, i.e., the degrees of freedom of S^2 will be $n-1$.

9.3. Tests of hypothesis for large samples

9.3.1 Introduction

If a sample of size n is drawn from a normal population with mean m and standard deviation σ , then sample mean \bar{x} is also distributed as normal with mean m and standard deviation $\frac{\sigma}{\sqrt{n}}$. This proposition holds good even if the population from which the sample is drawn is not normal provided the sample size is large (see central limit theorem 7.4). As \bar{x} is distributed with mean m and standard deviation $\frac{\sigma}{\sqrt{n}}$, the standard normal variate is given by

$$Z = \frac{\bar{x} - m}{\sigma/\sqrt{n}}$$

Hence, under the hypothesis

$$H_0 : m = m_0, \text{ the test statistic}$$

$$Z = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}} \text{ is also a standard normal variate.}$$

If \bar{x}_1 and \bar{x}_2 denote the sample means based on n_1 and n_2 observations from populations with means m_1 and m_2 and standard deviations σ_1 and σ_2 respectively, then, from 7.2.2 standard deviation (error) of $(\bar{x}_1 - \bar{x}_2)$ is given by

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Mean of $(\bar{x}_1 - \bar{x}_2)$ is given by $(m_1 - m_2)$

$$\text{Hence, } Z = \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is a standard normal variate. Under the hypothesis

$$H_0 : m_1 = m_2 \text{ i.e., } (m_1 - m_2) = 0,$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The above discussions can be generalised as follows :

Suppose under the null hypothesis, the sampling distribution of S is normal with mean m_s and standard deviation σ_s , then the standard normal variable is given by

$$Z = \frac{S - m_s}{\sigma_s}$$

The area under the normal curve between $m - 1.96\sigma$ and $m + 1.96\sigma$ is 0.95 (see 6.3.1). Hence, in the case of standard normal variable which has mean zero and variance 1, the area between -1.96 to 1.96 will be 0.95. This, if the hypothesis is true, Z value computed from the sample will be between -1.96 to 1.96 with probability of 0.95. On the other hand if computed value of Z lies outside the range -1.96 to 1.96 , it can be concluded that such a sample would arise with only probability of 0.05, if the null hypothesis was true. In this case it is inferred that Z differs significantly from the value expected under the hypothesis and hence the hypothesis is rejected.

In the tests involving normal distribution, the set of values of Z outside the range -1.96 to 1.96 constitutes the region of rejection or critical region (Fig. 1).

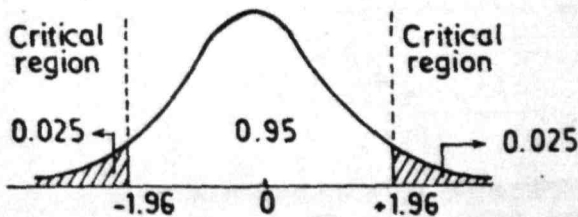


Fig. 1. Critical region, 5% level of significance

In the above discussion 5% level of significance was used. As mentioned earlier any level of significance (see 9.1.5) can be used. If 1% level of significance is used, the region of rejection will be outside the range - 2.58 to 2.58 (Fig. 2).

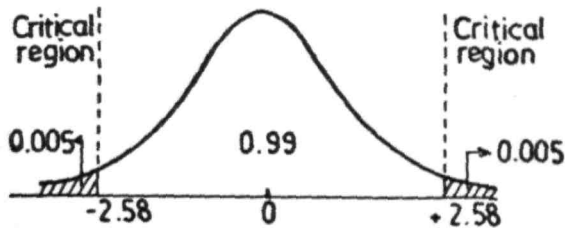


Fig. 2. Critical region, 1% level of significance

9.3.2 One-tailed and two-tailed tests

If the null hypothesis $H_0 : m = m_0$ is tested against $H_1 : m \neq m_0$ (which implies $m < m_0$ or $m > m_0$), then the interest is on extreme values of Z on both tails of the distribution. In such cases the critical region is on both the sides as shown in Figs. 1 and 2. Tests applied for such situations are called 'two-tailed' tests.

If the null hypothesis $H_0 : m = m_0$ is tested against $H_1 : m > m_0$, then the interest is in the extreme value to one side of the mean.

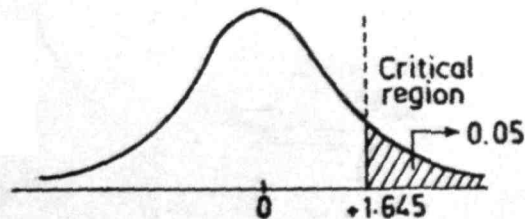


Fig. 3. One tailed test

In such cases the critical region will be to one side of the distribution as shown in Fig. 3. Tests applied to such situations are called 'one tailed' tests. It is to be noted that the critical value of Z at 5% and 1% level of significance for one tailed test are 1.645 and 2.33, whereas, these values are 1.96 and 2.58 for two tailed tests.

9.3.3 Test for single mean

Let x_1, x_2, \dots, x_n be the values of a variable X , in a random sample of size n from a population with mean m and variance σ^2 . On the basis of this sample, the hypothesis regarding the value m is tested. The null hypothesis tested is,

$$H_0 : m = m_0,$$

where m_0 is a specified value.

The following test statistic is computed

$$Z = \frac{(\bar{x} - m_0) \sqrt{n}}{\sigma}$$

where \bar{x} is the sample mean.

- If $|Z| \geq 1.96$, reject H_0 at 5% level of significance.
- If $|Z| \geq 2.58$, reject H_0 at 1% level of significance.

Example 1

A random sample of 144 fishes drawn from a certain species showed a mean length of 28 cm. Can this be considered as a sample from a population with mean 30 cm and standard deviation 16 cm?

Answer

$$H_0 : m = 30 \quad \text{--- } \bar{x}$$

Test statistic used is

$$Z = \frac{(\bar{x} - m_0) \sqrt{n}}{\sigma}$$

$$= \frac{(28 - 30) \sqrt{144}}{16}$$

$$= \frac{-24}{16} = -1.5$$

Since $|Z| < 1.96$, H_0 is not rejected.

Example 2

A company used to manufacture nylon twines with mean breaking strength of 3.5 kg and standard deviation 2 kg. The company now claims that by a newly developed process the mean breaking strength can be increased. A sample of 64 twines taken from this new process, gave mean of 4 kg. The standard deviation of the new process is assumed to be the same as the old process. Can the company's claim be accepted at 5% level of significance?

Answer

$$m = \bar{x}$$

$$H_0 : m = 3.5$$

$$H_1 : m > 3.5$$

The test statistic to be computed is

$$Z = \frac{(\bar{x} - m)}{\sigma / \sqrt{n}} = \frac{(4 - 3.5) \sqrt{64}}{2}$$

$$= \frac{0.5 \times 8}{2}$$

$$= 2$$

As $|Z| > 1.645$, the null hypothesis is rejected. Hence, the company's claim can be accepted.

9.3.4

Tests for equality of two population means

9.3.3

Let \bar{X}_1 be the mean of a sample of size n_1 from a population with mean m_1 and standard deviation σ_1 and let \bar{X}_2 be the mean of another sample of size n_2 from a population with mean m_2 and standard deviation σ_2 . To test the equality of population means the following null hypothesis is set up :

$$H_0 : m_1 = m_2$$

The test procedure is to calculate,

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

If $|Z| > 1.96$, reject H_0 at 5% level of significance.

If $|Z| > 2.58$, reject H_0 to 1% level of significance.

If σ_1 and σ_2 are not known the sample standard deviations are used to estimate them.

Note : If the samples have been drawn from populations with common standard deviation then

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \sigma}$$

Example 3 ✓

The mean length of 100 fish caught by sampling gear A was 8 cm with standard deviation of 2 cm, whereas the mean length of 120 fish caught by sampling gear B was 8.5 cm with standard deviation of 2.2 cm. Is there significant difference between the lengths of fish caught by the two gears at 5% level of significance?

Answer

H_0 : There is no significant difference between the length of fish caught by the two gears.

$$\begin{aligned}
 Z &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \\
 &= \frac{8 - 8.5}{\sqrt{\frac{(2)^2}{100} + \frac{(2.2)^2}{120}}} \\
 &= \frac{-0.5}{\sqrt{\frac{4}{100} + \frac{4.84}{120}}} \\
 &= \frac{-0.5}{\sqrt{0.04 + 0.0403}} \\
 &= \frac{-0.5}{\sqrt{0.0803}} \\
 &= -1.7645
 \end{aligned}$$

Since $|Z| < 1.96$ and also 2.58 , H_0 is not rejected at 5% and 1% level of significance.

9.4 Tests of hypothesis for small samples ($n < 30$)

When the size of the sample is small, the distribution of various statistics are far from normality and hence tests of hypothesis based on normal variate cannot be applied. In such cases tests of hypothesis based on exact sampling distribution of 't' and 'F' are applied. When applying these tests it is assumed that the population from which the sample is drawn is normal.

9.4.1 The t - distribution

The t - distribution is a sampling distribution derived from the parent normal distribution. This distribution is symmetrical about the mean but is slightly flatter than the normal distribution. Unlike the normal distribution it will be different for different size of the sample 'n' or the degrees

of freedom n-1. When the size of the sample is very small (< 30), the t - distribution markedly differs from normal distribution, but as n increases the t - distribution resembles more and more a normal distribution (figure 4). The values of 't' have been tabulated for different degrees of freedom at different levels of significance (Fisher and Yates, 1963)

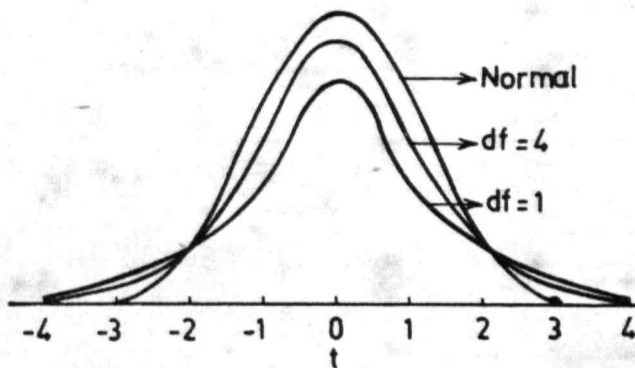


Fig.4. Students t distribution for varying degrees of freedom (df)

Test of hypothesis based on t distribution are discussed below :

9.4.1.1 Test for single mean

Let X_1, X_2, \dots, X_n be a random sample of size n drawn from a normal population with mean μ . Let \bar{x} and S^2 denote mean and variance of the sample. To test the hypothesis $H_0 : \mu = \mu_0$ the following test procedure is used :

Compute

$$t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{S} \quad \text{where } S = \sqrt{S^2}$$

This t follows t distribution with $n-1$ degree of freedom.

If $|t| >$ the table value of t at 5% level of significance, then reject H_0 at 5% level of significance.

If $|t| >$ the table value of t at 1% level of significance, then reject H_0 at 1% level of significance.

Example 4

A sample of 25 fingerlings drawn from a rearing tank showed a mean length of 75.8 mm and standard deviation of 10 mm. Is the data consistent with the claimed mean size of 80 mm?

H_0 : Sample is drawn from a population with mean 80 mm.
Calculate

$$\begin{aligned} t &= \frac{(\bar{X} - m_0) \sqrt{n}}{S} \\ &= \frac{(75.8 - 80) \sqrt{25}}{10} \\ &= \frac{(-4.2) 5}{10} = -2.1 \end{aligned}$$

The table values of t with 24 degrees of freedom are 2.064 at 5% and 2.797 at 1% level of significance. Since $|t| >$ the table value of t at 5% level H_0 is rejected at 5% level, but as $|t| <$ the table value of t at 1% level, H_0 is not rejected at 1% level of significance.

9.4.1.2 Testing of difference between two means (population variances assumed equal)

Let \bar{X} and S_1 be the mean and standard deviation of a sample of size n_1 from a normal population with mean m_1 and let \bar{Y} and S_2 be the mean and standard deviation of another sample of size n_2 from a normal population with mean m_2 . To test whether the population means differ significantly, the following null hypothesis is set up:

$$H_0 : \mu_1 = \mu_2$$

To test H_0 , calculate,

9.3

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which is distributed as t with $n_1 + n_2 - 2$ degrees of freedom.

S in the above expression is computed using the formula

$$S = \sqrt{\frac{\sum (X - \bar{X})^2 + \sum (Y - \bar{Y})^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}}$$

If $|t| >$ the table value of t at the specified level of significance, reject the hypothesis at that level.

Example 5

Weight was recorded separately for male and female one year old fish of species A. The mean weights of males and females are :

Sex	Sample Size	Mean weight (g)	Variance
Male	9	70	25
Female	11	61	16

Is there real difference in the average weight between the sexes?

Answer

H_0 : Samples come from the populations with the same mean. In other words, there is no significant difference between the mean weights of males and females.

To test the null hypothesis, calculate,

$$t = \frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Given, $\bar{X} = 70$, $\bar{Y} = 61$

$$S_1^2 = 25, S_2^2 = 16$$

$$n_1 = 9, n_2 = 11$$

First, calculate

$$\begin{aligned} S &= \sqrt{\frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}} \\ &= \sqrt{\frac{8 \times 25 + 10 \times 16}{18}} \\ &= \sqrt{\frac{200 + 160}{18}} \\ &= \sqrt{20} \\ &= 4.47 \end{aligned}$$

$$\begin{aligned} \text{Therefore, } t &= \frac{70 - 61}{(4.47) \sqrt{\frac{1}{9} + \frac{1}{11}}} \\ &= \frac{9}{(4.47) \sqrt{\frac{20}{99}}} \\ &= \frac{9}{(4.47) \sqrt{0.202}} \\ &= \frac{9}{(4.47) (0.4494)} = \frac{9}{2.009} = 4.48 \end{aligned}$$

The table value of t with 18 degrees of freedom is 2.101 at 5% and 2.878 at 1% level of significance. Since $|t| >$ the table values of t at 5% and 1% level of significance. H_0 is rejected i.e., there is significant difference between the mean weights of males and females.

- 9.4.1.3 Test of difference between two means of correlated observations (paired t - test):** When the two samples of equal size are drawn from two normal populations and these samples are not independent, then the paired t -test is used. Dependent samples arise, for instance, in experiments when an individual is tested first under one condition and then under another condition, so that there will be two observations for the same individual. Let n be the size of each of the two samples and d_1, d_2, \dots, d_n the difference between the corresponding members of the sample. Let \bar{d} denote the mean of differences and S the standard deviation of these differences.

The null hypothesis to be tested is

$$H_0 = \mu_1 = \mu_2$$

where μ_1 and μ_2 are means of 1st and 2nd population respectively.

To test this hypothesis, compute,

$$t = \frac{(\bar{d}) \sqrt{n}}{S}$$

It is distributed as t with $(n-1)$ degrees of freedom.

If $|t| >$ the table value at the specified level of significance, reject H_0 at that level.

Example 6

The following table gives the marks obtained by 9 students in two tests, one held at the beginning of a year and the other at the end of a year after intensive coaching. Do the data indicate that the students have benefitted by coaching?

Answer

H_0 : coaching has no effect.

Student	1	2	3	4	5	6	7	8	9
Test 1	55	60	65	75	49	25	35	18	61
Test 2	63	70	70	81	54	29	32	21	70
Difference (di)	8	10	5	6	5	4	-3	3	9

$$\sum di = 47$$

$$di^2 \quad \quad \quad 64 \quad 100 \quad 25 \quad 36 \quad 25 \quad 16 \quad 9 \quad 9 \quad 81$$

$$\sum di^2 = 365$$

$$\bar{d} = \sum \frac{di}{9} = \frac{47}{9} = 5.22$$

$$S = \sqrt{\frac{1}{n-1} (\sum di^2 - \frac{(\sum di)^2}{n})}$$

$$= \sqrt{\frac{1}{8} (365 - \frac{(47)^2}{9})} = \sqrt{\frac{1}{8} (365 - \frac{2209}{9})}$$

$$= \sqrt{\frac{1}{8} (365 - 245.44)} = \sqrt{\frac{119.56}{8}}$$

$$= \sqrt{14.945} = 3.866$$

$$\text{Therefore, } t = \frac{(\bar{d})\sqrt{n}}{S}$$

$$= \frac{(5.22)(\sqrt{9})}{3.866} = \frac{(5.22)(3)}{3.866}$$

$$= \frac{15.66}{3.866} = 4.05$$

The table value of t with 8 degrees of freedom is 2.306 at 5% level of significance and 3.355 at 1% level of significance.

Since $|t| >$ the table value of t at both 5% and 1% level of significance, H_0 is rejected. In other words it is concluded that the coaching has benefited the students.

9.4.1.4 Confidence limits for population mean μ

In chapter No. 8, computation of confidence limits for population mean μ , based on large samples using normal distribution was discussed. It was pointed out there that for samples with size less than 30, 't' distribution is used for computing confidence limits. The formula for computing confidence limits using 't' distribution is as follows :

$$\text{Upper limits} = \bar{X} + t \frac{S}{\sqrt{n}}$$

$$\text{Lower limit} = \bar{X} - t \frac{S}{\sqrt{n}}$$

Where \bar{X} stands for the sample mean
 S stands for sample standard deviation
 n stands for sample size
 t stands for the value of t with $n-1$ df which can be obtained from the table of t values, at 5% or 1% level of significance depending upon whether 95% or 99% confidence limits are computed.

Example 7

Following data refer to catch (in tons) per haul of one hour duration in a trawl survey off a certain coast.

1.2, 2.5, 1.0, 4.0, 3.0, 2.8, 0.6, 3.4, 2.5, 2.0

Compute mean catch per hour and also 95% confidence limits for catch per hour for the coast (population) under survey.

Answer

95% confidence limits are given by

$$\bar{X} \pm t \frac{S}{\sqrt{n}}$$

To calculate these confidence limits the following computations are to be made :

Haul No. :	1	2	3	4	5	6	7	8	9	10	Total
Catch/hour (X)	1.2	2.5	1.0	4.0	3.0	2.8	0.6	3.4	2	2.5	23.0
X ²	1.44	6.25	1.0	16.0	9.0	7.84	0.36	11.56	4.0	6.25	63.7

$$\text{Mean, } \bar{X} = \frac{23}{10} = 2.3$$

$$S^2 = \frac{1}{n-1} \left(\sum X^2 - \frac{(\sum X)^2}{n} \right) = \frac{1}{9} (10.80)$$

$$= 1.20, \text{ Hence, } S = 1.0954$$

From 't' table, the value of 't' with 9 d.f. at 5% level of significance is 2.2620.

$$t \frac{S}{\sqrt{n}} = \frac{(2.262)(1.0954)}{\sqrt{10}} = 0.7835$$

$$\text{Hence, upper limit} = \bar{X} + t \frac{S}{\sqrt{n}}$$

$$= 2.3 + 0.7835$$

$$= 3.0835$$

$$\begin{aligned}
 \text{Lower limit} &= \bar{X} - t \frac{S}{\sqrt{n}} \\
 &= 2.3 - 0.7835 \\
 &= 1.5165
 \end{aligned}$$

Thus mean catch per hour is expected to be between 1.5165 and 3.0835 tonnes.

9.5 The Chi-square (X^2) distribution

Theoretically, the X^2 distribution can be defined as the sum of squares of independent normal variates. If X_1, X_2, \dots, X_n are n independent standard normal variates, then sum of squares of these variates,

$X_1^2 + X_2^2 + \dots + X_n^2$ follows the X^2 distribution with n degrees of freedom. The shape of X^2 distribution depends on n , the degree of freedom which is also its mean (Fig.5). When n is small, the X^2 distribution is markedly different from normal distribution but as n increases the shape of the curve becomes more and more symmetrical and for $n > 30$,

it can be approximated by a normal distribution. The values of X^2 have been tabulated for different degrees of freedom at different levels of probability. (Fisher and Yates, 1963)

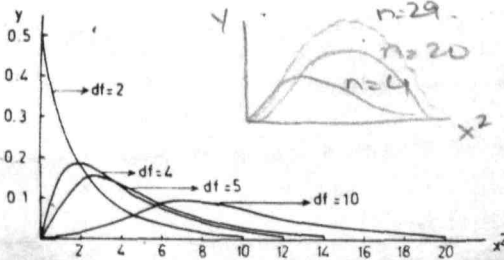


Fig 5 Chi-square distribution for different degrees of freedom (df)

Most data on biological investigations can be classified either as quantitative or qualitative (attribute) data. The statistical procedures discussed so far apply mostly to quantitative data. There are many instances in fisheries research, wherein attribute data describe the phenomenon under investigations more adequately than quantitative data. The chi-square test based on X^2 - distribution is commonly used for analysis of attribute data.

9.5.1 Test for fixed-ratio hypothesis

Many investigations are carried out to verify empirically some biological phenomena that are expected to occur under some given assumptions. For instance, a ratio of 3:1 is expected to occur in the F_2 generation of a cross between tall and dwarf plants. Whether this hypothesis of 3:1 ratio is substantiated by the actual observed data can be ascertained by χ^2 - test. This χ^2 test can be applied to test any fixed ratio hypothesis provided the expected ratio is specified before the investigation commences.

If O_i refers to observed frequency and E_i refers to the expected frequency based on the expected ratio hypothesis, then χ^2 is computed as follows:

$$\begin{aligned}\chi^2 &= \sum_1^k \frac{(O_i - E_i)^2}{E_i} \\ &= \sum_1^k \frac{O_i^2}{E_i} - n \quad \text{----- (1)}\end{aligned}$$

where n is the total number of observations and k is the number of classes. The χ^2 in (1) has $k-1$ degrees of freedom. In this test the expected frequency of each class should be more than 5. If any such frequency is small adjacent classes may be grouped, so that the expected frequency is more than 5.

If the calculated value of χ^2 is greater than the table value of χ^2 with $(k-1)$ df, at specified level of significance the null hypothesis of specified ratio is rejected.

Example 8

A sample of 500 fish observed for determining the sex ratio, indicated that 230 were male and 270 female. Do the observed data fit the expected ratio of 1:1 ?

Answer

H_0 : The observed data fit the ratio 1:1.

On the basis of this hypothesis of 1:1 ratio, 250 fish are expected in male and female classes. χ^2 is calculated as follows :

Sex	Frequency		O_i^2	$\frac{O_i^2}{E_i}$
	Observed (O_i)	Expected (E_i)		
Male	230	250	52900	211.60
Female	270	250	72900	291.60
Total				503.20

$$\begin{aligned}\chi^2 &= \sum \frac{O_i^2}{E_i} - n \\ &= 503.2 - 500 \\ &= 3.2\end{aligned}$$

The table value of χ^2 with 1 df at 5% level of significance is 3.841. As χ^2 computed is less than the table value of χ^2 , the hypothesis is not rejected.

9.5.2 Goodness of fit test for probability distributions

Another important application of χ^2 is in testing if a set of quantitative data follows a specific probability distribution. In this test actual frequency in each category (or class interval) are compared with the frequencies that could be theoretically expected if the data followed the hypothesized probability distribution. To perform this test following steps are followed :

- (i) Hypothesize the probability distribution to be fitted.
- (ii) Values of each parameter of selected probability distribution is estimated from the given data if not specified.

- (iii) Theoretical frequencies for each class are estimated based on the hypothesised probability distribution.
- (iv) The following chi-square test statistic is computed

$$\chi^2 = \sum_i^k \frac{O_i^2}{E_i} - n$$

It has $(k-1)$ df, where k is the number of classes.

- (v) If the expected frequency of any class is less than 5, the adjacent classes can be grouped to form a class, so that expected frequency is more than 5.

(vi) If the expected frequencies are calculated on the basis of certain parameters estimated from data, the degrees of freedom for χ^2 is not $(k-1)$ but is decreased by the number of parameters estimated.

- (vii) If χ^2 computed in step (iv) is greater than the tabular value of χ^2 with $(k-1)$ df at specified level significance, the null hypothesis that selected probability distribution is a good fit to the given data is rejected.

Example 9

Test whether the data on number of animals per square of a particular species of plankton given in example 5 of chapter 6 follows poisson distribution.

Answer

H_0 : Number of animals per square of a particular species of plankton follows Poisson distribution.

Expected frequencies using Poisson probability distribution have already been computed in example 5 of chapter 6. Hence based on observed and expected frequencies χ^2 can be computed as outlined below :

x	O_i	E_i	$\frac{O_i^2}{E_i}$
0	30	33.29	27.04
1	42	36.62	48.17
2	18	20.14	16.09
3	8	7.38	8.67
4	2	2.03	1.97
Total	100		101.94

$$\begin{aligned}
 \chi^2 &= \sum \frac{O_i^2}{E_i} - n \\
 &= 101.94 - 100 \\
 &= 1.94
 \end{aligned}$$

As the mean m of the distribution is estimated from the sample, number of degrees of freedom = $k-1-1 = 3$.

The table value of χ^2 with 3 df at 5% level of significance is 7.815

Since $\chi^2 = 1.94 < 7.815$

The null hypothesis is not rejected.

9.5.3 χ^2 - test for independence of attributes in 2 x 2 contingency table

Suppose that an attribute data of size n is classified according to two attributes, say, A and B and the attribute A is further subdivided into two classes A_1 and A_2 and the attribute B into B_1 and B_2 . Such attribute data can be presented in the form of a table called 2 x 2 contingency table as shown below.

Table 2 : 2 x 2 contingency table

B \ A	A ₁	A ₂	Total
B ₁	a	b	a+b
B ₂	c	d	c+d
Total	a+c	b+d	a+b+c+d = n

H₀ : The two attributes A and B are independent. It is tested by the X² test. A simple formula for computing X² of a 2 x 2 contingency table is given by,

$$X^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Where a, b, c and d are cell frequencies of 2 x 2 contingency table and n is the total frequency. This X² has 1 degree of freedom. If, the expected cell frequencies are large, the discrete distribution of probabilities of all frequencies approximate to normal distribution. This approximation holds good fairly well when the degrees of freedom are more than 1 and the expected frequency in the various classes is not small. As the degrees of freedom of X² statistic of 2 x 2 contingency table is 1, X² approximation in this case will not be satisfactory and leads to over estimation of significance. This is corrected by the method suggested by Yates which is known as 'Yates correction'. The correction consists of adding 1/2 to the observed minimum frequency and adjusting the other cell frequency for the observed marginal totals and then computing the X². Formula for X² using the Yates correction in a 2 x 2 contingency table is given by,

$$X^2 = \frac{n \left(\left| ad - bc \right| - \frac{n}{2} \right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

This correction is suitable when the expected frequency of classes is less than 5, but estimation with correction can do no harm even when the frequencies are large. Hence it is always better to use the correction as a matter of routine.

Example 10

In a series of experiments to test whether advanced stages of Mycobolus infection is cured by lime treatment, the following observations were found :

	Not cured	Cured	Total
Lime treated :	86	14	100
Untreated (control):	88	12	100
	174	26	200

Test whether lime has any effect in curing the infection.

Answer

Ho : There is no association between lime treatment and the curing of infection.

$$\chi^2 = \frac{n (ad - bc)^2}{(a+b) (c+d) (a+c) (b+d)}$$

where a = 86, b = 14, c = 88, d = 12

$$\begin{aligned} \chi^2 &= \frac{200 (86 \times 12 - 14 \times 88)^2}{100 \times 100 \times 174 \times 26} \\ &= \frac{200 (200)^2}{10000 \times 174 \times 26} = 0.1768 \end{aligned}$$

χ^2 with Yates correction

$$\begin{aligned} \chi^2 &= \frac{n (|ad - bc| - \frac{n}{2})^2}{(a+b) (c+d) (a+c) (b+d)} = \frac{200 (|86 \times 12 - 14 \times 88| - \frac{200}{2})^2}{100 \times 100 \times 174 \times 26} \\ &= 0.0442 \end{aligned}$$

Since X^2 calculated (with and without Yates correction) is less than the table value (3.84 at 5%, 6.64 at 1%), H_0 is not rejected.

9.5.4 Computation of X^2 in $r \times c$ contingency table

The $r \times c$ contingency table is an extension of 2×2 contingency table in which the data are classified into ' r ' rows and ' c ' columns (table 3). In this table the frequencies which occupy cells of the table are called 'cell frequencies' whereas row and column totals are called the 'marginal frequencies'.

Table 3 : A $r \times c$ contingency table

$A \backslash B$	B_1	B_2	..	B_j	..	B_c	Total
A_1	011	012	..	01j	..	01c	(A_1)
A_2	021	022	..	02j	..	02c	(A_2)
..
A_i	0i1	0i2	..	0ij	..	0ic	(A_i)
..
A_r	0r1	0r2	..	0rj	..	0rc	(A_r)
Total	(B_1)	(B_2)	..	(B_j)	..	(B_c)	n

As the table consists of ' r ' rows and ' c ' columns, there will be $r \times c$ observed frequencies, one in each cell. Corresponding to each observed frequency, there is expected frequency, computed based on certain hypothesis. Under the null hypothesis of no relationship or of independence between the attributes, expected frequency of each cell is computed

by multiplying totals of the row and column to which the cell belongs divided by the total number of observations. For instance, the expected frequency of the cell in 1st row and 2nd column is obtained by multiplying the 1st row total (A_1) with the 2nd column total (B_2) and then dividing by the total number of observations, 'n'. After calculating the expected frequencies for each cell, X^2 is computed using the formula,

$$X^2 = \sum \frac{O_i^2}{E_i} - n$$

which has (r-1) (c-1) degrees of freedom.

Example 11

In a fish tagging experiment, the length frequency of tagged fishes and recoveries were as under. Test whether the length distributions can be accepted as same?

	Length group (cm)					Total
	10-20	20-30	30-40	40-50	50-60	
Fishes tagged	108	140	256	358	111	1000
Fishes recovered	9	15	28	40	8	100

Answer

H_0 : There is no change in the length distribution

	Length (cm)					Total
	10-20	20-30	30-40	40-50	50-60	
Fishes tagged	108	140	256	385	111	1000
Fishes recovered	9	15	28	40	8	100
Total :	117	155	284	425	119	1100

To compute χ^2 statistic the following computations are to be made:

Observed frequency (O_i)	Expected Frequency (E_i)	$\frac{O_i^2}{E_i}$
1. 108	$\frac{1000 \times 117}{1100} = 106.36$	109.66
2. 140	$\frac{1000 \times 155}{1100} = 140.91$	139.10
3. 256	$\frac{1000 \times 284}{1100} = 258.18$	253.84
4. 385	$\frac{1000 \times 425}{1100} = 386.37$	383.63
5. 111	$\frac{1000 \times 119}{1100} = 108.18$	113.89
6. 9	$\frac{1000 \times 117}{1100} = 10.64$	7.61
7. 15	$\frac{1000 \times 155}{1100} = 14.09$	15.97
8. 28	$\frac{1000 \times 284}{1100} = 25.82$	30.36
9. 40	$\frac{1000 \times 425}{1100} = 38.63$	41.42
10. 8	$\frac{1000 \times 119}{1100} = 10.82$	5.92
Total		1101.40

$$\chi^2 = \sum \frac{O_i^2}{E_i} - n$$

$$= 1101.40 - 1,000 = 1.40$$

Table value of X^2 with 4 df at 5% level of significance is 9.488. As the calculated value of X^2 is less than the table value of X^2 , the null hypothesis is not rejected.

9.55 Test of hypothesis about a population variance

Let x_1, x_2, \dots, x_n be the values of a variable in a random sample of size n drawn from a normal population with variance σ^2 .

The null hypothesis to be tested is, $H_0: \sigma^2 = \sigma_0^2$, where σ_0^2 is a specified value. Test statistic used is

$$X^2 = \frac{(n-1) S^2}{\sigma^2} \quad \text{----- (II)}$$

where S^2 is the sample variance, X^2 in equation (II) is distributed as X^2 with $(n-1)$ df.

If alternative hypothesis to be tested is

(a) $H_1: \sigma^2 \neq \sigma_0^2$

then reject H_0 if

$$X_{cal}^2 > X_{\alpha/2}^2 \quad \text{or} \quad X^2 < \frac{X_{1-\alpha/2}^2}{2}$$

(b) $H_1: \sigma^2 > \sigma_0^2$

then reject H_0 is $X^2 > X_{\alpha}^2$

(c) $H_1: \sigma^2 < \sigma_0^2$,

reject H_0 if $X^2 < X_{1-\alpha}^2$

Example 12

A market survey conducted on 50 house holds, indicated that the average expenditure of house holds is Rs.40 per week on purchase of fish with

standard deviation of Rs.22. Can this data be considered as a sample from a population with variance of Rs.400, at 5% level of significance?

Answer

$$H_0 : \sigma^2 = 400$$

$$H_1 : \sigma^2 \neq 400$$

Test statistic to be computed is

$$\chi^2 = \frac{(n-1)(S)^2}{\sigma^2}$$

$$= \frac{(50-1)(22)^2}{400}$$

$$= \frac{49 \times 484}{400}$$

$$= 59.29$$

25.25

Reject H_0 if, $\chi^2 > \chi^2_{0.025}$ OR $\chi^2 < \chi^2_{0.975}$

Otherwise do not reject it

$$\text{For 49 df, } \chi^2_{0.025} = 70.222 \text{ and } \chi^2_{0.975} = 31.555$$

As χ^2 lies between 31.555 and 70.222, H_0 is not rejected.

9.6

The F distribution

Theoretically F distribution can be defined as the ratio of two independent χ^2 variates with n_1 and n_2 d.f. The shape of F distribution completely depends on n_1 and n_2 . As n_1 and n_2 increase without limit the F distribution approaches a normal distribution, if $n_1 = 1$ and n_2 increases without limit, F follows the t distribution. i.e., $F = t^2$. Thus F distribution embraces wide ranges of distributions like normal, χ^2 and t and lends itself to a large number of applications. Two important uses of 'F' are 'testing equality of two variances' and 'testing the equality of several means'. To arrive at significance, computed value of F has to be compared

deals

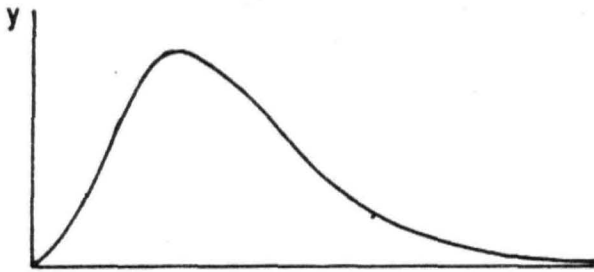


Fig. 6. F Distribution

with the table value of F . The values of F tabulated for different degrees of freedom and levels of significances are available (Fisher and Yates, 1963).

9.6.1 F test for testing equality of two variances

Let S_1^2 be the variance of a sample of size n_1 and S_2^2 be the variance of a sample of size n_2 . To test the null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ i.e., population variances are equal, the following test procedure is used :

$$\begin{aligned} \text{Compute : } F &= \frac{S_1^2}{S_2^2} \\ &= \frac{\frac{1}{n_1 - 1} \sum (x - \bar{x})^2}{\frac{1}{n_2 - 1} \sum (y - \bar{y})^2} \end{aligned}$$

This follows F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Generally larger sample variance is taken in the numerator while computing F .

Computed value of F is to be compared with F with $n_1 - 1$ and $n_2 - 1$ degrees of freedom at the desired level of significance to draw conclusions. If $F_{\text{cal}} > F_{\text{table}}$, reject H_0 at the selected level of significance.

Example 13

Data on daily fish landings recorded for 30 days in a landing centre showed variance of 60 kg whereas landings recorded for 25 days in another centre showed variance of 40 kg. Test whether the variability of daily fish landings is the same in two landing centres?

Answer

H_0 : Population variances are equal.

$$F = \frac{S_1^2}{S_2^2} = \frac{60}{40} = 1.5$$

Compare this value with the table value of F with 29 and 24 degrees of freedom. Table values are,

$$\begin{aligned} F &= 1.94 \dots \text{ at } 5\% \\ &= 2.58 \dots \text{ at } 1\% \end{aligned}$$

Since the calculated value of F is less than the table values of F , H_0 is not rejected at 5% and 1% level of significance.

9.6.2 F test for testing equality of several means (Analysis of variance technique) ANOVA

The t test enables us to test the significance of the difference between two population means. If there are three or more means, then to test whether these have come from the same population or not, the F -test is used and the method is generally referred to as 'Analysis of Variance' technique.

It is a systematic procedure of splitting the total variation into a number of components, each associated with a possible source of variability. This is done with the objective of assessing the relative importance of different sources of variability.

If the data is classified according to one attribute, the resulting table is called one way table and the analysis of variance applied to this set of data is known as 'analysis of variance - one way classification' or 'one-way analysis of variance'. If the data is classified in the form of a two-way table, then 'analysis of variance - two way classification' is applied. Thus the form of analysis of variance depends upon the nature of investigation from which the data are collected.

One-way analysis of variance is discussed here to explain the basic principles of this technique.

Let there be k classes (samples) A_1, A_2, \dots, A_k drawn from normal populations with mean m_1, m_2, \dots, m_k respectively with common variance σ^2 . Let n_i denote the number of observations in the i th sample, such that $\sum n_i = n$, the total number of observations.

The following mathematical model is assumed for the analysis.

$$x_{ij} = m + a_i + e_{ij}$$

where x_{ij} denotes the j th observation in the i th class ($i = 1, 2, \dots, k, j = 1, 2, \dots, n_i$),

m is the general mean

a_i is the effect of i th class

e_{ij} are independently, normally distributed with mean zero and variance σ^2 .

The null hypothesis to be tested is,

$$H_0 : m_1 = m_2 = \dots = m_k$$

The values of x_{ij} differ among themselves due to

- (i) Variation from class to class
- (ii) Variation within classes

The analysis of variance splits the total variation $\sum (x_{ij} - \bar{x})^2$ into components due to each of the sources of variability mentioned above. Sum of squares (SS) and the number of degrees of freedom (df), are computed for each of the sources. Dividing the SS by the corresponding df, variances (mean squares) of the respective components are obtained.

$$\text{Mean square between classes} = \frac{\sum n_i (\bar{x}_i - \bar{x})^2}{k-1} \dots \dots \text{(III)}$$

where \bar{x}_i is a mean of the i th class and \bar{x} is the grand mean.

$$\text{Mean square within classes} = \frac{\sum (x_{ij} - \bar{x}_i)^2}{n-k} \dots \dots \text{(IV)}$$

(Error mean square)

When the null hypothesis is true, the variance estimates of the two components given in (III) and (IV) are estimates of the same quantity, the population variance. As the two estimates are independent of each other, they may not give the identical value, but they are expected not to differ significantly, when the null hypothesis is true. Thus testing the null hypothesis of equality of several means, is equivalent to testing the equality of these two variances. As discussed earlier (see 9.6.1), the equality of two variances can be tested by F ratio. Hence, the null hypothesis is tested by computing the following statistic :

$$F = \frac{\text{Mean square between classes}}{\text{Mean square within classes}} \sim (k-1) \text{ and } (n-k) \text{ d.f.}$$

Compare this computed value of F with the table value of F with (k-1) and (n-k) df at a desired level of significance. If F calculated > F table, H_0 is rejected. Rejection of hypothesis means that classes (samples) come from populations with different means.

The results of analysis of variance are usually summarised in the following table called 'Analysis of variance' (ANOVA) table :

Table 4 : Analysis of variance table

Source of variation	df	SS	MS	F
Between classes	k-1	S_1	$\frac{S_1}{k-1} = M_1$	
Within classes	n-k	S_2	$\frac{S_2}{n-k} = M_2$	$F = \frac{M_1}{M_2}$
Total	n-1			

The analysis of variance technique forms the basis of analysis of experimental designs discussed in Chapter No. 12. For worked out examples readers may refer to Chapter No. 12.

Chapter 10

CORRELATION AND REGRESSION

10.1 Introduction

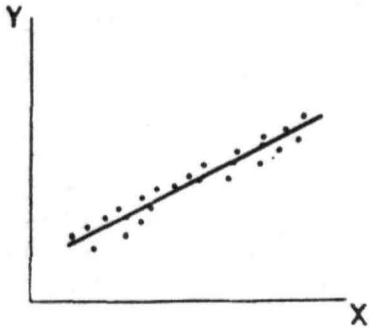
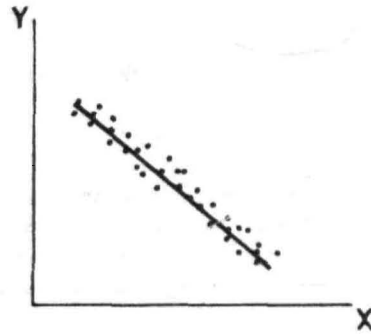
The statistical methods discussed so far are primarily intended to describe a single variable i.e., univariate populations. In this chapter the techniques that are useful in studying the relationships that exist when the data on two or more variables is available, are discussed.

If on the same individual, data on two variables say X and Y are listed, it is called a bivariate population. In this bivariate population, for every value of X, there is a corresponding value of Y. By treating these variables X and Y separately, measures of central tendency, dispersion etc., can be worked out. In addition to these measures it may be of interest to study the degree of relationship existing between the variables and the nature of their relationship. The study of the former aspect is referred to as 'correlation' and the latter as 'regression' analysis.

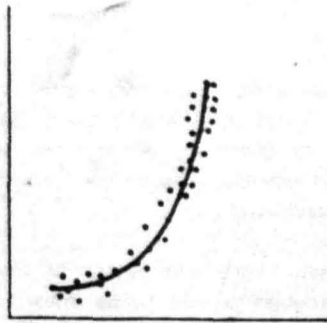
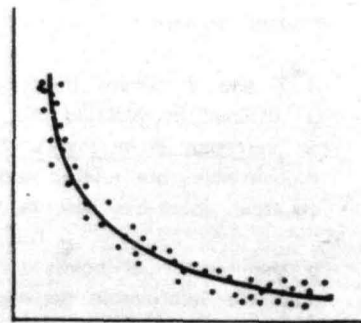
10.2 Scatter diagram

If X and Y denote the two variables under study, the scatter diagram is obtained by plotting the pairs of values of X and Y taking variables on cartesian co-ordinates. This diagram gives an indication of whether the variables are related and if so, the possible type of line or estimating equation which can describe the relationship.

If the scatter of points indicates that a line can better fit the data, then the relationship between the variables is said to be linear. Scatter diagrams in Fig. 1 and 2 are examples of linear relationship. In Fig. 1, X tends to increase as Y increases, the relationship between the variables is said to be direct and linear. In Fig. 2, X decreases as Y increases, the relationship between the variables is said to be inverse and linear.

**Fig. 1 : Direct linear****Fig. 2 : Inverse linear**

If the scatter of points indicates that a curve can better fit the data, then the relationship between the variables is said to be non-linear or curvilinear. Some curvilinear relationships are shown in Figures 3 and 4.

**Fig. 3 : Direct curvilinear****Fig. 4 : Inverse curvilinear**

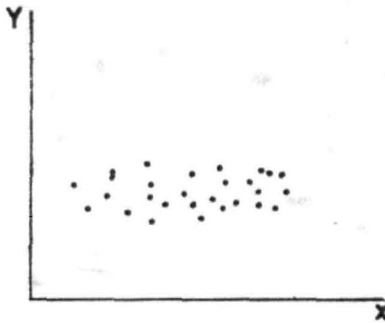


Fig. 5 : No relationship

If the scatter of points is as shown in Fig. 5, then there is little or no relationship between the variables.

10.3 Simple correlation

It is a statistical tool to study the degree of association or relationship existing between two variables, when the relationship is linear or approximately linear. The degree of relationship is quantified by a coefficient called the 'Karl Pearson's product moment correlation coefficient' or simply the 'correlation coefficient.' It is denoted by r . The working formula for r is given by

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right) \left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

In the above expression, X and Y denote the measurements on variables X and Y , n is the number of pairs of observations i.e. the sample size.

10.3.1 Properties of correlation coefficient

- (1) It is a pure number without units or dimensions.
- (2) It lies between -1 and 1 i.e., $-1 \leq r \leq 1$.
- (3) The correlation coefficient is independent of the origin and the scale of measurement of the variables.

Answer

	X	Y	XY	X ²	Y ²
1.	110	83	9130	12100	6889
2.	104	80	8320	10816	6400
3.	114	85	9690	12996	7225
4.	119	91	10829	14161	8281
5.	145	113	16385	21025	12769
6.	116	85	9860	13456	7225
7.	124	94	11656	15376	8836
8.	141	110	15510	19881	12100
9.	175	134	23450	30625	17956
10.	135	102	13770	18225	10404
11.	145	115	16675	21025	13225
12.	171	130	22230	29241	16900
13.	155	119	18445	24025	14161
14.	167	125	20875	27889	15625
15.	160	121	19360	25600	14641
Total	2081	1587	226185	296441	172637

$$\begin{aligned}
 r &= \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right) \left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right)}} \\
 &= \frac{226185 - \frac{(2081)(1587)}{15}}{\sqrt{\left(296441 - \frac{(2081)^2}{15}\right) \left(172637 - \frac{(1587)^2}{15}\right)}} \\
 &= \frac{6015.2}{\sqrt{(7736.94)(4732.4)}} = 0.9941
 \end{aligned}$$

10.3.2 Significance of the correlation coefficient

Let r be the observed correlation coefficient in a sample of n pairs of observations from a bivariate normal population. To test the hypothesis $H_0 : \rho = 0$, i.e. population correlation coefficient is zero, the following test procedure is used :

Compute :

$$t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}}$$

Which is distributed as t with $n-2$ df. If the calculated value of t is greater than the table value of t with $n-2$ df., at the desired level of significance, the correlation between the variables is significant. However, it is to be noted that significance of r is not an indication of the strength of relationship. It is simply a test to see whether ρ is equal to zero or not. The degree of the relationship between two variables can be measured by the square of the correlation coefficient r^2 (which is called the coefficient of determination). Unless r^2 is very high, one variable should not be used to forecast the other.

Example 2

The correlation between length and weight for a particular fish species is observed to be 0.7 from a sample of 18 specimens. Is it significant?

Answer

H_0 : Population correlation between length and weight is zero.

$$\begin{aligned} t &= \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.7 \sqrt{16}}{\sqrt{1-0.49}} = \frac{0.7 \sqrt{16}}{\sqrt{0.51}} \\ &= \frac{0.7 \times 4}{0.7141} = \frac{2.8}{0.7141} = 3.92 \end{aligned}$$

$$t_{16} (5\%) = 2.12, \quad t_{16} (1\%) = 2.92$$

Since the calculated value of t is greater than the table value of t at 5% and 1% level of significance, reject H_0 .

Hence, the correlation is highly significant.

Note : It is however, not necessary to carry out the " t " test described above for testing the significance of the correlation coefficient as ready made table of critical values of r for different degrees of freedom at 5% and 1% levels of significance is available (Fisher and Yates 1963). Compare the calculated value of r with the critical value of r from the table. If the calculated value of r is higher than or equal to the critical value, then correlation is significant.

10.4 ✓ Simple linear regression

If two variables are found to be highly correlated then a more useful approach would be to study the nature of their relationship. Regression analysis achieves this by formulating statistical models which can best describe these relationships. These models enable prediction of the value of one variable, called the dependent variable from the known values of the other variable(s). It differs from correlation in that regression estimates the nature of relationship where as the correlation coefficient estimates the degree or intensity of relationship.

[Simple linear regression deals with the study of linear relationships involving two variables, where as, the relationships among more than two variables are studied by the multiple regression techniques.]

10.4.1 Estimation of parameters a and b in the regression equation $Y = a+bX$

Scatter diagram gives some idea of the nature of relationship existing between the variables (see 10.2). If it indicates that the relationship is linear in nature, next step would be to develop a statistical model and proceed to estimate the underlying relationship. It is assumed that linear relationship of the form,

$$Y = a + bX + e \quad (I)$$

exists between the variables X and Y . In expression (I) e is a random variable (random error factor) assumed to be independently, randomly distributed with mean zero and variance σ^2 , 'a' and 'b' are constants (parameters). In this model it is assumed that each Y_i is normally distributed with mean $a + bX_i$ and constant variance σ^2 .

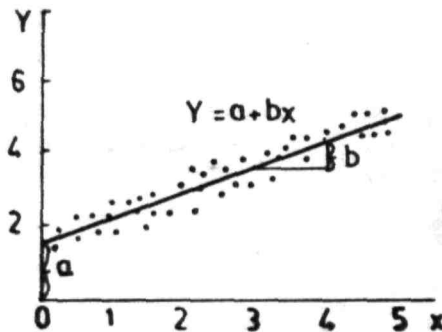


Fig. 6. Linear regression of Y on X

Fitting linear relationship of the form (I) is equivalent to estimating the constants a and b from the observed data. The best method that is used for estimation of 'a' and 'b' is the method of 'least squares'. In a popular way it only means that a line is found to which the total of squares of all distances from different points is minimum i.e. sum of e^2 is minimum. In other words search for the values of a and b which minimise,

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - a - bX_i)^2 \quad (II)$$

In the above expression n stands for the number of pairs of observations.

Estimates of parameters a and b which minimise (II) are obtained by the following formulae :

$$b = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

$$a = \bar{Y} - b\bar{X}$$

Estimated values of these constants are substituted in the equation $Y = a + bX$ to get the regression equation. From this equation the value of Y can be estimated for a given value of X .

10.4.2 Special names of the parameters

There are special names for the parameters 'a' and 'b'. The parameter 'a' is called the Y intercept. It is the value Y assumes when $X = 0$ (Fig. 6). The parameter 'b' is called the regression coefficient and gives the slope of the regression line, i.e., it shows how steep the line is. The regression coefficient indicates the rate of change in the dependent variable per unit change in the independent variable.

10.4.3 Variance about the regression line (deviations from regression)

The assumption behind the standard linear regression is that each Y_i is normally distributed with mean value $a + bX_i$ and with a constant variance σ^2 which is not dependent on the value of X_i . The formula for estimate of this variance is given by

$$S^2 = \frac{1}{n-2} \Sigma (Y_i - a - bX_i)^2$$

This forms the basis for an estimate of error in fitting the line. However, convenient formula to work out this variance is given by,

$$s^2 = \frac{1}{n-2} \left[\left(\sum Y^2 - \frac{(\sum Y)^2}{n} \right) - \frac{\left(\sum XY - \frac{(\sum X)(\sum Y)}{n} \right)^2}{\sum X^2 - \frac{(\sum X)^2}{n}} \right]$$

$$= \frac{1}{n-2} \left(\sum y^2 - \frac{(\sum xy)^2}{\sum x^2} \right)$$

where $\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$ is corrected sum of squares of Y

$\sum xy = \sum XY - \frac{(\sum X)(\sum Y)}{n}$ is corrected sum of cross products of X and Y.

$\sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n}$ is corrected sum of squares of X.

10.4.4 Two regression lines

If two variables X and Y are open to choice as to which affects which then 2 regression lines may be conceived. They are,

(i) Regression equation of Y on X

If Y is considered as dependent variable, then the regression equation of Y on X is given by

$$Y = a + bx$$

The regression coefficient b is called the regression coefficient of Y on X and is usually denoted by b_{yx} . In this equation a and b are so estimated as to minimise the residual variation (deviations from regression) of Y i.e. $\sum (Y_i - a - bX_i)^2$ is minimised.

(ii) Regression equation of X on Y

If X is considered as dependent variable then the regression equation is given by

$$X = a + bY$$

The regression coefficient ~~b~~ is called the regression coefficient of X on Y and is usually denoted by b_{yx} . In this equation a and b are so estimated as to minimise the residual variation of X i.e. $\sum (X_i - a - bY_i)^2$ is minimised. The values of 'a' and 'b' obtained in (i) and (ii) will usually be different.

10.4.5 Properties of regression lines

- (i) The regression lines intersect at point (\bar{X}, \bar{Y}) .
- (ii) If the variables are perfectly correlated, the regression lines co-incide.
- (iii) If the variables are not correlated the regression lines are perpendicular to each other.

10.4.6 Relation between correlation and regression coefficients

If b_{yx} is the regression coefficient in the regression equation of Y on X and b_{xy} is the regression coefficient in the regression equation of X on Y, then the correlation coefficient r is the square root of the product of b_{yx} and b_{xy} .

i.e.,
$$r = \sqrt{b_{yx} \cdot b_{xy}}$$

10.4.7 Test of significance of linearity of regression (significance of regression coefficient)

The significance of the linearity of regression is tested by,

- (i) the method of analysis of variance
or by
- (ii) t - test

Null hypothesis to be tested is

$H_0 : \beta = \alpha$, where β denotes the population regression coefficient.

(i) **Analysis of variance method**

In this method, the total sum of squares of dependent variable Y is split into two components. One due to regression and another due to errors of random sampling called deviations from regression or residual.

Thus,

Total sum of squares = Regression sum of squares + Residual sum of squares

If the regression is based on n observations, the total sum of squares has n-1 degrees of freedom (df), whereas regression sum of squares has 1 df and residual sum of squares n-2 df. Dividing the sum of squares (ss) by the corresponding df, respective mean squares (ms) are obtained. This information can be set down in the form of analysis of variance table as given below:

Source	df	ss	ms
Regression	1	$b \sum xy = S_1$	$\frac{S_1}{1} = m_1$
Residual (Deviations from regression)	n-2	$\sum y^2 - b(\sum xy) = S_2$	$\frac{S_2}{n-2} = m_2$
Total	n-1	$\sum y^2$	

$$\text{In the above table } \sum xy = \frac{\sum XY (\sum X)(\sum Y)}{n}$$

$$\sum y^2 = \frac{\sum Y^2 - (\sum Y)^2}{n}$$

The residual mean square indicates the variation not accounted by the linear regression and therefore measures the uncontrolled

variation that affects Y values. Significance of regression coefficients is tested by comparing mean square due to regression with residual mean square, using F ratio.

$$F = \frac{\text{ms due to regression}}{\text{residual ms}} = \frac{m_1}{m_2}$$

This is distributed as F with 1, n-2 df. If the calculated value of F is more than the table value of F at the desired level of significance, it is concluded that regression is statistically significant.

(ii) **t - test**

Alternatively, test of linearity of regression can be carried out with the procedure outlined below.

Compute ,

$$t = \frac{b - \beta}{s_b} = \frac{b - 0}{s_b} = \frac{b}{s_b}$$

which is distributed as t with n-2 df. In the above expression, s_b is the standard deviation of regression coefficient and is the square root of

$$s_b^2 = \frac{\text{Residual ms}}{\sum (X - \bar{X})^2} = \frac{[\sum y^2 - (\sum xy)^2 / (\sum x^2)] / (n-2)}{\sum x^2}$$

Where $\sum y^2$, $\sum x^2$ are corrected sum of squares of y and x and $\sum xy$ is corrected sum of products of X and Y. (See 10.4.3).

If the calculated value of $|t|$ is more than the table value of t at the desired level of significance, the null hypothesis is rejected.

Example 3

The data on fish yield tested under 5 stocking densities are given below:

S.No.	1	2	3	4	5
Fingerlings (^{'000} /ha)	2	3	4	5	6
Fish yield (t/ha)	2.5	3.6	4.4	5.0	5.4

- Compute :**
- the regression equation of fish yield on stocking density.
 - Estimate fish yield for stocking density of 4,200/ha.
 - Mean square due to deviations from regression.
 - Test whether the regression coefficient is significant.

Answer

	Fingerlings (^{'000} /ha) X	Fish yield (t/ha) Y	XY	X ²	Y ²
(1)	2	2.5	5.0	4	6.25
(2)	3	3.6	10.8	9	12.96
(3)	4	4.4	17.6	16	19.36
(4)	5	5.0	25.0	25	25.10
(5)	6	5.4	32.4	36	29.16
Total	20	20.9	90.8	90	92.73

$$b = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{90.8 - \frac{(20)(20.9)}{5}}{90 - \frac{(20)^2}{5}}$$

$$= \frac{7.2}{10} = 0.72$$

$$a = \bar{Y} - b\bar{X} = 4.18 - (0.72)4 = 1.3$$

$$\text{Hence, } Y = 1.3 + 0.72 X$$

- (ii) To obtain the estimate of fish yield when the stocking density is 4,200, put $X = 4.2$ in the above equation.

$$\begin{aligned} \text{i.e., } Y &= 1.3 + 0.72 (4.2) \\ &= 4.32 \text{ tons} \end{aligned}$$

- (iii) Mean square due to deviations from regression is given by

$$\begin{aligned} S^2 &= \frac{1}{n-2} \left(\sum y^2 - \frac{(\sum xy)^2}{\sum x^2} \right) \\ &= \frac{1}{3} \left(5.37 - \frac{(7.2)^2}{10} \right) = \frac{1}{3} (0.19) = 0.06 \end{aligned}$$

- (iv) The null hypothesis to be tested is

$$H_0 : \beta = 0$$

The test statistic used is

$$\begin{aligned} t &= \frac{b}{\sqrt{\frac{S^2}{\sum x^2}}} = \frac{0.72}{\sqrt{\frac{0.06}{10}}} = \frac{0.72}{\sqrt{0.006}} \\ &= \frac{0.72}{0.077} = 9.35 \end{aligned}$$

Table value of t at 5% level with 3 df is 3.18. Since computed t value of 9.35 is greater than the table value, H_0 is rejected.

10.4.8 Homogeneity of two regression coefficients

$$\text{Let } Y = a_1 + b_1 X \dots \dots (I)$$

$$\text{and } Y = a_2 + b_2 X \dots \dots (II)$$

be two regression lines of Y on X for the same set of characters X and Y for say male (set I) and female (set II) group of fishes. To test the null hypothesis that population regression coefficients of the two regression lines are the same i.e. $H_0 : \beta_1 = \beta_2$, the following test statistic is used

$$t = \frac{|b_1 - b_2|}{\sqrt{S_p^2 \left(\frac{1}{\sum x_1^2} + \frac{1}{\sum x_2^2} \right)}} \quad (\text{III})$$

This is distributed as t with $n_1 + n_2 - 4$ degrees of freedom. $\sum x_1^2$, $\sum x_2^2$ are corrected sums of squares as defined in 10.4.3 for 1st and 2nd set of data.

If S_1^2 and S_2^2 are variances about the regression line (deviations from regression) of the 1st and 2nd set of data based on n_1 and n_2 observations respectively, then

$$S_p^2 = \frac{(n_1 - 2) S_1^2 + (n_2 - 2) S_2^2}{n_1 + n_2 - 4}$$

Example 4

The relationship between the standard length (X) and body depth (Y) was studied by linear regression for male and female fishes. The following data were obtained.

	Sample size	Corrected sum of squares and products				
		x^2	y^2	xy	S^2	b
Male	12	64	7.5	21	0.061	0.33
Female	15	68	9.5	24	0.080	0.35

Test the hypothesis that population regression coefficients of both the regression lines are the same.

Answer

The null hypothesis to be tested is,

$$H_0 : \beta_1 = \beta_2$$

To test this hypothesis, the following test statistic is used :

$$t = \frac{|b_1 - b_2|}{\sqrt{S_p^2 \left(\frac{1}{\Sigma x_1^2} + \frac{1}{\Sigma x_2^2} \right)}}$$

It is distributed as t with $n_1 + n_2 - 4$ df

Compute,

$$S_p^2 = \frac{(n_1 - 2) S_1^2 + (n_2 - 2) S_2^2}{n_1 + n_2 - 4} = \frac{10 \times 0.061 + 13 \times 0.080}{12 + 15 - 4} = 0.0717$$

$$\Sigma x_1^2 = 64, \quad \Sigma x_2^2 = 68, \quad b_1 = 0.33, \quad b_2 = 0.35$$

Substituting these values, t is computed as :

$$\begin{aligned} t &= \frac{|0.33 - 0.35|}{\sqrt{0.0717 \left(\frac{1}{64} + \frac{1}{68} \right)}} = \frac{0.02}{\sqrt{0.00217}} \\ &= \frac{0.02}{0.0466} = 0.4292 \end{aligned}$$

The table value of t with 23 df is 2.069 at 5% and 2.807 at 1% level of significance. Since $|t| <$ the table value of t at 5% and 1% level of significance, the null hypothesis is not rejected.

10.4.9 Linearizing transformation

Both the correlation and regression techniques discussed earlier are based on the assumption that linear relationship exists between the variables. However, there are many cases in fisheries investigations, where relation-

ship between the variables is not linear i.e. non-linear. Some of these non-linear relationships can be brought to linear form, using certain transformations. An example of length-weight relationship which can be transformed to the linear form is discussed here.

10.4.9.1 Length-weight relationship

The relationship between body weight (W) and body length (L) in fishes has been empirically observed to be of the form

$$W = aL^b \quad (IV)$$

This equation is not in linear form. The parameters 'a' and 'b' are almost universally estimated by fishery workers by transforming the equation (IV) to logarithmic form and applying the least squares technique. Thus the equation actually used is,

$$\log W = \log a + b \log L$$

The above method tacitly assumes the following multiplicative error model :

$$W = a.L^b e \quad (V)$$

where a and b are constants and e is a random error factor.

Taking logarithm on both sides of (V) gives rise to

$$\begin{aligned} \log W &= \log a + b \log L + \log e \\ \text{i.e. } Y &= A + BX + E \quad (VI) \end{aligned}$$

where $Y = \log W$, $X = \log L$, $A = \log a$, $b = B$, $E = \log e$

Expression (VI) is in the linear form. If it is assumed that E is distributed normally with mean zero and variance σ^2 , then the estimate of A and B can be obtained by the method of least squares discussed earlier in section 10.4.1, using the formula

$$B = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

$$A = \bar{Y} - B\bar{X}$$

In the above expressions $Y = \log W$ and $X = \log L$, \bar{Y} and \bar{X} denote means of Y and X values respectively. The B value gives an estimate of b , whereas conventionally, 'a' is estimated as e^A or $\exp(A)$. This method however gives biased estimate of 'a'. To compensate for the bias the 'a' value obtained is multiplied by the following correction factor :

$$\text{Correction factor} = e^{S^2/2} = \exp\left(\frac{S^2}{2}\right)$$

Where S^2 is an estimate of variance of deviations from regression (see 10.4.3). Hence, corrected $a = \exp\left(A + \frac{S^2}{2}\right)$.

If common logarithms are used, $a = \text{Antilog}\left(A + \frac{S^2}{2}\right)$

10.4.9.2 Applications of length-weight relationship

- (i) **It is useful in estimating weight of fish for a given length.**

As length of fish can be measured more easily and accurately than weight in landing centres as well as on board the vessels in the sea, it is convenient to estimate weight from predetermined length-weight relationships.

- (ii) **It is useful in determining condition factor**

In order to compare weight and length in a particular sample or individual, condition factors are employed. Fulton's condition factor (K) is calculated as,

$$K = \frac{W}{L^3}$$

where W and L are the observed total weight and length of a fish. It is the value of 'a' in length-weight relationship, $W = aL^b$, when $b = 3$. If the fish is heavier, at a given length, the larger

is the factor K , implying better is the condition of fish. K greater than 1, indicates general well being of the fish is good. Fulton's condition factor, is suitable for comparing differences related to sex, season or place of capture. Even when b differs from 3, Fulton's condition factor may be used, if fish are approximately of the same length. If the length range is large, the following formula is used :

$$K^I = \frac{W}{L^b}$$

Alternatively, the condition factor is computed as the ratio of observed weight to estimated weight.

$$K^{II} = \frac{W}{\hat{W}}$$

where \hat{W} is the estimated weight based on length-weight relationship $W = aL^b$.

Example 5

Total length (cm) and weight (gm) recorded on a sample of 12 fish are given below :

S.No.	1	2	3	4	5	6	7	8	9	10	11	12
Length (cm)	17.6	19	27.2	20.2	18.6	14.8	21.1	16.8	21.4	13.2	23.7	24.6
Weight (g)	25	32	110	42	30	10	45	20	48	8	75	82

(i) Fit length-weight relationship of the type $W = aL^b$, where W is weight and L is length of fish.

(ii) Test whether b differs significantly from 3.

Answer

S.No.	Length (cm) (L)	Weight (g) (W)	X=log L	Y=log W	XY	X ²
1.	17.6	25	1.2455	1.3979	1.7411	1.5513
2.	19	32	1.2786	1.5052	1.9248	1.6353
3.	27.2	110	1.4346	2.0414	2.5221	1.5265
4.	20.2	42	1.3054	1.6232	2.1189	1.7041
5.	18.6	30	1.2695	1.4771	1.8752	1.6116
6.	14.8	10	1.1703	1.0000	1.1703	1.3696
7.	21.1	45	1.3243	1.6532	2.1893	1.7538
8.	16.8	20	1.2253	1.3010	1.5941	1.5014
9.	21.4	48	1.3304	1.6812	2.2367	1.7700
10.	13.2	8	1.1206	0.9031	1.0120	1.2557
11.	23.7	75	1.3747	1.8751	2.5777	1.8898
12.	24.6	82	1.3909	1.9138	2.6619	1.9346
			15.4701	18.3722	24.0304	20.0341

(i) Length-weight relationship

$$\bar{X} = 1.2892 \quad \bar{Y} = 1.5310$$

$$\begin{aligned}
 b &= \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} \\
 &= \frac{24.0304 - \frac{(15.4701)(18.3722)}{12}}{20.0347 - \frac{(15.4701)^2}{12}} \\
 &= \frac{24.0304 - 23.6850}{20.037 - 19.9437} = \frac{0.3454}{0.0910} = 3.7956
 \end{aligned}$$

$$\begin{aligned}
 A &= \bar{Y} - b\bar{X} \\
 &= 1.5310 - 4.8933 = -3.3623
 \end{aligned}$$

Deviations from regression

$$S^2 = \frac{1}{n-2} \left[\left(\sum Y^2 - \frac{(\sum Y)^2}{n} \right) - \frac{(\sum XY - (\sum X) \cdot (\sum Y)/n)^2}{\sum X^2 - \frac{(\sum X)^2}{n}} \right]$$

$$= \frac{1}{10} \left(1.3219 - \frac{0.1193}{0.0910} \right) = 0.0011$$

$$\begin{aligned}
 a &= \text{Antilog} \left(\frac{A + S^2}{2} \right) \\
 &= \text{Antilog} \left(\frac{-3.3623 + 0.0011}{2} \right) = \text{Antilog} (-3.3568) \\
 &= 0.0004.
 \end{aligned}$$

The length-weight relationship is therefore given by,

$$W = 0.0004 L^{3.7956}$$

- (ii) To test whether the sample regression coefficient ($b = 3.7956$) comes from a population with the regression coefficient $\beta = 3$, the following null hypothesis is set up: $H_0: \beta = 3$

The test statistic used is,

$$t = \frac{|b - 3|}{s_b}$$

This is distributed as t with $n-2$ df.

$$s_b^2 = \frac{\text{Deviation from regression}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

$$= \frac{0.0011}{0.0910} = 0.0121$$

$$\text{Hence } s_b = \sqrt{0.0121} = 0.11$$

$$\text{Hence } t = \frac{|3.7956 - 3|}{0.11} = \frac{0.7956}{0.11} = 7.23$$

The table values of t at 5% and 1% level of significance are 2.228 and 3.169 respectively. As t calculated is 7.23, which is more than the table values of t at both 5% and 1% level of significance, the null hypothesis is rejected.

10.4.10 Applications of linear regression analysis (Ricker, 1973)

Some situations wherein linear regression approach is of value in fisheries research are given below :

1. Conversion between different length measurements, i.e., from, say, total length to standard length or to fork length and so on.
2. Calculation of fish lengths from scale measurements.
3. Fitting weight-length relationship.
4. Estimating parameters of the von Bertalanffy growth curve.
5. Splitting of total mortality coefficient Z in to M , the natural mortality coefficient and F the fishing mortality coefficient using the regression of Z on effort f .
6. Catch curve method of estimating instantaneous total mortality rate making use of the relation between abundance (number) of fish and their corresponding age.

7. Schaefer's method of population analysis based on the regression of catch per unit of effort on effort.
8. Relation of fecundity to body weight using the regression of number of eggs (F) on body weight (W).
9. Routine metabolic rate of fishes through the regression of oxygen consumption quiescent of fishes (Q) on body weight W .

Chapter 11

SAMPLING METHODS

11.1 Introduction

In the earlier discussion on collection of data, (chapter, 2), the advantages of collecting required information on a sample, a part of the population were discussed. The information obtained from a sample is then used to describe or estimate certain characteristics or parameters of the whole population. Drawing inferences about a population based on a sample is an age old practice, though the scientific approach to the problem is of recent origin. A consumer at a provision store inspects a handful of rice in order to form a conclusion about the quality of rice in the whole bag or the housewife tastes a spoonful of soup to draw conclusions regarding the whole quantity in the kettle. These are some examples of uses of sampling procedures in every day life. If uniformity exists among the units (individuals) of population, then any sample chosen will give almost the same result. If there is a great variation among units of the population then a proper method of selecting the sample has to be used to draw reliable conclusions about the population. Modern statistical sampling methods such as the 'probability random sampling methods' give a definite procedure for selecting a sample from a population. A distinct advantage of random sampling procedure is its ability to provide an estimate of the sampling error based on the sample itself, which forms the basis for ascertaining the reliability of the estimate. There are different methods of random sampling, and the important ones are described below.

11.2 ✓ Simple random sampling

In this method a sample is drawn unit by unit with equal probability of selection for every unit at each draw. Every possible sample of required size has the same chance of being chosen in this method. Simple random sampling can be selected using either the lottery method or using random number tables which was discussed in chapter No. 2. Selection through random number tables is convenient when the population under study is large.

11.2.1 Estimation of population parameters

Once the sample has been drawn, the next step is to estimate the population parameters based on the sample observations. The main parameters of interest are population mean and total.

Let us suppose that a population of size N is being sampled for some characteristic, say, X . Let X_1, X_2, \dots, X_N be the values of the character on N units of the population.

Further, suppose that a sample of n individuals is selected by simple random sampling with values x_1, x_2, \dots, x_n . The unbiased estimate of the population mean is given by,

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

The variance of sample mean is estimated using the formula

$$\begin{aligned} \text{Var}(\bar{x}) &= \left(\frac{1}{n} - \frac{1}{N} \right) S^2 \\ &= \frac{(N-n)}{N} \frac{S^2}{n} = \left(1 - \frac{n}{N} \right) \frac{S^2}{n} \\ &= \frac{S^2}{n}, \text{ when } N \text{ is large as compared to } n \end{aligned}$$

In the above expression,

$$S^2 = \frac{1}{n-1} \left(\sum (x_i - \bar{x})^2 \right) \text{ is the sample variance which is the unbiased estimate of population variance. It is clear that the variance of } \bar{x} \text{ depends upon sample size and variability present in the population.}$$

$$\text{Std error of Mean} = \sqrt{\text{Var}(\bar{x})}$$

The estimate of the population total is : $X = N\bar{x}$

$$\text{Var}(X) = N^2 \text{Var}(\bar{x})$$

Example 1

Fishes are landed at a certain small landing centre throughout the year. Twenty days were randomly selected from 365 days of a year and fish landed (in tonnes) during these days were recorded which are given below.

Weight in kg : 30, 42, 25, 32, 48, 32, 40, 28, 30, 20
18, 31, 15, 28, 25, 30, 35, 40, 22, 20

Estimate (i) the average fish landings per day, and (ii) total fish landings during the year.

Answer

(i) **Mean fish landings per day**

$$= \frac{30 + 42 + \dots + 22 + 20}{20} = \frac{591}{20} = 29.55$$

$$S^2 = \frac{1}{n-1} \left(\sum x^2 - \frac{(\sum x)^2}{n} \right)$$

$$= \frac{1}{19} (18853 - 17464.05)$$

$$= \frac{1388.95}{19} = 73.1026$$

$$\text{Var}(\bar{x}) = \frac{S^2}{n} \left(1 - \frac{n}{N} \right)$$

$$= \frac{73.1026}{20} \left(1 - \frac{20}{365} \right) = 3.6551 (0.9452)$$

$$= 3.4548$$

Standard error of mean = 1.8587
(S.E.)

$$\begin{aligned}
 \text{(ii) Estimate of total fish landing } X &= N\bar{x} \\
 &= 365 (29.55) \\
 &= 10785.75
 \end{aligned}$$

$$\begin{aligned}
 \text{Standard error of total fish landings} &= N \times (\text{SE of mean}) \\
 &= 365 \times 1.8587 \\
 &= 678.4255
 \end{aligned}$$

11.3 Stratified random sampling

As has been mentioned in simple random sampling that the variance of sample mean depends on the size of the sample and the variability of the population. Therefore, the only way of increasing the precision of an estimate apart from the size of the sample is to devise sampling procedures which will effectively reduce the variability. One such procedure known as 'stratified sampling' consists of dividing the population into 'classes' or 'strata', each relatively homogenous and drawing random samples of known sizes, one each from different strata. Then the estimates are made for each of the strata and combined by a proper weightage to obtain the estimate for the whole population. The variance of this estimate is obtained by combining the variances of the estimates within each stratum. This combined estimate of variance will be small as, within stratum variances will tend to be small as each of them come from relatively homogenous stratum.

Stratified random sampling method assumes that the structure of the population necessary to demarcate the strata is known, which may not be true in many situations. For better planning and execution of the survey work, stratification is some times done based on geographical proximity.

11.3.1 Estimation of mean and variance

If the population under study is divided into k strata with sizes N_1, N_2, \dots, N_k respectively, then estimate of the population mean is given by

$$\bar{x} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2 + \dots + N_k \bar{x}_k}{N_1 + N_2 + \dots + N_k} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2 + \dots + N_k \bar{x}_k}{N}$$

$\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ are means of 1st, 2nd ... kth stratum

$\bar{x}_i = \frac{1}{n_i} \cdot (\sum x_i)$ where n_i is the i th stratum sample size.

$$\text{Var}(\bar{x}) = \frac{N_1^2 \text{Var}(\bar{x}_1) + N_2^2 \text{Var}(\bar{x}_2) + \dots + N_k^2 \text{Var}(\bar{x}_k)}{N^2}$$

The estimate of population total is given by

$$X = N\bar{x}$$

and its variance is

$$\text{Var}(X) = N^2 \text{Var}(\bar{x})$$

11.3.2 Allocation of sample size

The advantage of stratified sampling can be increased by allocating the sample size n into various strata in the best proportion. Following are the important methods of allocation of samples to different strata.

(i) Equal samples for each stratum

The total sample size n is divided equally among all the strata, i.e., for the i th stratum, $n_i = \frac{n}{k}$ if there are k strata.

(ii) Proportional allocation

In this method the total sample size n is allocated to different strata in proportion to their size, i.e., for the i th stratum,

$$n_i = n \frac{N_i}{N}$$

(iii) Optimum allocation

In some cases it may be required to conduct a sample survey with a fixed budget, but with varying costs of selecting the

sampling units from different strata. In such situations the allocation of the number of units to each stratum is done with a view to minimizing the sampling variance for the given fixed cost or the cost of survey is minimum for the specified value of the sampling variance. Allocation of the sample for the i th stratum is given by,

$$n_i = n \frac{N_i S_i / \sqrt{c_i}}{\sum N_i S_i / \sqrt{c_i}}$$

where N_i is the size of the i th stratum, S_i is the population standard deviation for the i th stratum, c_i is the cost of obtaining a single observation from the i th stratum.

(iv) **Neyman allocation**

This method of allocation is used when the cost of selecting the sampling unit does not differ from stratum to stratum. The sample size for the i th stratum is given by,

$$n_i = \frac{N_i S_i}{\sum N_i S_i} n$$

This indicates that the total sample size n is allocated in proportion to $N_i S_i$, that is, take more individuals from the strata that are large and are highly variable.

Some times it may so happen that the sample size n_i estimated by Neyman allocation may be larger than the corresponding N_i of the i th stratum. Such a type of situation arises only when overall sampling fraction ($\frac{n}{N}$) is substantial, and one stratum is much variable than the others. The procedure that can be adopted in such situations is outlined in Cochran (1963).

11.3.3 Advantages of stratified random sampling

There are many advantages with stratified random sampling and the important ones are given below :

- (i) Stratified random sampling gives better cross section of the population than simple random sampling.
- (ii) Precision of the estimated character is likely to be higher in stratified random sampling than in simple random sampling.
- (iii) For physical or administrative reasons it is easier to collect the data using this sampling technique.

Example 2

In a certain state there are 250 fishing villages. It is known that in 120 villages the fish landings are below 50 kg, in 80 villages the landings are between 50 and 60 kg and in 50 villages the landings are above 60 kg per boat. A random sample of 12, 8 and 5 villages was drawn respectively from the 3 groups of villages and the catches in kg per boat recorded in these selected villages are given below. Estimate the catch in kg per fishing boat with 95% confidence limits.

Stratum (group) I : 25, 20, 38, 35, 39, 43, -26, 29, 42, 23, 32, 36
 Stratum (group) II : 51, 52, 58, 54, 51, 55, 59, 57
 Stratum (group) III : 72, 68, 61, 66, 79

Answer

For stratum I : $N_1 = 120, n_1 = 12$

$$\text{Mean, } \bar{x}_1 = \frac{\sum x}{n_1} = \frac{388}{12} = 32.33$$

$$S_1^2 = \frac{1}{(n_1 - 1)} \left(\sum x^2 - \frac{(\sum x)^2}{n_1} \right)$$

$$= \frac{1}{11} (13194 - 12545.333) = 58.9697$$

$$\text{Var}(\bar{x}_1) = \left(\frac{1}{n_1} - \frac{1}{N_1} \right) S_1^2 = \left(\frac{1}{12} - \frac{1}{120} \right) 58.9697$$

$$= (0.0833 - 0.0083) 58.9697 = (0.075) 58.9697 = 4.3931$$

Stratum II : $N_2 = 80, n_2 = 8$

$$\text{Mean, } \bar{x}_2 = \frac{\sum x}{n_2} = \frac{437}{8} = 54.625$$

$$\begin{aligned} S_2^2 &= \frac{1}{n_2 - 1} \left(\sum x^2 - \frac{(\sum x)^2}{n_2} \right) \\ &= \frac{1}{7} (23941 - 23871.125) = \frac{69.875}{7} = 9.982 \end{aligned}$$

$$\begin{aligned} \text{Var}(\bar{x}_2) &= \left(\frac{1}{n_2} - \frac{1}{N_2} \right) S_2^2 \\ &= \left(\frac{1}{8} - \frac{1}{80} \right) 9.982 = (0.125 - 0.0125) 9.982 \\ &= (0.1125) (9.982) = 1.123 \end{aligned}$$

Stratum III : $N_3 = 50, n_3 = 5$

$$\text{Mean, } \bar{x}_3 = \frac{\sum x}{n_3} = \frac{346}{5} = 69.2$$

$$\begin{aligned} S_3^2 &= \frac{1}{n_3 - 1} \left(\sum x^2 - \frac{(\sum x)^2}{n_3} \right) \\ &= \frac{1}{4} (24126 - 23943.2) = \frac{182.8}{4} = 45.7 \end{aligned}$$

$$\begin{aligned} \text{Var}(\bar{x}_3) &= \left(\frac{1}{n_3} - \frac{1}{N_3} \right) S_3^2 \\ &= \left(\frac{1}{5} - \frac{1}{50} \right) 45.7 = (0.2 - 0.02) 45.7 \\ &= (0.18) (45.7) = 8.226 \end{aligned}$$

$$\begin{aligned}\bar{x} &= \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2 + N_3 \bar{x}_3}{N} \\ &= \frac{11809.6}{250} = 46.8384 \\ \text{Var}(\bar{x}) &= \frac{N_1^2 \text{Var}(\bar{x}_1) + N_2^2 \text{Var}(\bar{x}_2) + N_3^2 \text{Var}(\bar{x}_3)}{N^2} \\ &= \frac{14400 \times 4.3937 + 6400 \times 1.123 + 2500 \times 8.226}{62500} = 1.4562\end{aligned}$$

$$\begin{aligned}\text{Standard error} &= \sqrt{\text{Var} \bar{x}} = 1.2067 \\ (\text{S.E.})\end{aligned}$$

$$\begin{aligned}\text{Estimated total} &= N_1 \bar{x}_1 + N_2 \bar{x}_2 + N_3 \bar{x}_3 \\ &= 11709.6\end{aligned}$$

The 95% confidence limits for mean are,

$$\begin{aligned}&\bar{X} \pm 1.96 \text{ S.E. of mean} \\ &= 46.8384 \pm 1.96 \times 1.2067 \\ &= 46.8384 \pm 2.3651\end{aligned}$$

The confidence interval is (44.4733, 49.2035)

11.4 Systematic sample

The method of sampling in which only the first unit is selected at random and the rest being selected according to a predetermined pattern is known as 'systematic sampling'. Suppose that a population consists of N units, serially numbered from 1 to N and a sample of n is desired to be drawn from it. Further, let it be possible to express N as $N = nk$, k being an integer (i.e., N is an integral multiple of k). This $k = \frac{N}{n}$, is called the sampling interval and the sample drawn with this interval is called a 1 in k sample. Drawing a systematic sample consists of selecting a unit at random from the first k units and then selecting every k^{th} unit in the population thereafter. For example, suppose we have a population of size $N = 60$, and it is required to draw a sample of size $n = 5$. In this case $k = \frac{N}{n} = \frac{60}{5} = 12$. Hence, this is 1 in 12 sample.

Therefore, draw a unit at random from the first 12 units. Suppose this turns out to be the 8th unit, then the sample consists of the 8th unit and every 12th unit thereafter, i.e., units with serial number 8, 20, 32, 44 and 56, constitute a sample.

When $N = nk$, the sample mean of a systematic sample selected with the above procedure provides an unbiased estimate of the population mean. However, when $N \neq nk$ the mean of a sample selected with the above procedure, gives a biased estimate of the population mean. If $n > 50$ the bias is negligible and can be ignored. An alternative method of drawing a sample when $N \neq nk$, which gives unbiased estimate of the population mean, is also available. For details readers may refer to Cochran (1963).

11.4.1 Estimation of population mean and population total

When $N = nk$, as mentioned earlier, sample mean provides an unbiased estimate of the population mean. Sample mean is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

An approximate and biased estimate of variance of \bar{X} is given by

$$\begin{aligned} \text{Var}(\bar{x}) &= \frac{N-n}{2Nn(n-1)} \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2 \\ &= \frac{(k-1)}{2nk(n-1)} \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2 \end{aligned}$$

The estimate of population total is given by

$$X = N\bar{x}$$

and its variance estimate is

$$\text{Var}(X) = \frac{N^2(k-1)}{2nk(n-1)} \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2$$

If the arrangement of the units in the population is random, then

$$\text{Var}(\bar{x}) = \frac{S^2}{n} \left(\frac{1-n}{N} \right)$$

11.4.2 Advantages

- (i) Easy to draw samples and requires less time. Hence it is operationally convenient to execute the survey.
- (ii) Sample is spread out more evenly over the population.
- (iii) It is likely to be more precise than simple random sampling for many populations.

This sampling procedure has been successfully employed in forest surveys, in fish catch estimation surveys, in milk yield surveys, etc. Systematic sampling is widely used with other sampling methods such as stratified random sampling, cluster sampling, etc. For instance, in estimating the total marine fish catches of our country, systematic sampling is used with stratified multistage random sampling.

11.4.3 Disadvantages

- (i) There is no reliable method for estimating the variance of mean of a systematic sample from the sample data.
- (ii) Systematic sampling has to be used carefully when there is a periodicity in the population, as its efficiency depends upon the choice of the sampling interval.

Example 3

At a particular landing centre, data on fish landings were recorded during a month (of 30 days) on 10 days selected by systematic sample with a sampling interval of 3 days. Below are given the landings (in tons) of the 10 sampled days.

Day	1	2	3	4	5	6	7	8	9	10	Total
Landing(t)	3.5	4.5	3.2	3	4.2	4	3.5	5	4	3.3	38.2

Give an estimate of the total fish landings of the centre for the month and also an approximate standard error.

Answer

Given $N = 30$, $k = 3$, $n = 10$. An estimate of the total landings is given by

$$X = N\bar{x} = \frac{N}{n} \sum x_i = \frac{30}{10} (38.2) = 114.60$$

$$\text{Var } V(X) = \frac{N^2 (k-1)}{2nk(n-1)} \sum_{i=1}^{n-1} (\bar{x}_{i+1} - \bar{x}_i)^2 = 27.3333$$

$$\text{Standard error of the estimate } X = \sqrt{\text{Var}(X)} = \sqrt{27.3333} = 5.23$$

11.5 Cluster sampling

If the basic sampling unit in a population is to be found in groups or clusters, then from the operational point of view the sampling may be carried out by selecting a sample of clusters and observing all the units of each selected cluster. This type of sampling is known as cluster sampling.

It is less costly than simple random sampling. However, cluster sampling is less efficient than simple random sampling due to the fact that individual units within a cluster tend to be similar. Efficiency of cluster sampling can be increased by increasing the size of the sample. In many situations, it costs less to take considerably larger cluster samples than to take smaller simple random samples with the same precision. Hence, it is generally expected that efficiency per unit cost will be more in cluster sampling than in simple random sampling.

10.5.1 Estimation of population mean

Let N and n denote respectively the number of clusters in the population and in the sample respectively. Let M_i denote number of units in the

ith cluster. For simplicity consider the case of equal clusters, i.e. $M_i = M$ for all i . Then an unbiased estimate of the population mean is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n \bar{x}_i$$

$$\bar{x}_i = \frac{1}{M} \sum x_{ij} \text{ is the mean of } i\text{th cluster.}$$

Estimate of the variance of \bar{x} is given by

$$\text{Var } \bar{x} = \left(\frac{1}{n} - \frac{1}{N} \right) \left(\frac{1}{n-1} \right) \sum_{i=1}^n (\bar{x}_i - \bar{x})^2$$

Estimate of the population total is :

$$X = N \bar{x}$$

$$\text{and Var } (X) = N^2 \text{Var } \bar{x}$$

For details regarding estimation of population parameters when clusters are of unequal size, and selection is with replacement and unequal probability, readers may refer to Sukhatme (1954).

Example 4

For recording catch data of a marine landing centre a calendar month of 30 days was divided into 15 clusters of 2 consecutive days each and 7 clusters were randomly selected from these 15 clusters. The catch (in tons) data for the selected clusters are given below.

	Cluster No.	1	2	3	4	5	6	7
Catch (t)	1st day	6	7	6.5	8.2	7.4	6.6	8.5
	2nd day	8	6.2	8.3	10	8.0	8.8	7.1

- (i) Estimate the average catch per day for the landing centre alongwith its standard error.
- (ii) Estimate the total catch for the landing centre during the month.

Answer

Cluster	Catch (t)		mean (\bar{x}_i)
	1st day	2nd day	
1	6	8	7.0
2	7	6.2	6.6
3	6.5	8.3	7.4
4	8.2	10.0	9.1
5	7.4	8.0	7.7
6	6.6	8.8	7.7
7	8.5	7.1	7.8
Total			53.3

$$(i) \quad \text{Average catch per day } \bar{x} = \frac{\sum \bar{x}_i}{n} = \frac{53.3}{7} = 7.6143$$

$$\begin{aligned} \text{Var } (\bar{x}) &= \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{n-1} \left(\sum \bar{x}_i^2 - n\bar{x}^2 \right) \\ &= \left(\frac{1}{7} - \frac{1}{15} \right) \frac{1}{6} (409.55 - 405.8432) \\ &= \left(\frac{1}{7} - \frac{1}{15} \right) \left(\frac{3.7068}{6} \right) \\ &= (0.0762) (0.6178) = 0.0471 \end{aligned}$$

$$\text{Standard error} = \sqrt{\text{Var } (\bar{x})}$$

$$= 0.217$$

$$\begin{aligned} (ii) \quad &\text{Estimate of the total catch during the month} \\ &= (\text{number of days of the month}) \times (\text{average catch per day}) \\ &= 30 \times 7.6143 \\ &= 228.429 \text{ tons} \\ &\text{Standard error of the total catch,} \\ &= 30 \times 0.217 \\ &= 6.51 \end{aligned}$$

11.6 Subsampling or two stage sampling

A sampling method in which the sample is selected in stages, is called subsampling. In this method, the sampling units at each stage are subsampled from the units chosen at the previous stage. The first stage (primary) units are selected from the population. From each of these selected first stage units, the second stage units are selected. As the sample is taken in two stages subsampling is some times called two stage sampling. For instance, in estimating the total marine catches along the coastline of a certain state, certain landing centres (first stage units) on certain days are selected at random. Then from these selected landing centres, some boats landing the catches (second stage units) are selected for recording the catch data.

Sample can of course be selected in more than two stages in which it is called multistage sampling. For instance in the above example, if it was decided to record the length of fishes also, then the sample fishes (say 200 fish) can be taken from each selected boat. In this example, the sample is taken in 3 stages and hence called 3 stage or multistage sampling.

Usually individuals within the same primary unit are likely to resemble each other. Hence more number of primary units may be selected with few individuals from each.

11.6.1 Estimation of population mean (Equal first stage units)

Let the population be composed of NM elements grouped in to N first stage units of M second stage units each. Let n denote the number of first-stage units in the sample and m the number of second stage units selected from each selected first-stage units.

The estimate of the mean of any sampled first stage unit is given by

$$\bar{x}_i = \frac{1}{m} \sum_{j=1}^m x_{ij}$$

where x_{ij} is the value of the j th individual in the i th first-stage unit.

The estimate of the population mean is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n \bar{x}_i$$

Estimate of variance of sample mean is given by

$$\text{Var}(\bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{N} \left(\frac{1}{m} - \frac{1}{M}\right) s_w^2$$

$$\text{where } s_b^2 = \frac{\sum_{i=1}^n (\bar{x}_i - \bar{x})^2}{n-1}$$

$$s_w^2 = \frac{1}{n} \sum_{i=1}^n s_i^2$$

$$s_i^2 = \frac{\sum_{j=1}^m (x_{ij} - \bar{x}_i)^2}{m-1}$$

when $(N-n)/N$ and $(M-m)/M$ can be taken as unity, then

$$\text{var}(\bar{x}) = \frac{s_b^2}{n}$$

when $(M-m)/M$ alone can be taken as unity, then

$$\text{var}(\bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{Nm} s_w^2$$

For details regarding estimation of population mean, when the first stage units are unequal, readers may refer to Sukhatme (1954).

11.6.2 Advantages

The main advantages of subsampling are

- (i) Only the units of the population selected at any stage need be listed for sampling at the next stage, i.e., complete list of units of the population is not needed

- (ii) This sampling method allows for the use of different selection procedures at different stages.

It is because of this flexibility multistage sampling procedure is most commonly used in large sample surveys. For example, in estimating the marine fish catches, we use a multistage design in stratified random sampling with systematic selection of days in a month.

- (iii) Reduction in cost and ease of administration in the survey work.

Example 5

Four boats were selected randomly from the 100 boats that landed the catch on a given day in a landing centre and 20 mackerel from each selected boat were taken for recording length measurements. The following data were obtained.

Boat 1					
Length (cm)	18,	17.5,	20,	19,	18.5,
	21,	18,	14.5,	20,	15,
	16.5,	16,	14,	17,	19,
	20,	18.5,	15,	17,	20,
Boat 2	20,	18.5,	20.5,	19,	18,
	16,	15,	18.5,	20,	20.5,
	16.5,	21,	18.5,	19,	20,
	18,	20,	17,	20.5,	18,
Boat 3	15,	17,	18,	18.5,	20,
	12.5,	12,	14.5,	15,	16,
	17,	16,	19.5,	14,	18,
	16,	12,	17.5,	18,	18,
Boat 4	17,	18.5,	21,	20.5,	19,
	20,	14.5,	19,	20,	17,
	17.5,	18,	22,	16,	18,
	19,	16.5,	19,	17,	20

Estimate the mean length of mackerel during the day's landings, and its standard error.

Answer

Let x_{ij} denote the length, n the number of 1st stage units in the sample and m the number of second stage units sampled from each selected first stage unit.

Boat	1	2	3	4
$\sum \sum x_{ij}$	354.5	374.5	324.5	369.5
$\sum \sum x_{ij}^2$	6364.25	7064.75	5371.25	6891.2503
Mean, \bar{x}_i	17.725	18.725	16.225	18.475
Variance, s_i^2	= 4.2494	2.7493	5.5913	3.4073

Given, $n = 4$, $m = 20$

$$\text{Mean length, } \bar{x} = \frac{\sum \bar{x}_i}{n} = \frac{71.15}{4} = 18.7875$$

Variance between boats is given by,

$$\begin{aligned} s_b^2 &= \frac{1}{n-1} \sum_{i=1}^4 (\bar{x}_i - \bar{x})^2 \\ &= 1.2656 \\ s_w^2 &= \frac{1}{n} \sum s_i^2 = \frac{15.9973}{4} = 3.9993 \end{aligned}$$

Variance of sample mean (\bar{x}), when the number of 2nd stage units in the population is large relative to the 2nd stage units in the sample, is given by

$$\begin{aligned} \text{Var } (\bar{x}) &= \left(\frac{1}{n} - \frac{1}{N} \right) s_b^2 + \frac{s_w^2}{Nm} \\ &= 0.2400 \times 1.2656 + 0.0020 \\ &= 0.3037 + 0.0020 \\ &= 0.3057 \end{aligned}$$

$$\begin{aligned} \text{Standard error of sample mean } (\bar{x}) \\ &= \sqrt{0.3057} = 0.5529 \end{aligned}$$

11.7 Sampling design to estimate total marine fish landings

Our country has a long coast line of about 7515 km and there are about 1400 landing centres scattered along the coast. The sampling design developed and practiced by the Central Marine Fisheries Research Institute (CMFRI) provides the estimate of total marine fish landings for the entire nation. The sampling design adopted for the purpose is 'stratified multistage random sampling' the stratification being over space and time. Each maritime state is divided into several zones on the basis of geographical consideration and fishing practices.

11.7.1 Sampling for first stage units

Nine landing centres are selected at random from each zone for recording fish landings. A month is divided into 3 groups of 10 consecutive days. From the first ten days group, a day is selected randomly such that it falls within the first five days. Then 6 consecutive days from the selected day onwards are considered and these 6 days are grouped into 3 clusters of 2 consecutive days each. From the 2nd and 3rd group of 10 days, 3 clusters of two days each are chosen systematically with a sampling interval of 10 days. To illustrate, suppose that the 4th date (day) was selected from the group of first 10 days. Then 6 consecutive days from the selected day will be 4, 5, 6, 7, 8, 9 and these days are grouped into 3 clusters of 2 consecutive days, i.e., dates 4 and 5 will form one cluster, while 6 and 7 and 8 and 9 will form the other two clusters. From the 2nd group of 10 days, 6 days are systematically selected with a sampling interval of 10 days from the first date selected from the group of first ten days. Thus 6 days in the 2nd group will be 14, 15, 16, 17, 18 and 19 forming 3 clusters of dates 14 and 15, 16 and 17 and 18 and 19. From the 3rd group, 6 days are selected with the sampling interval of 10 days from the first day selected in the 2nd group, i.e., the dates will be 24, 25, 26, 27, 28 and 29 whose clusters are 24 and 25, 26 and 27, 28 and 29. Thus there are 9 clusters of two days each in a month. These 9 clusters are allotted to the 9 selected landing centres. On the first day of observation data are collected from 12 to 18 hours and the next day 6 to 12 hours. The data on night landings are collected by enquiry covering the period from 18 hours of the first day to 6 hours of the

next day. Thus a 24 hour period is covered for a landing centre. This forms the landing centre day and is the first stage sampling unit.

11.7.2 Sampling for second stage units

On the day of observation at the selected landing centre, if the total number of fishing units that land their catches is 10 or less, then, the data on all the units is collected. If the number of fishing units exceeds 10, a sample of boats is selected in a predetermined manner. Thus fishing units form the 2nd stage units on which data on species wise catch, effort, craft, and gear etc., are recorded.

11.7.3 Sampling for 3rd stage units

At the 3rd stage, samples of commercially important species are taken from the selected second stage units for biological observations.

11.7.4 Estimation of total landings

Based on the data collected from the selected fishing units, the total landings for the landing centre day are estimated. From these the monthly estimates for each year on a zonal, district and state basis are worked out together with the corresponding sampling errors.

11.8. Estimation of inland fish catch

As there is no standardised sampling technique for estimation of inland fish catch, the estimates at state level are computed based on various considerations such as, market arrivals, water area leased, lease value etc., differing from state to state. Some studies have been carried out by Indian Statistical Institute (1960-61, 1963) and by National Sample Survey Organisation (1962-63, 1973-75) to evolve suitable sampling methodology for estimation of inland fish catch, but without much success. Method of estimation of estuarine fish catch for Hooghly System and Mahanadi has been worked out by Central Inland Fisheries Research Institute (CIFRI), Barrackpore (Pillay and Gosh, 1962; Shetty and Gosh, 1963). In a recent attempt (1984) by CIFRI, Barrackpore and IASRI, New Delhi, a sampling design for estimation of fishery resources and catch has been suggested.

Chapter 12

BASIC EXPERIMENTAL DESIGNS

12.1 Introduction

Controlled experimentation is an important technique for collection of reliable data in aquaculture. In any controlled experiment, however meticulously planned, the response observations are affected not only by the action of treatments but also by some extraneous factors. In order to obtain reliable estimates of treatment effects and to draw valid inferences from experiments, the effect of extraneous factors has to be quantified and segregated from rest of the effects due to treatments. This can only be accomplished by designing the experiments suitably. Several designs are available in statistical literature to aid in proper planning and designing of experiments. Basic experimental designs useful in aquaculture experiments are described in this chapter.

12.2 Terminology

The terms which are frequently used in experimental designs are given below.

12.2.1 Experiment

An experiment is a planned enquiry to obtain new facts or to confirm or to deny the facts established earlier.

12.2.2 Treatment

The object of comparison is termed as treatment. For example, treatments may be different stocking rates, feeds, fish species, etc.

12.2.3 Experimental unit

An experimental unit is the unit of material to which a treatment is applied. For example, the experimental unit may be a pond, a fish, an animal, a piece of land, etc.

12.2.4 Experimental error

Experimental error is a measure of variation that exists among the experimental units treated alike. There are two main sources of experimental error.

- (i) inherent variability in the experimental material (unit) to which treatments are applied.
- (ii) variability due to lack of uniformity in the physical conduct of experiment or in other words, failure to standardise the experimental technique.

Experimental error provides a basis for the confidence to be placed in the results obtained from the experiment. Therefore, it is necessary to control the experimental error. Replication and local control, which will be discussed later, under the principles of experimental design, are helpful in reducing the experimental error.

12.3 Basic principles of experimental designs

The three basic principles of experimental design are (1) Randomization (2) Replication and (3) Local control. These 3 principles are the minimum requirements for any valid experimental design.

12.3.1 Randomization

Allocation of treatments to various experimental units by a random process is called randomization. Thus randomization ensures that all the experimental units have an equal chance of receiving a particular treatment. Its function is to provide unbiased estimates of treatment means and experimental error.

12.3.2 Replication

Repetition of the treatment under investigation is known as replication. Its function is to provide an estimate of experimental error and to improve the precision of the treatment effects and hence of the experiment. A minimum of 2 replications are required to estimate the

experimental error. As the precision of the experiment increases with the increase in the number of replication, maximum number of replications (feasible) should always be tried.

12.3.3 Local control

Local control is a device of grouping experimental units into groups of relatively homogeneous units. Local control helps in reducing the experimental error.

12.4 Experimental designs

The number and nature of the treatments proposed to be included in the experiment help in selecting the appropriate experimental design. Commonly used designs are given below.

12.4.1 Completely randomized design (CRD)

This is the simplest of all designs which uses two principles of experimental designs, namely, replication and randomization. In this design each treatment is applied randomly to a few experimental ponds or units. It is not necessary that the number of replications for each treatment be the same. However, to estimate the treatment effects with equal precision it is better to have each treatment replicated equally. This design is suitable only when the experimental units receiving treatments are homogeneous (e.g., cement cisterns). Hence, the design is mostly used in laboratory experiments.

12.4.1.1 Layout of the design

The term layout refers to the placement of treatments on the experimental units. Suppose it is planned to compare 4 treatments A, B, C and D, with 5 replicates of each treatment. Then 20 experimental units are required. These experimental units are then numbered from 1 to 20. Draw 5 numbers which are less than 20, using random number tables or by drawing lots. They may turn out to be, say, 8, 19, 7, 12, and 3. The experimental units bearing these numbers will receive say, treatment A.

The second set of 5 random numbers to which say treatment B will be applied is drawn, which may turn out to be say, 1,4,13, 18 and 20. The third set of 5 random numbers is drawn which may turn out to be 9,15,17,6 and 14 and the experimental units bearing these numbers receive say treatment C. Treatment D is applied to the remaining 5 experimental units. One of the possible layouts of the design is shown in Figure 1.

1 _B	2 _D	3 _A	4 _B
5 _D	6 _C	7 _A	8 _A
9 _C	10 _D	11 _D	12 _A
13 _B	14 _C	15 _C	16 _D
17 _C	18 _B	19 _A	20 _B

Fig. 1. Layout of completely randomized design

12.4.1.2 Analysis

Let x_{ij} denote the observation on the i th treatment from j th replication ($i = 1, 2, \dots, t, j = 1, 2, \dots, r_i$). For analysis of this design, the following additive model is used :

$$x_{ij} = m + a_i + e_{ij}$$

where m is the general mean

a_i is the effect due to i th treatment

e_{ij} is experimental error which is independently normally distributed with mean zero and variance σ^2

Null hypothesis to be tested is,

H_0 : There is no difference among treatment effects

Analysis of variance technique is used to test this hypothesis.

Tabulate the data collected from an experiment as shown in Table 1.

Table 1. Data from a completely randomized design

Treatments		Observed yield from experimental Units					Total
(A)	1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	T_1
(B)	2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	T_2
(C)	3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	T_3
(D)	4	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	T_4

The following computations are required :

i. Correction factor (CF) = $\frac{(G)^2}{n}$

Where G is the grand total of all the observations and 'n' is the total number of observations, i.e., $n = rt$ where 'r' is number of replications and 't' is the number of treatments.

ii. Total sum of square (TSS)

$$\begin{aligned}
 &= \text{Sum of squares of all observations} - CF \\
 &= (x_{11}^2 + x_{12}^2 + \dots + x_{44}^2 + x_{45}^2) - CF \\
 &= \sum \sum x_{ij}^2 - CF
 \end{aligned}$$

III. Treatment sum of squares (TRSS)

(when the number of replications is the same for each treatment)

$$= \frac{T_1^2 + T_2^2 + T_3^2 + T_4^2}{r} - CF$$

$$= \frac{\sum T_i^2}{r} - CF$$

Treatment sum of squares when the number of replications is different for each treatment is given by,

$$TRSS = \frac{T_1^2}{r_1} + \frac{T_2^2}{r_2} + \frac{T_3^2}{r_3} + \frac{T_4^2}{r_4} - CF$$

where T_1, T_2, T_3 and T_4 are yield totals for the 1st, 2nd, 3rd and 4th treatments respectively.

IV. Error sum of squares (ESS)

$$= TSS - TRSS$$

$$= II - III$$

These computations can be summarized in the form of analysis of variance (ANOVA) table. The ANOVA for CRD with 't' treatments and 'r' replications is given below :

Source of variation	df	Sum of squares (SS)	Mean square (MS) = $\frac{SS}{df}$	F
Treatments	t-1	TRSS	$\frac{TRSS}{t-1} = M_1$	$F(\text{cal}) = \frac{M_1}{E}$
Error	t(r-1)	ESS	$\frac{ESS}{t(r-1)} = E$	
Total	rt-1	TSS		

To test the hypothesis H_0 , $F(\text{cal})$ has to be compared with the table value of F with $(t-1)$ and $t(r-1)$ df., at a desired level of significance, generally at 5% and 1% levels of significance.

If $F(\text{cal}) > F$ table reject H_0

When $F(\text{cal}) > F$ table F value is said to be significant. If it is significant at 5%, one asterisk is put on F value, whereas if it is significant at 1% two asterisks are put. When H_0 is rejected, the treatment means that differ significantly may be found out. This is done by computing the critical difference (CD). The formula for computing CD when the number of replications is the same for each treatment is given by,

$$CD = \left(\sqrt{\frac{2 E}{r}} \right) t$$

The value of t is obtained from the t -tables at 5% level of significance with the error degrees of freedom.

If the number of replications is not the same for each treatment then CD for comparing two treatments which have been replicated r_i and r_j times is given by,

$$\left(\sqrt{\left(\frac{1}{r_i} + \frac{1}{r_j} \right) E} \right) t$$

12.4.1.3 Advantages and disadvantages

The main advantages are

- (i) The design is very flexible and can be used for any number of replications. Replications can vary between the treatments.
- (ii) The design allows the maximum number of degrees of freedom for error (the precision of small experiments increases with error degrees of freedom).
- (iii) The statistical analysis is simple.

- (iv) Unequal number of replications for the various treatments does not affect the simplicity of the statistical analysis.

The main disadvantage of the design is that it is usually suitable only for small number of treatments and for homogeneous experimental material.

Example 1

Five test diets were tested against the growth performance of a certain fish in plastic pools for a period of 1 month. The daily feed provided was 50% of the total weight of 40 fry kept in each plastic pool. The experimental design used was completely randomized design and each treatment was replicated 4 times.

Growth performance is given below :

Treatments (test diets)	Net gain in weight (g)/fish				Total	Mean
	Rep. 1	Rep. 2	Rep. 3	Rep. 4		
A	0.95	0.85	0.85	0.90	3.55	0.89
B	0.43	0.45	0.40	0.42	1.70	0.43
B	0.70	0.90	0.75	0.70	3.05	0.76
D	1.00	0.95	0.90	0.90	3.75	0.94
E	0.90	1.00	0.95	0.95	3.80	0.95
Grand Total					15.85	

Ho : There is no significant difference among treatment means.

i. Correction factor (CF) = $\frac{(15.85)^2}{20} = 12.5611$

ii. Total sum of squares (TSS)

$$= (0.95)^2 + (0.85)^2 + \dots + (0.95)^2 + (0.95)^2 - CF$$

$$= 13.3713 - 12.5611 = 0.8102$$

III. Treatment sum of squares (TRSS)

$$= \frac{(3.55)^2 + (1.70)^2 + (3.05)^2 + (3.75)^2 + (3.80)^2}{4} - CF$$

$$= 13.3244 - 12.5611 = 0.7633$$

IV. Error sum of squares (ESS) = TSS - TRSS = 0.0469.

These computations can be summarized in the form of analysis of variance table given below.

Source of variation	d.f.	SS	MS	F
Treatments	4	0.7633	0.19080	$F = \frac{0.1900}{0.0031} = 61.5484^{**}$
Error	15	0.0469	0.0031=E	
Total	19	0.8102		

From F table it is found that F with, 4 and 15 df

= 3.06 at 5%

= 4.89 at 1% level of significance.

F value is significant at 1% as $F(\text{cal}) > 4.89$. Hence, the hypothesis is rejected and the conclusion is that the mean gain in weight for the test diets differed significantly. It will be of interest to know which of the treatment means differs significantly. Compute CD using the formula,

$$CD = \left(\sqrt{\frac{2E}{r}} \right) t$$

$$= \sqrt{\left(2 \frac{(0.0031)}{4} \right)} 2.131$$

$$= 0.08$$

Arrange the treatment means in the descending order

Treatment (Test diet)	E	D	A	<u>C</u>	B
Mean gain in weight	0.95	0.94	0.89	0.76	<u>0.43</u>

The treatments which do not differ have been joined by a line. In this experiment treatment means of E, D and A do not differ by more than 0.08 (CD value). Hence they have been joined by a line. But the difference between treatment means E and C, D and C and A and C is more than the CD values, indicating that treatment C differs from E, D and A. Similar interpretation holds good for the other treatments also.

Conclusions

Test diet E recorded maximum net gain in weight (0.95 g/fish) and the minimum gain (0.43 g/fish) was recorded by test diet B. The effect of diet E on growth performance was on par with the effect of diets D and A and these 3 diets were significantly superior to C and B.

12.4.2 Randomised complete block design

The completely randomized design, discussed earlier is suitable only when all the experimental units (e.g. ponds) are homogeneous, which is difficult to ensure in many cases. However, it may be possible to get a group (block) of homogeneous units (e.g. ponds in a row or ponds of similar size etc.). After grouping the experimental units in to homogeneous blocks, the treatments are allocated randomly to the experimental units within the block such that each treatment appears once in each block. The number of experimental units in each block equals the number of treatments. Such type of experimental plan is called 'Randomised complete Block Design (RBD)'. In this design fresh randomization is needed for each block, i.e., randomization is restricted to a block. This is the difference between RBD and CRD.

The design is more suitable when there is one source of variability as it can be controlled by suitable blocking. For instance the design is particularly useful if soil fertility varies in one direction.

The blocks or replications do not necessarily mean blocks in fields/farms. In fish feeding experiments for instance, blocks may be made up of fish of the same age or weight or some other factor. RBD can be used in experiments for comparing the efficiency of different feeds,

effect of stocking densities, different doses of fertiliser, water level, etc., on fish yield or on any other character of interest. This design is very popular in agricultural field experiments.

12A.2.1 Layout of the design

To illustrate the layout of the RBD consider an experiment in which 4 feeds, A, B, C, and D are to be compared for the growth of a major carp from the stage the fry are going to be stocked in stocking ponds. If it is decided to use 5 replications, then there have to be 5 blocks each containing 4 stocking ponds. These ponds are numbered from 1 to 4 in each block. For randomizing treatments in block I, numbers less than or equal to 4 are randomly drawn from random number tables or by drawing lots. If the numbers drawn appear in the following order 3, 1, 2, 4 then allot treatment A to pond number 3, B to pond number 1, C to pond number 2 and D to pond number 4. Likewise treatments can be randomized within block numbers II, III, IV and V by drawing fresh set of 4 numbers randomly for each block separately. Layout of the design is given in figure 2, which is one of the many possible layouts.

Block

I	¹ B	² C	³ A	⁴ D
II	¹ D	² A	³ C	⁴ B
III	¹ A	² D	³ B	⁴ C
IV	¹ B	² C	³ A	⁴ D
V	¹ C	² A	³ D	⁴ B

Fig. 2 : A layout of the randomised complete block design.

12.4.2.2 Analysis

Let x_{ij} denote the observation on the i th treatment ($i = 1, 2 \dots t$) in the j th replication ($j = 1, 2 \dots r$). The following additive model is assumed :

$$x_{ij} = m + a_i + b_j + e_{ij}$$

where m is the general mean

a_i is the effect due to i th treatment

b_j is the effect due to j th block

e_{ij} is experimental error which is assumed to be independently and normally distributed with mean zero and variance σ^2 .

The null hypothesis to be tested is,

H_0 : There is no difference among treatment effects

Analysis of variance technique is used to test this hypothesis.

The data collected are arranged in the following form :

Treatments	Replications					Total
	I	II	III	IV	V	
A	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	T ₁
B	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	T ₂
C	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	T ₃
D	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	T ₄
Total	R_1	R_2	R_3	R_4	R_5	G ₁

Compute,

$$I. \quad \text{Correction Factor (CF)} = \frac{G^2}{n}$$

where G is the grand total of all the observations $n = rt$ is the total number of observations.

$$\begin{aligned} II. \quad & \text{Total sum of squares (TSS)} \\ & = \text{sum of squares of all observations} - CF \\ & = (x_{11}^2 + x_{12}^2 + \dots + x_{44}^2 + x_{45}^2) - CF \\ & = \sum \sum x_{ij}^2 - CF \end{aligned}$$

$$\begin{aligned} III. \quad & \text{Replication sum of squares (RSS)} \\ & = \frac{R_1^2 + R_2^2 + R_3^2 + R_4^2 + R_5^2}{\text{number of treatments}} - CF \\ & = \sum \frac{R_j^2}{t} - CF \end{aligned}$$

where R_1, R_2, R_3, R_4 and R_5 totals of 1st, 2nd, 3rd, 4th and 5th replications respectively, t is the number of treatments.

$$\begin{aligned} IV. \quad & \text{Treatment sum of squares (TRSS)} \\ & = \frac{T_1^2 + T_2^2 + T_3^2 + T_4^2}{\text{Number of replications}} - CF \\ & = \sum \frac{T_i^2}{r} - CF, \end{aligned}$$

where T_1, T_2, T_3 and T_4 are totals of the 1st, 2nd, 3rd and 4th treatments respectively and r is the number of replications.

$$\begin{aligned} V. \quad & \text{Error sum of squares (ESS)} \\ & = \text{TSS} - (\text{RSS} + \text{TRSS}) \\ & = II - (III + IV) \end{aligned}$$

These computations can be summarised in the form of ANOVA table.

Source of variation	df	SS	MS	F
Replications	r-1	RSS	$\frac{RSS}{r-1} = M_1$	$F(\text{Cal}) = \frac{M_2}{E}$
Treatments	t-1	TRSS	$\frac{TRSS}{t-1} = M_2$	
Error	(r-1)(t-1)	ESS	$\frac{ESS}{(r-1)(t-1)} = E$	
Total	n-1	TSS		

For testing the hypothesis, the value of $F(\text{Cal})$ has to be compared with the table value of F with $(t-1)$ and $(r-1)(t-1)$ df at a desired level of significance.

If $F(\text{cal}) > F \text{ table}$, reject H_0

When H_0 is rejected, it may be of interest to know which of the treatment effects differ significantly. This is done by calculating CD using the formula,

$$CD = \left(\sqrt{\frac{2E}{r}} \right) t$$

The value of t is obtained from t tables at 5% level of significance with error degrees of freedom.

The treatments for which the means differ by CD value or more will be considered as differing significantly.

12.4.2.3 Advantages and disadvantages

The main advantages are,

- (i) Design is more accurate than CRD for most types of experimental work as variation due to blocks can be eliminated from experimental error, thereby reducing the error.
- (ii) Design is flexible as no restrictions are placed on the number of treatments or on the number of replications in an experiment.
- (iii) The statistical analysis is simple.

The main disadvantage of the design is that it is not suitable for experiments with large number of treatments since the blocks become too larger and lose their homogeneity.

Example 2

To study the effect of stocking density on a certain fish species, an experiment was conducted with 6 different stocking densities (treatments) of fingerlings in ponds of size 0.02 ha using randomized block design with 4 replications. All cultural practices except the stocking densities were kept the same. Harvesting was done after 6 months of stocking. The yield t/ha for different stocking densities is given below. Find out whether there is significant difference among yields obtained at different stocking densities.

Stocking Densities (per hectare)	Replications				Total
	R I	R II	R III	R IV	
20,000	3.6	2.8	3.0	4.0	13.4
30,000	4.8	4.2	4.0	5.6	18.6
40,000	6.0	5.7	5.2	6.2	23.1
50,000	6.6	6.4	5.4	6.5	25.9
60,000	7.0	6.5	5.9	7.0	26.4
70,000	7.1	6.8	6.0	7.2	27.1
	35.1	32.4	29.5	36.5	133.5

H_0 : There is no significant difference among treatment effects.

The following computations are required to test the hypothesis:

$$\begin{aligned}
 \text{I. CF} &= \frac{(133.5)^2}{24} = 742.5938 \\
 \text{II. TSS} &= (3.6)^2 + (2.8)^2 + \dots + (6)^2 + (7.2)^2 - \text{CF} \\
 &= 783.29 - \text{CF} \\
 &= 40.6962 \\
 \text{III. RSS} &= \frac{(35.1)^2 + (32.4)^2 + (29.5)^2 + (36.5)^2}{6} - \text{CF} \\
 &= 747.3783 - \text{CF} \\
 &= 4.7845 \\
 \text{IV. TRSS} &= \frac{(13.4)^2 + (18.6)^2 + \dots + (27.1)^2}{4} - \text{CF} \\
 &= \frac{3110.51}{4} - \text{CF} \\
 &= 35.0337 \\
 \text{V. ESS} &= \text{TSS} - (\text{RSS} + \text{TRSS}) \\
 &= 40.6962 - (4.7845 + 35.0337) \\
 &= 0.878
 \end{aligned}$$

These computations can be summarized in the form of ANOVA table given below.

Source of variation	d.f.	SS	MS	F
Replication	3	4.7845	1.5948	
Treatments	5	35.0337	7.0067	F(cal) = 119.77**
Error	15	0.870	0.0585	
Total	23	40.6962		

** Significant at 1%

F table = 2.90 at 5% level of significance
 = 4.56 at 1% level of significance

The calculated value of F is significant at 1% level of significance, indicating that the mean yield per pond for all the treatments differed significantly.

To find out which of these treatments differ significantly CD was worked out. It was found to be 0.36

Treatments (stocking density/ha)	70,000	60,000	50,000	40,000	30,000	20,000
Mean yield (per pond)	<u>6.78</u>	<u>6.6</u>	<u>6.23</u>	<u>5.78</u>	<u>4.65</u>	<u>3.35</u>

Conclusions

Stocking density 20,000/ha recorded the minimum yield, whereas 70,000/ha recorded the maximum yield. There was no significant difference between the yields obtained at 60,000 and 70,000/ha densities. The yield from these stocking densities was significantly higher than that from all other densities tested in the experiment. The yields obtained at 50,000/ha; 40,000/ha; 30,000/ha; and 20,000/ha differed significantly from one another.

12.4.3 Latin square design (LSD)

Variation due to one factor of variability can be effectively controlled by adopting RBD. Often there is variation in respect of more than one, for example, pond size, depth, shape etc. The variation in respect of two factors can be controlled by using the 'Latin Square Design (LSD)'. In this design two restrictions are imposed by forming blocks in two directions, row-wise and column-wise.

In this design, the number of treatments equals the number of replications. If 'r' stands for the number of treatments as well as for the number of replications of each treatment, then the total number of

experimental units required for this design is $r \times r = r^2$. These r^2 units are arranged in 'r' rows and 'r' columns, each corresponding to different sources of variation. Then the 'r' treatments are assigned to these r^2 experimental units in such a way that each treatment appears only once in each row and in each column. In RBD there was one restriction that each treatment must appear once in each block, LSD differs from RBD, in that, two restrictions are imposed namely each treatment must appear once in each row and each column. LSD is preferred to RBD when there are two sources of variability.

12.4.3.1 Layout of the design

Treatments have to be assigned to experimental units in such a way that every treatment occurs only once in each row and once in each column. This can be done in large number of ways. The way in which it has to be done has to be decided randomly.

There are 2 possible arrangements for 2×2 latin square of which one can be selected randomly as a layout of the design. For 3×3 latin square there are 12 possible arrangements of which one can be selected randomly as the layout for conducting the experiment. The number of possible arrangements increases with the increase in size of the latin square. For 5×5 latin square there are 161280 possible arrangements. Complete enumeration and selection for such a large number of squares is tedious and hence in 'statistical tables for biological, agricultural and medical research', Fisher and Yates (1963) have given set of squares for 4×4 ; 5×5 and 6×6 from which all possible arrangements could be obtained by permuting rows, columns and letters. From these set of squares first select a square at random. Then, in the case of 4×4 and 5×5 permute all rows except the first of the selected square, and all columns. Alternatively permute all rows except the first and assign the letters to the treatments at random. For 6×6 squares permute all rows and columns at random and then assign the treatments randomly to the letters. Only 4 squares for 7×7 and one square each for 8×8 to 12×12 latin squares are available (Fisher & Yates, 1963). For squares of these sizes it is enough, if we take only given square and permute randomly all rows, columns and treatments.

One of the possible arrangements of 5 x 5 latin squares is given in Figure 3.

		Columns					
		A	E	D	C	B	
		D	B	A	E	C	A, B, C, D and E are treat- ments
Rows		B	A	C	D	E	
		C	D	E	B	A	
		E	C	D	A	D	

Fig. 3 : A Layout of latin square design

12.4.3.2 Analysis

For analysis of this design the following additive model is used :

$$x_{ijk} = m + a_i + b_j + c_k + e_{ijk}$$

where x_{ijk} is the observation on i th treatment in j th row and k th column ($i, j, k = 1, 2, \dots, r$)

m is the general mean effect

a_i is the effect due to i th treatment

b_j is the effect due to j th row

c_k is the effect due to k th column

e_{ijk} is experimental error which is assumed to be independently and normally distributed with mean zero and variance σ^2 .

The null hypothesis to be tested is that there is no difference among treatment effects.

The analysis of data proceeds almost in a similar way as in the case of RBD. But here we will have Row sum of squares (ROSS) which is computed using the row totals, column sum of squares (COSS) using the column totals and treatment sum of squares (TRSS) using the treatment totals.

The following computations are required for testing the null hypothesis:

$$I. \quad \text{Correction Factor (CF)} = \frac{(G)^2}{r}$$

where G is the grand total of all the observations, r = Number of treatments = Number of replications.

$$II. \quad \text{Total sum of squares (TSS)} \\ = \text{Sum of squares of all observations} - CF \\ = \sum \sum \sum x_{ijk}^2 - CF$$

$$III. \quad \text{ROSS} = \frac{\sum R_j^2}{r} - CF$$

where R_j is the total of the j th row

$$IV. \quad \text{COSS} = \frac{\sum C_k^2}{r} - CF$$

$$V. \quad \text{TRSS} = \frac{\sum T_i^2}{r} - CF$$

where T_i is the total of observations of the i th treatment.

$$VI. \quad \text{Error sum of squares (ESS)} \\ = \text{TSS} - (\text{ROSS} + \text{COSS} + \text{TRSS}) \\ = II - (III + IV + V)$$

These computations can be summarized in the form of ANOVA table as given below:

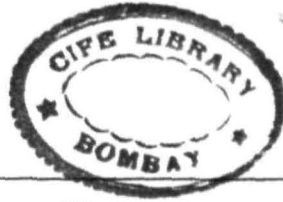


Table : ANOVA table for LSD

Source of Variation	d.f.	SS	MS
Rows	$r-1$	ROSS	$\frac{ROSS}{r-1} = M_1$
Columns	$r-1$	COSS	$\frac{COSS}{r-1} = M_2$
Treatments	$r-1$	TRSS	$\frac{TRSS}{r-1} = M_3, F(\text{cal}) = \frac{M_3}{E}$
Error	$(r-1)(r-2)$	ESS	$\frac{ESS}{(r-1)(r-2)} = E$
Total	$r^2 - 1$		

For testing the hypothesis, $F(\text{cal})$ has to be compared with the table value of F with $(r-1)$ and $(r-1)(r-2)$ df at say 5% and 1% level of significance.

If $F(\text{cal}) > F \text{ table}$, reject H_0

When H_0 is rejected, treatment effects that differ significantly may be found out. This is done by calculating CD using the formula,

$$CD = \left(\sqrt{\frac{2E}{r}} \right)_t$$

The value of t is obtained from t -tables at 5% level of significance with error degrees of freedom.

The treatments for which means differ by this CD value or more will be considered as differing significantly.

12.3.3.3 Advantages and disadvantages

The advantages are,

- i) With a two-way grouping, the latin square design controls more variation than CRD and RBD, which results in a small error mean square.
- ii) Statistical analysis is simple

The disadvantages are

- i) Due to its rigid nature, the number of replications should be equal to the number of treatments, this design becomes impracticable for large number of treatments. The design is seldom used for more than 10 to 12 treatments.
- ii) For less than 5 treatments, latin square design may not be as efficient as RBD or CRD as it does not provide sufficient number of degrees of freedom for reliable estimation of experimental error.

Example 3

Five test diets (A, B, C, D and E) were tested for the growth performance of a certain fish species for a period of 6 months adopting the latin square design taking initial weight of fish as rows and initial age as columns. The net gain in weight (g) per fish is given below, Find out whether there is significant difference among yields obtained from test diets A, B, C, D and E.

Initial age

		Initial age					Total
		D	E	C	B	A	
Initial weight	D	35	33	31	29	26	154
	C	50	47	41	46	31	215
	E	43	30	25	37	35	170
	B	40	39	40	25	38	182
	A	27	25	21	31	25	129
Total	195	174	158	168	155	850	

Answer :

H_0 : There is no significant difference among treatment effects.

$$1. \quad CF = \frac{G^2}{r^2} = \frac{(850)^2}{25} = 28900$$

$$\begin{aligned}
 2. \quad TSS &= \sum \sum \sum x_{ijk}^2 - CF \\
 &= 30398 - 28900 \\
 &= 1498 \dots\dots\dots (I)
 \end{aligned}$$

$$\begin{aligned}
 3. \quad \text{ROSS} &= \frac{\sum R_j^2}{r} - \text{CF} \\
 &= \frac{148606}{5} - \text{CF} \\
 &= 821.1 \dots\dots\dots (II)
 \end{aligned}$$

$$\begin{aligned}
 4. \quad \text{COSS} &= \frac{\sum C_k^2}{r} - \text{CF} \\
 &= \frac{145514}{5} - \text{CF} \\
 &= 202.8 \dots\dots\dots (III)
 \end{aligned}$$

$$5. \quad \text{TRSS} = \frac{\sum T_i^2}{r} - \text{CF}$$

Where T_1, T_2, T_3, T_4 and T_5 are totals of A, B, C, D, and E treatments respectively. In the given example

$$T_1 = 149, T_2 = 150, T_3 = 186, T_4 = 178, T_5 = 187$$

$$\begin{aligned}
 \text{Hence, TRSS} &= \frac{\sum T_i^2}{r} - \text{CF} \\
 &= \frac{145950}{5} - \text{CF} \\
 &= 290 \dots\dots\dots (IV)
 \end{aligned}$$

$$\begin{aligned}
 6. \quad \text{ESS} &= \text{Total SS} - (\text{ROSS} + \text{COSS} + \text{TRSS}) \\
 &= (I) - (II + III + IV) \\
 &= 1498 - (821.2 + 202.8 + 290) \\
 &= 1498 - 1314 \\
 &= 184
 \end{aligned}$$

Analysis of variance table

Source	d.f.	SS	MS	F
Rows (Initial weight)	4	821.20	205.30	13.39**
Columns (Initial age)	4	202.80	50.70	3.31*
Treatments	4	290.00	72.50	4.73*
Error	12	184.00	15.33	
Total	24			

* - Significant at 5%

** - Significant at 1%

F table = 3.26 at 5% level of significance

= 5.41 at 1% level of significance

Conclusions

- i) F value is significant at 1% for rows indicating that mean gain in weight varied depending upon the initial weight.
- ii) F value is significant at 5% level of significance for treatments, indicating that mean gain in weight differs significantly for test diets A, B, C, D and E.

T2.5 Advanced Designs

Designs discussed above are the 3 basic experimental designs. For advanced designs such as confounded, split plot, composite, residual effect designs, etc., readers may refer to Cochran and Cox (1957) and Nigam & Gupta (1979).

Chapter 13

TIME SERIES

13.1 Introduction

Time series is a series of observations, constituting a statistical data, observed at different units of time, such as years, months, days etc. For instance, time series may represent the fish production of a country over the years, prices of fish over the months, growth of fish over the weeks etc. The basic assumption in the time series analysis is that those factors which have influenced the observations, in the past and present will continue to influence more or less in the same manner in future. Therefore, the objective of time series is to identify and isolate these factors for predictive purpose.

13.2 Components of time series

Observations of time series vary with time, due to the effect of certain factors. These factors are generally referred to as 'components of time series'. They are,

- i) Trend or secular trend
- ii) Seasonal fluctuations
- iii) Cyclic fluctuations
- iv) Irregular fluctuations

Every observation of the time series is assumed to be the joint effect of these four components. Mathematically, $Y = T \times S \times C \times I$

Where T, S, C, and I indicate the effect of trend, seasonal fluctuations, cyclic fluctuations and irregular fluctuations, respectively. When the data are recorded annually, then an observation is expressed as

$$Y = T \times C \times I$$

13.2.1 Trend

Most of the time series exhibit a general tendency to increase or decrease over a long period of time. This basic tendency is called the 'trend' or secular trend of a time series (Fig.1).

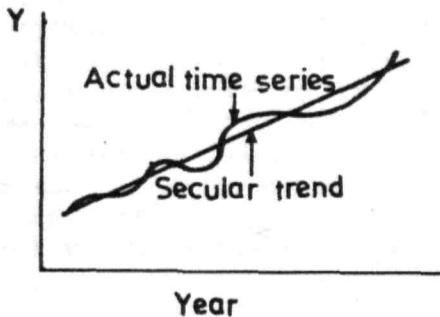


Fig. 1. Secular trend

Thus trend is the overall change taking place in time series over a long period of time. It reflects the effect of forces that constitute gradual growth or decline without sudden reversal of directions. Some examples of secular trend are, steady increase of population over a period of time, steady increase of inland fish production over the last few years.

13.2.2 Seasonal fluctuations

Seasonal fluctuations refer to regular and periodic variations that occur in a time series over a short period. These fluctuations repeat at regular intervals of time, say every day, every month etc.

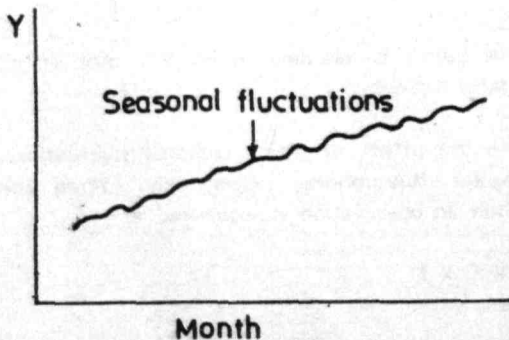


Fig. 2. Seasonal fluctuations

Generally the periodicity of seasonal fluctuations is less than one year. Weather conditions, customs, traditions and habits of people etc., cause seasonal fluctuations. Some examples of seasonal fluctuations are

- i) Phytoplankton production will be more during day time.
- ii) Landings of some marine fishes are more during lunar phases.

13.2.3 Cyclic fluctuations

Most of time series on economic activities are influenced by the periods of prosperity and depression. In times of prosperity, production, sales, employment etc., are high and in the times of depression, the opposite is true. Thus the periods of prosperity and depression cause upward

and downward movements in time series. These movements or fluctuations are called 'cyclic fluctuations'. These fluctuations differ from seasonal fluctuations in that they are of longer duration than a year and they do not generally exhibit regularity in their occurrence.

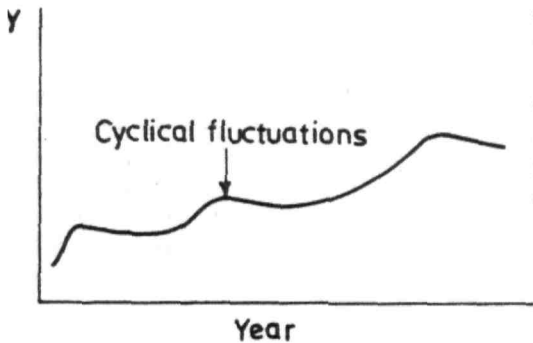


Fig. 3. Cyclical fluctuations

13.2.4 Irregular fluctuations or random fluctuations

Fluctuations caused by chance events such as wars, floods, strikes etc., are called irregular fluctuations. These fluctuations are unpredictable and their effect lasts generally for a short period. For example strike by fishermen will push down fish production, a fire in a departmental store will influence sales,

13.3 Analysis of time series

The process of separating out the different components of the time series and studying them individually is known as analysis of time series.

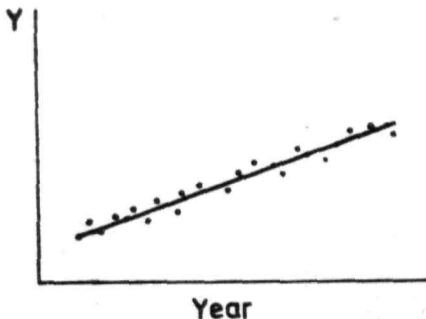
13.3.1 Estimation of trend

The following are three important methods of measuring the trend :

- i) Free hand method or graphic method
- ii) Method of moving averages
- iii) Method of least squares

13.3.1.1 Free hand method or graphic method

This is the most simple method of estimating the trend values. In this method first, the original observations of the time series are plotted on a graph paper. Then a smooth curve passing through many of



these points is drawn. This curve describes the trend. This method is quite easy to understand and does not involve any mathematical complications and saves time. The main drawback of this method is, it does not lead to unique results as different persons may draw different trend lines for the same set of data.

Fig. 4. Eye fitted trend

13.3.1.2 Method of moving averages

The method of moving averages is a simple but effective method of measuring the secular trend in a time series. It consists of calculating the simple arithmetic mean by taking specified number of observations at a time say 3, 4, 5 etc., and writing it in the centre of these observations. Then repeat the process by adding the next observation and dropping the initial observation till all the observations of the series are exhausted. The number of observations taken at a time is called the 'period of moving average'. These averages are called moving averages as the process of taking the averages goes on moving from beginning of the table to the end.

An advantage of this method is that it reduces the variations.

The drawback of this method is that it does not give the values for a few observations at the beginning of the series and for a few observations at the end of the series.

Example 1

Find the trend values for the following data on total fish production in the country by i) 3 yearly and ii) 5 yearly moving averages. Plot the given data and the trend values.

Year	1974	1975	1976	1977	1978	1979	1980
Production (lakh tons)	22.56,	22.66,	21.74,	23.12,	23.06,	23.40,	24.42

Answer

Year	Fish Production (lakh tons)	3 yearly moving average	5 yearly moving average
1974	22.56	-	-
1975	22.66	22.32*	-
1976	21.74	22.51	22.63**
1977	23.12	22.64	22.80
1978	23.06	23.19	23.15
1979	23.40	23.63	-
1980	24.42	-	-

$$*22.32 = \frac{22.56 + 22.66 + 21.74}{3}$$

$$**22.63 = \frac{22.56 + 22.66 + 21.74 + 23.12 + 23.06}{5}$$

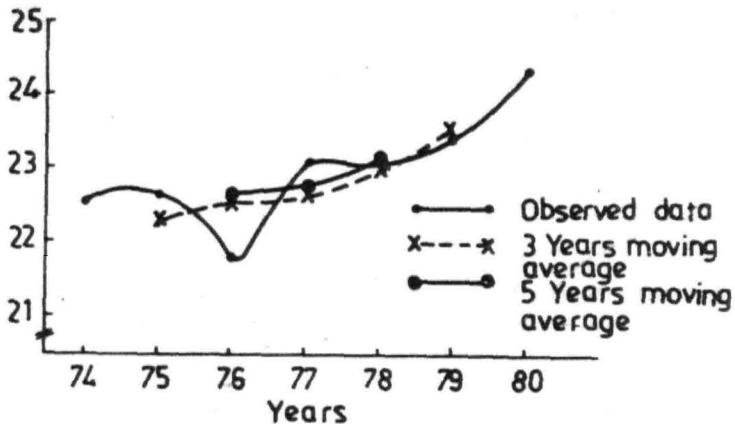


Fig. 5. Trend by moving average method

13.3.1.3 Method of least squares

The method of least squares described in chapter 10 can be used to fit appropriate trend line for the given time series data. Appropriate trend may be linear (straight line) or exponential or quadratic (parabolic) depending upon the variable under study. The method is objective and highly efficient. Trend line worked out by this method can be used for predicting immediate future trend.

i) Linear trend

Linear trend is estimated by the equation,

$$Y = a + bt \dots \dots \dots (1)$$

Where Y is the variable under study and t is the time period. The parameters of the equation 'a' and 'b' are estimated by the method of least squares discussed in Chapter 10. Computations can be made easier by choosing t such that $\sum t = 0$. In such cases 'a' and 'b' are estimated using the following formulae :

$$a = \frac{\Sigma Y}{n}$$

$$b = \frac{\Sigma Yt}{\Sigma t^2}$$

Example 2

Fit a linear trend to the data given in example 1, graph the original and trend values, estimate the production for 1981.

Answer

In the given example $n = 7$, therefore middle most year is taken as zero and other values are written as shown in the table.

Year	Fish production (lakh tonnes) (Y).	t	tY	t ²
1974	22.56	-3	67.68	9
1975	22.66	-2	45.32	4
1976	21.74	-1	21.74	1
1977	23.12	0	0	0
1978	23.06	1	23.06	1
1979	23.40	2	46.80	4
1980	24.42	3	73.26	9
	160.96	0	8.38	28

$$a = \frac{\Sigma Y}{n} = \frac{160.96}{7} = 22.99$$

$$b = \frac{\Sigma tY}{\Sigma t^2} = \frac{8.38}{28} = 0.299$$

Hence, $Y = 22.99 + 0.299 t \dots \dots \dots$ (II)

Trend values for different years are estimated using equation (II). To get trend value for 1974, put $t = -3$ in equation (II), which gives $Y = 22.093$. To get trend value for 1975, put $t = -2$ in equation (II) and so on. Trend values for different years are given below :

Year	Observed	Trend value (Estimated)
1974	22.56	22.093
1975	22.66	22.392
1976	21.74	22.691
1977	23.12	22.990
1978	23.06	23.289
1979	23.40	23.588
1980	24.42	23.887

Estimated production for 1981 is obtained by putting $t = 4$ in equation (II).

$$\begin{aligned} \text{i.e., } Y &= 22.99 + 0.299 \cdot (4) \\ &= 24.186 \end{aligned}$$

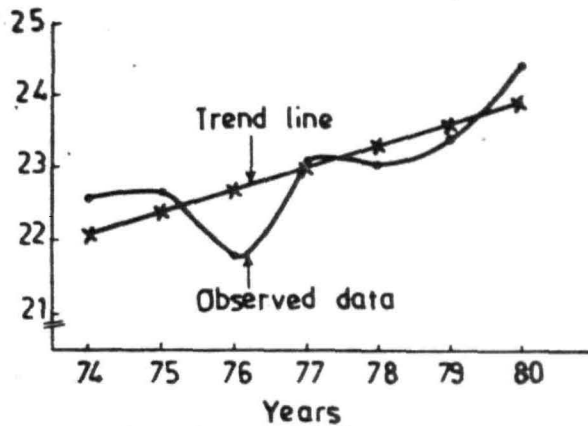


Fig. 6. Linear trend

Example 3

Export of marine products in thousand tons are given below. Fit linear trend line and estimate the exports for 1981.

Year	1975	1976	1977	1978	1979	1980
Quantity Exported ('000 t)	54.56	66.75	65.97	86.89	86.40	75.58

Answer

In this example, number of observations, $n = 6$, therefore two middle most years are taken as $t = -1$ and $t = 1$ and remaining values of t are as shown in the table, so as to make $\sum t = 0$.

Year	Quantity (Exported) ('000t) (Y)	t	tY	t ²
1975	54.46	-5	-272.3	25
1976	66.75	-3	-200.25	9
1977	65.97	-1	- 65.97	1
1978	86.89	1	86.89	1
1979	86.40	3	259.20	9
1980	75.58	5	377.90	25
	436.05	0	185.47	70

$$a = \frac{\sum Y}{n} = \frac{436.05}{6} = 72.675$$

$$b = \frac{\sum tY}{\sum t^2} = \frac{185.47}{70} = 2.65$$

Hence, the trend line is $Y = 72.675 + 2.65 t$. Estimated exports for the year 1981 are

$$Y = 72.675 + 2.65 (7) = 91.225$$

ii) Exponential trend

The linear trend is suitable when the variable under study is changing on an average by equal absolute amounts in each time period. However, if the absolute amount of a variable increases more rapidly in the later time periods than the earlier one, exponential trend will be appropriate instead of linear trend.

Exponential trend is estimated by the

$$Y = ab^t \dots \dots \dots (III)$$

Where Y is the variable under study and t is the time point, a and b are constants. Equation (III) is not in the linear form. However, it can be brought to the linear form by logarithmic transformation assuming multiplicative error model (see 10.4.9.1).

Taking logarithm on both sides of (III) gives,

$$\log Y = \log a + (\log b)t \text{ i.e., } Y^I = A + Bt \dots \dots \dots (IV)$$

Where $Y^I = \log Y$, $A = \log a$, $B = \log b$.

As equation (IV) is in linear form, A and B are estimated as in the case of linear trend (equation I), choosing t such that $\sum t = 0$.

$$A = \log a = \frac{\sum Y^I}{n} = \frac{\log Y}{n}$$

$$B = \log b = \frac{\sum Y^I t}{\sum t^2} = \frac{\sum (\log Y)t}{\sum t^2}$$

Example 4

Fit an exponential trend line for the data given in example 1. Estimate the production for the year 1981.

Answer

Year	Fish prod	$Y^I = \log Y$	t	$Y^I t$	t^2
1974	22.56	1.3534	-3	-4.0602	9
1975	22.66	1.3553	-2	-2.7106	4
1976	21.74	1.3373	-1	-1.3373	1
1977	23.12	1.3640	0	0	0
1978	23.06	1.3628	1	1.3628	1
1979	23.40	1.3692	2	2.7384	4
1980	24.42	1.3878	3	4.1634	9
		9.5298	0	0.1565	28

$$A = \frac{\sum Y^1}{n} = \frac{9.5298}{7} = 1.3614$$

$$B = \frac{\sum Y^1 t}{\sum t^2} = \frac{0.1565}{28} = 0.0056$$

Thus the fitted trend line can be expressed as,

$$\log Y = 1.3614 + 0.0056t$$

To estimate the production for 1981, put $t = 4$

$$\log Y = 1.3614 + .0056xy$$

$$\log Y = 1.3838$$

$$Y = 24.1991$$

iii) Quadratic trend

Straight line may provide trend of a time series reasonably well, for short periods of time. However, for longer periods a curve of some sort may be suitable to describe the trend. A 'Quadratic model', or 'second degree polynomial' or 'parabolic curve' is the simplest of curvilinear models. It is discussed here to estimate the quadratic trend, which is given by,

$$Y = a + bt + ct^2 \dots \dots (V)$$

Where Y is the variable under study and t is the time period. The parameters a, b and c of the equation (V) are estimated using the method of least squares. This is achieved by solving the following 3 normal equations :

$$\sum Y = na + b \sum t + c \sum t^2$$

$$\sum tY = a \sum t + b \sum t^2 + c \sum t^3$$

$$\sum t^2 Y = a \sum t^2 + b \sum t^3 + c \sum t^4$$

If t is chosen such that $\Sigma t = 0$, then the above normal equations reduce to

$$\Sigma Y = na + c \Sigma t^2 \quad (VI)$$

$$\Sigma tY = b \Sigma t^2 \quad (VII)$$

$$\Sigma t^2 Y = a \Sigma t^2 + c \Sigma t^4 \quad (VIII)$$

The parameter 'b' is estimated from equation VII as,

$$b = \frac{\Sigma tY}{\Sigma t^2}$$

The parameters 'a' and 'c' are estimated by solving the equations (VI) and (VIII).

Fitting of quadratic trend is explained with the help of data given in example 1.

Year	Y	t	tY	t ²	t ² Y	t ⁴
1974	22.56	-3	67.68	9	203.04	81
1975	22.66	-2	45.32	4	90.64	16
1976	21.74	-1	21.74	1	21.74	1
1977	23.12	0	0	0	0	0
1978	23.06	1	23.06	1	23.06	1
1979	23.40	2	46.80	4	93.60	16
1980	24.42	3	73.26	9	219.78	81
	160.96	0	8.38	28	651.86	196

$$b = \frac{\Sigma tY}{\Sigma t^2}$$

Parameters a and c are estimated by solving the equations.

$$\Sigma Y = na + c\Sigma t^2$$

$$\Sigma t^2 Y = a \Sigma t^2 + c \Sigma t^4$$

i.e., solve

$$160.96 = 7a + 28c \quad \text{--- (IX)}$$

$$651.86 = 28a + 196c \quad \text{--- (X)}$$

Multiply (IX) by 4 and subtract (X) from it.

$$\text{i.e., } 643.84 = 28a + 112c$$

$$\underline{-651.86 = -28a - 196c}$$

$$\text{i.e., } -8.02 = -84c$$

Hence, $c = 0.0955$

Substituting the value of c in (IX), the value of a is estimated as

$$a = 22.6124$$

Hence, the estimated quadratic trend is

$$Y = 22.6124 + 0.2993t + 0.0955t^2$$

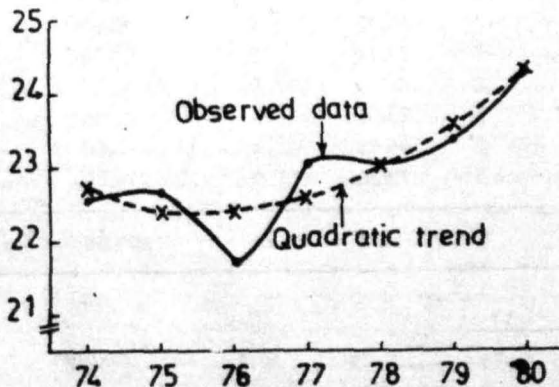


Fig. 7. Quadratic trend

Chapter 14

INDEX NUMBERS

14.1 Introduction

An index number is a statistical device which measures the relative change in the values of groups (set) of related variables at two different situations, time or place. The object of these index numbers is to measure the changes that have occurred in prices, production, cost of living etc. Business-men and economists are able to describe and analyse business and economic situations quantitatively with the help of these indices. That is why index numbers are some times called 'Economic indicators or barometers'. Just as barometers measure atmospheric pressure, index numbers measure changes occurring in economic field. Although index numbers are mainly used in business and economics, they can also be used in other fields.

14.2 Types of index numbers

Index numbers can be broadly classified into 3 categories namely,

1. Price index numbers - to compare the changes in prices
2. Quantity index numbers - to compare the changes in quantity
3. Value index numbers - to compare the changes in value.

14.3 Base period and current period

Index number helps to compare the changes in the values of variables at 2 different periods. For comparison, one of the periods must be a standard one. The period which is considered as standard is known as 'base period', the other one a 'current period' (given period). Usually the price of a commodity in the base period is denoted by p_o and quantity by q_o . In the case of current period, price of a commodity is denoted by p_n , and quantity by q_n . Price index numbers are generally denoted by P_{on} and quantity index numbers by Q_{on} .

14.4 Price relatives

Price relative is defined as the ratio of the price of a single commodity in the current period to its price in the base period.

$$\begin{aligned} \text{Price relative} &= \frac{\text{Price of a commodity in current period}}{\text{Price of a commodity in base period}} \\ &= \frac{P_n}{P_o} \quad \text{usually expressed in \%} \end{aligned}$$

Note that the price relative for a given period relative to the same period is always 100% i.e., the price relative corresponding to base period is always 100. In statistical language 1970 = 100 indicates that the year 1970 is taken as the base period.

Example 1

The price of rohu fish during 1976 and 1980 was Rs.9 and Rs.12 per kg. respectively. Compute price relative taking 1976 as the base year.

Answer

$$\text{Price relative} = \frac{\text{Price in current year}}{\text{Price in base year}} = \frac{12}{9} = 1.33$$

The result indicates that the price of rohu in 1980 was 133% of that in 1976 i.e., it increased by 33%.

14.5 Quantity relatives

Instead of comparing the price of a commodity if we are interested in comparing quantities; we use quantity relative

$$\text{Quantity relative} = \frac{q_n}{q_o}$$

and is usually expressed in %.

Example 2

Export of marine products in our country during the year 1975 and 1980 was 54,463 and 75,583 tonnes respectively. Compute quantity relative taking 1975 as base year.

Answer

$$\begin{aligned} \text{Quantity relative} &= \frac{\text{Quantity exported in the current year (1980)}}{\text{Quantity exported in base year (1975)}} \\ &= \frac{75,583}{54,463} = 1.39\% = 139 \end{aligned}$$

The result indicates that the export of marine products in 1980 has increased by 39%, as compared to 1975.

14.6 Necessity of single index number

Usually the comparisons of prices/quantities of large groups of commodities will be of interest rather than of a single commodity. For example, in computing cost of living index it will be of interest not only to compare prices of milk in one period with another, but also to compare prices of fish, eggs, bread, etc., so as to obtain some general picture. Of course, price relatives of all these commodities can merely be listed. This would not be satisfactory, what is required is a single price index number which would compare over all price change in two periods. Averages such as arithmetic mean and geometric mean are generally used to summarize a large amount of information and to arrive at a single index number. Depending upon the type of average used there are different methods for computing index numbers.

The computational procedures of price or quantity or value index numbers are almost similar. The computation of price index numbers is described here.

14.7 Construction of price index numbers

There are unweighted and weighted price index numbers. Methods of constructing these are described below.

14.7.1 Unweighted price index numbers

14.7.1.1 Simple aggregative method

This is the simplest method of constructing index numbers. This method consists of working out the ratio of sum of prices of commodities of current year to that of sum of prices of commodities of base year. Usually this ratio is expressed in percentage.

If $P_{n1}, P_{n2}, \dots, P_{nk}$ denote the prices of 1st, 2nd, kth commodity in the current year and $P_{o1}, P_{o2}, \dots, P_{ok}$ denote the respective prices in the base year, then simple aggregative index number is given by

$$\begin{aligned} P_{on} &= \frac{P_{n1} + P_{n2} + \dots + P_{nk}}{P_{o1} + P_{o2} + \dots + P_{ok}} \\ &= \frac{\sum P_n}{\sum P_o} \times 100 \end{aligned}$$

Although this method has the advantage of being easy to apply, it has 2 disadvantages which make it unsatisfactory

- (i) It does not take into account the relative importance of different commodities. Thus according to this method, equal importance is given to salt and sugar and fish
- (ii) The particular unit of measurement used in price quotations such as kilograms, quintals etc., affect the value of the index number.

14.7.1.2 Simple average of relatives method

In this method, first the price relatives are computed by dividing the price of the commodity in current year by price of the commodity in the base year. Then these price relatives are averaged to get single index number by using averages like arithmetic mean and geometric mean.

Let $P_{n1}, P_{n2}, \dots, P_{nk}$ denote the price of 1st, 2nd \dots \dots kth commodity in current year and $P_{o1}, P_{o2}, \dots, P_{ok}$ their respective prices in base year.

$$\text{Price relative for 1st commodity is } \frac{P_{n1}}{P_{o1}}$$

$$\text{Price relative for 2nd commodity is } \frac{P_{n2}}{P_{o2}} \quad \text{and so on.}$$

Thus price index number using arithmetic mean as average is

$$\begin{aligned} P_{on} &= \frac{1}{k} \left(\frac{P_{n1}}{P_{o1}} + \frac{P_{n2}}{P_{o2}} + \dots + \frac{P_{nk}}{P_{ok}} \right) \times 100 \\ &= \frac{1}{k} \left(\sum \frac{P_n}{P_o} \right) \times 100 \end{aligned}$$

The price index number using geometric mean as average is

$$P_{on} = \left(\frac{P_{n1}}{P_{o1}} \times \frac{P_{n2}}{P_{o2}} \times \dots \times \frac{P_{nk}}{P_{ok}} \right)^{1/k}$$

$$\log P_{on} = \frac{1}{k} \left(\log \frac{P_{n1}}{P_{o1}} + \log \frac{P_{n2}}{P_{o2}} + \dots + \log \frac{P_{nk}}{P_{ok}} \right)$$

$$= \frac{1}{k} \cdot \left(\sum \log \frac{P_n}{P_0} \right)$$

$$\text{Hence, } P_{on} = \text{Anti log} \left(\frac{1}{k} \cdot \sum \log \frac{P_n}{P_0} \right)$$

It is usually expressed in percentage. The disadvantage of this method is that it does not take into account the relative importance of various commodities.

14.7.2 Weighted index numbers

14.7.2.1 Weighted aggregative method

To overcome the disadvantages of the simple aggregative method, prices of each commodity are assigned weightage by a suitable factor often taken as the quantity of commodity sold during the base year or of current year. Weightage indicates the importance of the particular commodity. The formulae that arise when the base year or current year quantities are used as weights, are discussed below.

i) Laspeyre's index number

In this index number base year quantities are used as weights. If q_0 stands for the quantity of commodity sold during the base year, then Laspeyre's index number is given by

$$P_{on} (\text{Lasp}) = \frac{\sum \frac{P_n q_0}{P_0 q_0}}{\sum \frac{P_0 q_0}{P_0 q_0}} \times 100$$

ii) Paasche's index number

In this index number current year quantities are used as weights. If q_n stands for quantity of commodity sold during current year, the Paasche's index number is given by

$$\text{Pon (Paas)} = \frac{\sum P_n q_n}{\sum P_o q_n} \times 100$$

iii) **Fisher's index number**

It is the geometric mean of Laspeyre's and Paasche's index number. Fisher's index number is given by

$$\begin{aligned} \text{Pon (F)} &= \sqrt{\text{Pon (Lasp)} \text{Pon (Paas)}} \\ &= \sqrt{\frac{\sum P_n q_o \cdot \sum P_n q_n}{\sum P_o q_o \cdot \sum P_o q_n}} \end{aligned}$$

iv) **Marshall - Edge worth index number**

In this index number weights are taken as arithmetic mean of base year and current year quantities i.e.,

$$\text{Weights will be } \left(\frac{q_o + q_n}{2} \right)$$

Marshall-Edge worth price index number is given by

$$\begin{aligned} \text{Pon (M-E)} &= \frac{\sum P_n \left(\frac{q_o + q_n}{2} \right)}{\sum P_o \left(\frac{q_o + q_n}{2} \right)} \times 100 \\ &= \frac{\sum P_n (q_o + q_n)}{\sum P_o (q_o + q_n)} \times 100 \end{aligned}$$

14.7.2.2 Weighted average of relatives method

To overcome the disadvantages of simple average of relatives method weighted average of relatives method is used. Each price relative is assigned weight by the total value of the commodity in terms of some monetary unit such as Rupee. Since the value of the commodity is obtained by multiplying the price p of the commodity by quantity q , the weights are given by pq .

We have 2 formulae depending on whether base year values ($p_0 q_0$) or current year values ($p_n q_n$) are used as weights

1. When base year values are used as weights,

$$P_{on} = \frac{\sum \frac{p_n}{p_0} p_0 q_0}{\sum p_0 q_0} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$$

Which is equivalent to Laspeyre's index number

2. When current year values are used as weights

$$P_{on} = \frac{\sum \frac{p_n}{p_0} (p_n q_n)}{\sum p_n q_n} \times 100$$

$$= \frac{\sum PW}{\sum W} \times 100, \quad \text{where } P = \frac{p_n}{p_0}, W = p_n q_n$$

14.8 Quantity index numbers

Formula for quantity index numbers can be obtained by interchanging p by q and q by p in the formula of price index numbers discussed earlier.

14.8.1 Simple quantity index numbersi) **Simple aggregative method**

$$Q_{on} = \frac{\sum q_n}{\sum q_o} \times 100$$

ii) **Simple average of relatives**

$$Q_{on} = \frac{1}{k} \sum \frac{q_n}{q_o} \times 100$$

14.8.2 Weighted quantity index numbersi) **Laspeyre's quantity index number**

$$Q_{on} (\text{Lasp}) = \frac{\sum q_n p_o}{\sum q_o p_o} \times 100$$

ii) **Paasche's quantity index number**

$$Q_{on} (\text{Paas}) = \frac{\sum q_n p_n}{\sum q_o p_n} \times 100$$

iii) **Fisher's quantity index number**

$$Q_{on} (F) = \sqrt{\frac{(\text{Laspeyre's quantity index number}) (\text{Paasche's quantity index number})}{}}$$

iv) **Marshall - Edge worth quantity index number**

$$Q_{on} (M-E) = \frac{\sum q_n (p_o + p_n)}{\sum q_o (p_o + p_n)} \times 100$$

Example 3

Calculate price index number using Laspeyre's, Paasche's, Marshall Edge worth and Fisher's index number formulae for the following data on selected items of marine products exported during 1979 taking 1975 as the base year.

Item	Price (Rs. 000)		Quantity exported (000 t)	
	1975	1979	1975	1979
1. Frozen shrimp	23.8	41.4	46.5	51.1
2. Frozen froglegs	24.0	21.7	2.0	2.9
3. Frozen cuttle	28.9	27.6	1.2	1.5
4. Fish & fillets Fresh & frozen fish	12.0	5.9	0.3	22.6
5. Dried fish	4.1	4.7	2.4	3.4

Answer

Let P_o & P_n denote price in base and current year respectively and q_o & q_n denote the quantities in base and current year respectively. The following computations have to be made :

Item	P_o	P_n	q_o	q_n	$P_o q_o$	$P_o q_n$	$P_n q_o$	$P_n q_n$
1. Frozen shrimp	23.8	41.4	46.5	51.1	1106.7	1216.2	1925.1	2115.5
2. Frozen froglegs	24.0	21.7	2.0	2.9	48.0	69.6	43.4	62.9
3. Frozen cuttle	28.9	27.6	1.2	1.5	34.7	43.4	33.1	41.4
4. Fresh & Frozen fish	12.0	5.9	0.3	22.6	3.6	271.2	1.8	133.3
5. Dried fish	4.1	4.7	2.4	3.4	9.8	13.9	11.3	16.0
T O T A L					1202.8	1614.3	2014.7	2369.1

i) **Laspeyre's index number**

$$= \frac{\sum P_n q_o}{\sum P_o q_o} \times 100$$

$$= \frac{2014.7}{1202.8} \times 100$$

$$= 1.68 \times 100$$

$$= 168\%$$

ii) **Paasche's index number**

$$= \frac{\sum P_n q_n}{\sum P_o q_n} \times 100$$

$$\begin{aligned}
 &= \frac{2369.1}{1614.3} \times 100 \\
 &= 1.47 \times 100 \\
 &= 147\%
 \end{aligned}$$

iii) **Marshall - Edge worth index number**

$$\begin{aligned}
 &= \frac{\sum P_n (q_o + q_n)}{\sum P_o (q_o + q_n)} \times 100 \\
 &= \frac{\sum P_n q_o + \sum P_n q_n}{\sum P_o q_o + \sum P_o q_n} \times 100 \\
 &= \frac{2014.7 + 2369.1}{1202.8 + 1614.3} \times 100 \\
 &= \frac{4383.8}{2817.1} \times 100 \\
 &= 1.56 \times 100 \\
 &= 156\%
 \end{aligned}$$

iv) **Fisher's index number**

$$\begin{aligned}
 &= \sqrt{(\text{Laspeyre's}) (\text{Paasche's}) \text{ index number}} \\
 &= \sqrt{(1.68) (1.47)} \\
 &= \sqrt{2.4696} \\
 &= 1.57 \\
 &= 157\%
 \end{aligned}$$

14.9 Value index number

The value index number using simple aggregative method is given by

$$V_{on} = \frac{\sum P_n q_n}{\sum P_o q_o} \times 100$$

Where $\sum P_o q_o$ = total value of all commodities in the base period.

$\sum P_n q_n$ = total value of all commodities in the given period (current)

Weighted aggregative value index numbers can be constructed using appropriate weights to indicate the relative importance of the commodities on the lines similar to the construction of weighted aggregative price and quantity index numbers.

14.10 Tests for consistency of an index number

According to Prof. Fisher, a good index number is the one which satisfies the following two tests i) Time reversal test and ii) Factor reversal test.

14.10.1 Time reversal test

According to this test any index number formula to be accurate should be time consistent i.e., the same picture of change in the price level should be obtained if the base period and current period are interchanged. Consider a particular commodity, say, fish. if the price of fish is doubled from 1970 (Rs.4) to 1983 (Rs. 8), then price relative for year 1983 taking 1970 as base year will be 2.00. Similarly price relative for year 1970 taking 1983 as base year will be $1/2 = 0.50$. Thus, one is the reciprocal of the other and the product of 2 (0.5) = 1. This is obviously for each individual commodity and according to this test, it should be true for index number. In symbols this test says $P_{no} = 1$ i.e., the product of index number obtained by interchanging the base year and the current year (subscript o & n) with the original index number should be one.

This test is satisfied by simple aggregative index number, index number based on geometric mean, Marshall - Edge worth index number and Fisher's index number

14.10.2 Factor reversal test

The index number obtained by inter changing the factors p's (prices) and q's (quantities) occurring in a price index number formula, when multiplied by original price index number should give value index number i.e., the product of price index number and quantity index number should be equal to value index number.

i.e., $P_{01} Q_{10} = V_{01}$

Fisher's index number is the only index number which satisfies this test.

As Fisher's index number satisfies both the time and factor reversal tests it is called the ideal index number.

14.11 Cost of living index number (consumer price index number)

The (whole sale) general price index numbers measure variations only in the general level of prices. These variations do not throw light on the effects of rise and fall of prices on the cost (standard) of living of different classes of people. Therefore, to overcome this inability of general price index numbers, special type of index numbers known as 'cost of living index numbers' are computed.

Cost of living index number studies the effect of changes in prices of commodities on the people (consumers). This index number is designed to measure the increase in the cost of expenses of maintaining the same standard of living as that of base year. Since different groups of people consume different types of commodity and that also in different proportion, it becomes necessary to compute separate index number for different groups of people and for different areas. Cost of living index number is therefore always associated with a well defined class or group of people.

Cost of living index number is some times referred to as 'consumer retail price index number'. This index number is of special interest as it is generally used for fixation of salaries and wages of employees and industrial workers.

Most commonly used formula for computation of cost of living index number is Laspeyre's formula which is given by

$$\frac{\sum P_n q_o}{\sum P_o q_o} \times 100$$

Construction of index number by this method necessitates collection of information on the quantities (quantities consumed per average family) of items. But in practice it is difficult to find such average quantities of consumption in all cases. To overcome this difficulty, the following formula is used

$$\frac{\sum \frac{P_n}{P_o} \times 100 (P_o q_o)}{\sum P_o q_o}$$

$$= \frac{\sum PW}{\sum W} \cdot$$

$$\text{Where } P = \frac{P_n}{P_o} \times 100$$

$$W \text{ represents weights} = P_o q_o$$

Example 4

Compute the cost of living index number for the following data :

Answer

Group	Group index $P = \frac{P_n}{P_o} \times 100$	Weights	PW
Food	200	48	9600
Clothes	150	22	3300
Rent	100	12	1200
Fuel	125	10	1250
Miscellaneous	174	8	1400
		100	16750

Answer

Cost of living index number

$$= \frac{\sum PW}{\sum W}$$

$$= \frac{16750}{100}$$

$$= 167.5$$

14.12 Basic requirements in the construction of index numbers

i) The purpose and scope of index numbers

Define clearly the purpose of index numbers, as most of the later problems like commodities to be included, selection of the base period etc., will depend upon the purpose. For instance, if an

index number is being constructed on the cost of living of fishermen, then food and other items should include such consumer goods which are important to this class.

The scope of index numbers refers to area to be covered and time taken in to consideration.

ii) **Selection of the commodities to be included**

It is practically not possible to include all commodities which are available in the market, as it involves more time, money and labour. Therefore, it is necessary to select some important commodities from the available ones. The selection of commodities should be done with great care so that the index number constructed reflects the purpose of its use. It is also necessary that the commodities included should be easily recognisable.

iii) **Sources of collecting required data on price and quantity**

Reliability of an index number depends on the accuracy of data used in its construction. As price of an item varies from place to place and even from shop to shop, it is necessary to select representative places and persons from whom the data have to be obtained. Usually the places where the particular commodity is purchased or sold in large quantities are chosen. These can also be obtained from reports published by the Government departments or from standard trade journals. The Government Departments, Central Statistical Organisation are the major sources of reliable data, besides the Indian Merchants Chambers or Chamber of Trade and Commerce.

iv) **Choice of base period**

The period which is selected as the base year should be economically stable. The period chosen should be free from wars, floods, famines etc. It should not be far away from the current year. If base period is far away from the current year, there is a possibility that the commodity which was very popular during the base year might have become out of taste in the current year. The base period chosen may be a single year or it may be an average of two or three years.

v) **Choice of appropriate weights**

The purpose for which the index number is being constructed is an important consideration in choice of weights. The weights assigned should reflect the relative importance of different items, that are included in the construction of index numbers. Usually the weights chosen for price index numbers are quantities of base year/current year. Some times the value weights are also used.

14.13 Uses of index numbers

1. Index numbers are useful in comparing changes in production, price, imports, exports etc., to study general economic conditions and to plan activities such as production of goods, stock of goods etc.
2. Index numbers are useful to Government in framing economic policies on taxation, imports and exports, grant of licences to new firms, banks etc.
3. Cost of living index numbers are useful in fixation of salary wages and grant of allowances to employees and industrial workers.
4. Index numbers are used as price deflators to measure the real changes in economic magnitudes keeping prices constant.

REFERENCES AND READING LIST

- Berenson, M.L. and D.M. Levine (1983) Basic business statistics. 2nd edition. Prentice - Hall, Inc., Englewood Cliffs, New Jersey
- Cochran, W.G. (1963) Sampling techniques. 2nd edition. John Wiley & Sons Inc., New York.
- Cochran, W.G. and G.M. Cox (1957) Experimental designs. 2nd edition. Asia Publishing house, New Delhi.
- Federer, W.T. (1977) Experimental design. Third Indian reprint (G). Oxford & IBH Publishing Co., New Delhi.
- Fisher, R.A. and F. Yates (1963) Statistical tables for biological, agricultural and medical research. 6th edition. Longman Publication.
- Levin, R.L. and D.S. Rubin (1983) Business Statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
- Nigam, A.K. and V.K. Gupta (1979) Hand book on analysis of agricultural experiments. Indian Agricultural Statistics Research Institute, New Delhi.
- Ostle, B. (1974) Statistics in research. 2nd edition. Oxford & IBH Publishing Co., Bombay.
- Panse, V.G. and P.V. Sukhatme (1978) Statistical methods for agricultural workers. Indian Council of Agricultural Research, New Delhi.
- Pillay, T.V.R. and K.K. Ghosh (1962). The bagnet fishery of the Hooghly-matiah Estuarine System (West Bengal). Ind. J. Fish. 9, 71-79.

- Raghava Rao, D. (1983) Statistical techniques in agricultural and biological research. Oxford & IBH Publishing Co., Bombay
- Ricker, W.E. (1973) Linear regression in fishery research. J. Fish. Res. Board, Canada. 30:409-434
- Shetty, H.P.C. and K.K.Ghosh (1963). On the collection of capture fisheries statistics in the Mahanadi estuary. Ind.J.Fish. 10(A), 48-58
- Singh, D., P.Singh and P.Kumar (1978) Hand book on sampling methods. Indian Agricultural Statistics Research Institute, New Delhi.
- Snedecor, G.W. and W.G.Cochran (1967) Statistical methods. 6th edition. Oxford & IBH Publishing Co., Bombay
- Spiegel, M.R. (1981) Theory and problems of statistics. Schaum Publishing Co., New York.
- Srivastava, U.K., G.V.Shenoy and S.C.Sharma (1985) Quantitative techniques. 2nd reprint. Wiley eastern limited, New Delhi.
- Stockton, J.R. and C.T.Clark (1975) Introduction to business and economics statistics. South-western publishing Co., England.
- Sukhatme, P.V. (1954) Sampling theory of surveys with applications. Ind. Soc. Agril. Stat., New Delhi.
- Wolf, C.M. (1968) Principles of biometry. D.Van Nostrand Co., Inc., Princeton.
- Yamane, T. (1967) Elementary sampling theory. Prentice-Hall international, Inc., London.

DATE DUE

CENTRAL INSTITUTE OF FISHERIES EDUCATION

VERSOVA ANDHERI BOMBAY-400 061.

(ICAR)

1. Books may be retained for a period not exceeding 15 days.
2. Books may be renewed on request at the discretion of the Librarian.
3. Dog-eared the pages of a book, marking or writing therein with ink or pencil, tearing or taking out its pages or otherwise damaging it, will constitute an injury to a book.
4. Any such injury to a book is a serious offence; Unless a borrower points out the injury at the time of borrowing the book, he shall be required to replace the book or pay its price.

Help to keep this book fresh and clean

LIBRARY
CENTRAL INSTITUTE
OF FISHERIES EDUCATION BOMBAY
(ICAR)

CENTRAL INSTITUTE OF FISHERIES EDUCATION
VERSOVA MUMBAI - 400 061.
(ICAR)

1. Books may be retained for a period not exceeding 15 days.
2. Books may be renewed on request at the discretion of the Librarian.
3. Dog-earing the pages of a book, marking or writing therein with ink or pencil, tearing or taking out its pages or otherwise damaging it will constitute an injury to a book.
4. Any such injury to a book is a serious offence; Unless a borrower points out the injury at the time of borrowing the book, he shall be required to replace the book or pay its price.

Help to keep this book fresh and clean

Due Date Slip

**LIBRARY
CENTRAL INSTITUTE
OF FISHERIES EDUCATION, MUMBAI.
(ICAR)**

Acc. No. **7064^D** Class No. **310**

Please return the book on or before the last date stamped. A fine will be charged if the book is not returned in time.

~~14/10/97~~
~~10/12/97~~
~~4/1/98~~
~~30/1/98~~
~~22/98~~
~~13/98~~
~~16/98~~
~~9/5/98~~
~~28/5/98~~
~~13/6/98~~

~~14/5/98~~
~~13/6/98~~
~~9/2/98~~
~~19/2/99~~
~~30/1/99~~
~~28/2/2000~~
~~25/3/2000~~
~~24/3/00~~
~~30/11/2001~~



D7064