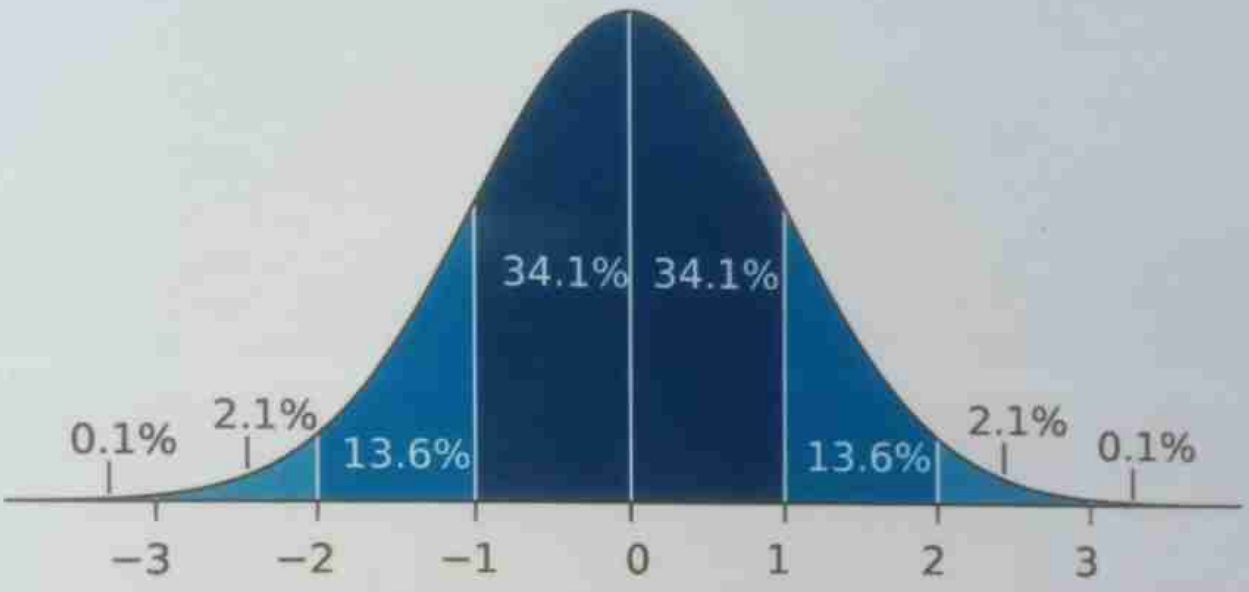
A textbook of
Statistics
For B.S.

محمد بلال مبصر



Muhammad Jamil
Muhammad Abdullah

Statistical thinking will one day be as necessary a qualification for efficient citizenship as the ability to read and write.

H G Wells

ARJ Publicalions
0333 6964525
0331 7707339

Price: Rs. 585

A TEXT BOOK OF

Statistics

BS LEVEL

***MUHAMMAD* JAMIL**
Govt. Graduate College Samundri.
***MUHAMMAD* ABDULLAH**
Govt. Associate College Renala Khurd

All rights reserved.

Published by: Mushtaq printers, Lahore

Printed by:

Edition: 3rd

Price: Rs. 585

Preface

Statistics, as we all know is compulsory in almost all BS levels including natural and social sciences. The chief aim of writing this book is to meet the requirement of the students of B.S. level of various universities in Pakistan.

The text has been written in simple and understandable language that can easily be grasped by students and other readers alike. To make the subject easy, self-explanatory solved examples have been added.

Every effort has been put in to provide latest and up-to-date information of the subject.

Suggestions from all corners would be highly welcomed and incorporated in subsequent editions.

Muhammad Jamil
profjamil786@gmail.com

Muhammad Abduulah
abdullahstat120@gmail.com
02-04-2023.

ACKNOWLEDGEMENT

All Praise to Almighty Allah Who has blessed us with courage and strength to work for the betterment of students of various BS levels in the subject of Statistics.

Our sincerest thanks go to our colleagues and friends, without their enthusiastic support it would have not been possible to bring out the book this year. We also owe a deep debt of gratitude to

Prof. Dr. Nasir Abbas,
Govt. Graduate College, Jhang.

Prof. Dr. Muhammad Kashif,
Department of Mathematics and Statistics, University of Agriculture, Faisalabad

Prof. Dr. Tabassum Sultana,
Principal Govt. Graduate College For Women, D-Type Colony, Faisalabad.

Prof. Dr. Irfan Aslam,
Govt. Islamia Graduate College, Railway Road Lahore

Prof. Muhammad Ashraf,
Govt. Graduate College Samanabad Faisalabd.

Prof. M Anwar Ahmad Bajwa,
Govt. Municipal Graduate College Faisalabad..

Prof. Muhammad Nadeem Jamil,
Govt. Graduate College Jhang

Prof. Muhammad Farooq Yasir,
Govt. Graduate College Gojra.

Prof. Walait Ali Sabir,
Govt. Municipal Graduate College Faisalabad.

Prof. Shahid Majeed,
Govt. Graduate College Toba Tek Singh.

Prof. Muhammad Hanif Rana,
Govt. Graduate College Sahiwal.

Prof. Muhammad Anwar Ali Tahir,
Govt. Graduate College Jhang.

Prof. Muhammad Yousaf Khan,
Govt. Graduate College Bahawalnagar.

Prof. Mrs. Saeeda Bano,
Govt. Emerson College Multan.

Prof. Mrs. Noor ul Huda,
Govt. Graduate College Peoples colony No. 2 Faisalabad.

Prof. Mrs. Asna Butt,
Govt. Graduate College for women Peoples Colony No. 2 Faisalabad.

Prof. Mrs. Rabia Noreen,
Govt. Islamia Graduate College for women Eid Gah Road, Faisalabad.

Prof. Shafaqat Ali,
Department of Mathematics National University of Modern Languages, H9, Islamabad.

Dr. Zahid Bashir,
The University of Faisalabad.

We are also thankful to those to whom we know or do not know but who hav helped us a lot directly or indirectly, we thank all of them.

Contents

Chapter	Title	Page
1	Introduction	7
2	Presentation of the data	31
3	Measures of central tendency	73
4	Measures of dispersion	107
5	Probability	143
6	Probability distributions	177
7	Sampling	205
8	Testing of hypothesis (t and Z test)	235
9	Testing of hypothesis (Chi square test)	265
10	Regression and correlation	277
11	ANOVA and Experimental designs	323
12	Index numbers	349
13	Time series	375
	Additional exercise	388
	Statistical Tables	402

CHAPTER 1

INTRODUCTION

Students of other fields consider statistics as uninterested, non-beneficial and dry subject, due to its non-conceptual teaching.

Learning Goals

(i) Be familiar with the subject of statistics.
(ii) Be familiar with the techniques used for collection of the data.
(iii) Be familiar with the nature and scale of a variable.

1.1 Statistics

Statistics is a multidisciplinary science, concerned with scientific methods for **collecting, organizing (presenting), summarizing**, and **analyzing** of data, as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

1.2 Meaning of statistics

History tells us the word statistics has been derived from the Latin word "Status" or Italian word "Statista" or "German word "Statistik" or the French word "Statistique" each of which means "Political State".

At present there are three different meanings of the word statistics.

01) In singular sense it means that field of study which deals with **collection, presentation, analysis** and **interpretation** of numerical data.

02) In plural sense it refers to numerical facts and figures which are collected by a systematic method.

03) In another sense it is plural of the word statistic, a value computed from a sample.

1.3 Descriptive and inferential statistics

Statistics as a subject may be divided into descriptive and inferential statistics.

1.3.1 Descriptive statistics

The area of statistics that describes and analyses a given set of data without drawing any conclusions or inferences about it, is called descriptive statistics. It is also known as deductive statistics.

1.3.2 Inferential statistics

The branch of statistics that deals with drawing inferences about the characteristics of the population on the basis of sample information is called as inferential statistics. It is also known as inductive statistics.

Inferential Statistics includes the estimation of population parameters and testing of hypothesis. It is based on probability theory.

1.4 (a) Theoretical Statistics

Theoretical statistics is that branch of statistics in which we formulate statistical methods, formulas and rules to find every day problems' solution in the areas of physical and social sciences. It is also called mathematical statistics.

1.4 (b) Applied statistics

Applied statistics is that branch of statistics in which we find everyday problems' solution in the areas of physical and social sciences by using statistical methods.

Statistics adopts different names when it combines with other disciplines of study.

Other disciplines	New name of Statistics	Other fields
Economics	Econometrics	Managerial economics
Life sciences	Biostatistics/biometry	Bioinformatics, Genetics
Chemistry	Chemometrics	Analytical chemistry
Geography	Geoinformatics Geostatistics	
Psychology	Psychometrics	
Physics	Physical statistics	Statistical mechanics

1.4.1 Life sciences and statistics (Biostatistics)

Application of statistical methods to understand the discipline of the life sciences is termed as biostatistics. It is also known as biometry.

Biostatistics/statistics is helpful in understanding the physiology and anatomy. Genetics that is an important area of life sciences is defined as the game of probability. The term regression was originated by Sir Francis Galton (a geneticist) in his paper about inheritance of stature. Bioinformatics based mainly on statistics.

1.4.2 Statistics and economics

Statistical perhaps plays most important role in understanding the economics. Index number, regression analysis, correlation and time series specially are evolved for the sake of economics. Without statistics economics is nothing. Managerial economics and econometrics totally based on application of statistics in economics.

1.4.3 Statistics and chemistry

Chemometrics is application of statistical methods in the field of chemistry. Analytical chemistry is another example for the importance of statistics in chemistry. A researcher or a student of chemistry cannot be independent of factor analysis, principal component analysis and discriminant analysis. Average, measure of dispersion and testing of hypothesis are very important for data analysis of chemistry as in other fields.

1.4.4 Statistics and computer

Nowadays computer technology mainly depends on mathematics and statistics. Especially probability and logistic regression play an important role in artificial intelligence and machine learning process. Well-known features "spell check" and 'auto correction" in internet search engines based on probability theory. These features give better results if Bayesian approach to probability is applied. Data scientist is a hybrid of computer engineer and statistician.

1.4.5 Statistics and geography

Statistics applied in geography and geology is known as Geostatistics. Different methods such as correlation technique is applied to estimate the amount of any substance in a

mines by drilling a borehole. This technique is called "Kriging" named after a South African mining engineer Danie krige.

1.5 Population

The totality of observations with which we are concerned is called a population.
Examples: (i) All BS students in a college, (ii) All the books in a library, (iii) All corona patients in a hospital, (iv) Blood of a body and (iv) All the trees in a forest.

1.5.1 Parameter

Any numerical value calculated from the population is called parameter or population parameter. Parameter is constant. It is denoted by Greek letter as μ, σ, ρ etc.

1.6 Sample

A sample is a representative part of population which is selected to obtain the information concerning the characteristics of a population.
Examples: (i) Punjab food authority picking a handful of wheat from the bags to check the quality, (ii) A spoon of ice cream selected from a pot to taste the flavour and (iii) A syringe of blood for the test of the patient etc.

1.6.1 Representative Sample

If a sample represents all the qualities of a population it is known as representative sample.

1.6.2 Statistic

Any numerical value calculated from the sample is called statistic or sample statistic. Statistic varies from sample to sample therefore, it is variable. It is denoted by Latin letter as $\bar{X}$, S, r etc.
Note: Population and sample sizes are denoted by N and n respectively.

1.7 Importance of statistics

In almost every field of research whether it belongs to physical or social sciences, analysis of data and calculation of uncertainty are very important aspects. Following are the importance of statistics in short to emphasize its importance.

(01) It simplifies complex data to make it easily understandable.
(02) It presents facts in a numerical form.
(03) It facilitates comparison of data.
(04) It studies relationship among different facts.
(05) It helps in making predictions.
(06) It helps to draw inferences about the characteristics of the population by using sample information.
(07) It helps in formulating the policies.
(08) It helps in testing the laws of other sciences.

1.8 Limitations of statistics

Some limitations of statistics are
(01) Statistics only deals with aggregate of facts.
(02) Statistical results are valid (true) only on the average or in the long run.
(03) Statistics deals with facts that can be numerically measured.
(04) A person who has an expert knowledge of statistics can handle statistical data efficiently.
(05) Statistics provides only the tools for analysis.

1.9 Constant

Any value that does not change but remains fixed is called constant.
For example: The number of days in a week, $\pi = 3.1415$, $g = 9.8\ m/sec^2$ and $e = 2.718281$ etc.
The constants are usually represented by first alphabets of the English language as, a, b, c etc.

1.10 Variable

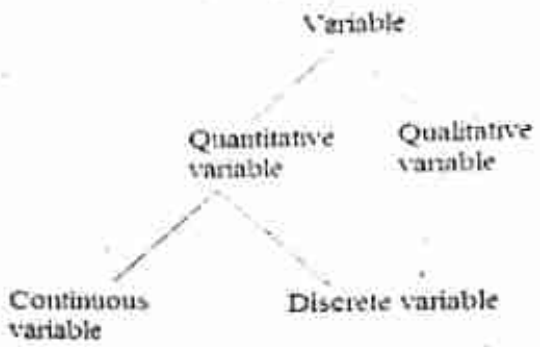
A characteristic that varies from individual to individual, place to place or time to time is called a variable. Variables are usually represented by last alphabets of the English language as X, Y and Z etc.
Examples:
(i) Age of a person
(ii) Weight of a person

03) Weight of a person
04) Eye colour of a person
05) RAM in a computer
06) Volume of a liquid
07) Temperature of a place
08) Number of petals in a flower
09) Demand of a commodity
10) Marital status of an individual
11) Hard disk size of a computer
12) IQ level of a student
13) Behavior of an animal during a specific time interval
14) Utility of an item used
15) Price of a commodity
16) Brand of a computer
17) Income of a person
18) Amount of an agent in the blood of a person.

1.10.1 Types of variable

There are two types of variable

(i) Qualitative variable (ii) Quantitative variable



1.10.2 Qualitative variable

A qualitative variable is determined when characteristics of interest results in a nonnumeric value. It is also called categorical variable or attribute. For example marital status, gender, pain level or personality type etc.

1.10.3 Quantitative variable

A quantitative variable is determined when characteristics of interest results in a numerical value. In other words physical characteristic that varies is known as quantitative variable. It is simply called a variable, for example age, weight, shoe size or temperature etc.

There are two types of quantitative variable

(i) Discrete variable (ii) continuous variable

1.10.4 Discrete variable

A quantitative variable that can assume only specified values is called discrete variable. Examples are shoe size, collar size and number of petals in a flower etc.

1.10.5 Continuous variable

A quantitative variable that can assume any value within a given range is called continuous variable. Examples are temperature, height, weight and age etc.

1.11 Scale of Measurements

Scale of measurement is defined as a classification that tells about the nature of the data / variable. It was developed by Stanley Smith Steven. There are four types of measurement scales.

1.11.1 Nominal scale: Qualitative data/variables are measured on nominal scale. It labels the variable without any quantitative value. For example data on gender, hair colour and caste follow nominal scale because these are just names to categories the data. It is also called categorical variable scale.

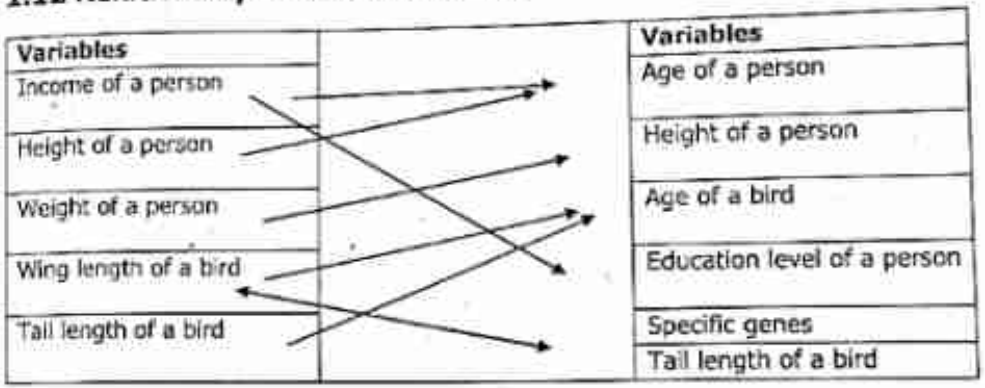
1.11.2 Ordinal scale: Qualitative data/variables with ordering nature are measured on ordinal scale. For example satisfaction level, grade in an examination and education level of a person follow ordinal scale because they can be ordered.

1.11.3 Interval scale: Quantitative data/variables with equal intervals without true zero are measured on interval scale. True zero also known as absolute zero means origin of scale. For example temperature follow interval scale.

1.11.4 Ratio scale: Quantitative data/variables with equal intervals and true zero (No number exists below zero) are measured on ratio scale. For example height, weight and length follow ratio scale.

Scale	Properties
Nominal	It is use for the purpose of only classification.
Ordinal	It is use for the purpose of classification and ranking.
Interval	It is use for the purpose of classification and ranking without true zero.
Ratio	It is use for the purpose of classification and ranking with true zero.

1.12 Relationship between variables



Explanation:

Income of a person depends on (is function of) his age and education level.
Height of a person depends on (is function of) his age and genes.
Weight of a person depends on (is function of) his height.
Wing length of a bird depends on (is function of) his age.
Tail length of a bird depends on (is function of) his age.
Wing length and tail length of a bird depends upon each other.

1.12.1 Dependent variable

A variable that depends on other variable(s) is known as a dependent variable or a variable being tested and measured in a scientific experiment. It is also called as explained, regressand, effect and predictand variable.

1.12.2 Independent variable

A variable that represents the potential reason for change in dependent variable or a variable that is changed or controlled in a scientific experiment to test the effects on the dependent variable is known as an independent variable. It is also called as explanatory, regressor, cause and predictor variable.

1.13 Data

Data is plural of datum which literally means to give or something given. Series of information is called data.

1.13.1 Types of data

With respect to nature

01) Qualitative data.
02) Quantitative data.

With respect to collection method:

01) Primary data.
02) Secondary data.

With respect to formation:

01) Cross sectional data
02) Time series
03) Pooled data

1.13.1.1 Primary data

Data observed or collected directly from first-hand experience is known as primary data. Primary data has not been published yet and is more reliable, authentic and objective. Primary data has not been changed or altered by any one, therefore its validity is greater than secondary data.

1.13.1.2 Sources of primary data

Sources for primary data are limited and at times it becomes difficult to obtain data from primary source due to scarcity of population or lack of cooperation. Following are some of the sources of primary data.

Direct / Indirect Investigation

Experiments: Experiments conducted in laboratories and in any scientific research study are basic sources of primary data.

Survey: Survey is commonly used method in social sciences to collect the primary data. Methods used in survey are:

Questionnaire: A list of open ended or closed ended questions is called questionnaire. It is the most commonly used method in survey.

Interview: A face-to-face conversation with the respondent is called interview. This method can be used for collection of the required data.

Registration: An act of recording a name or information on an official/un-official list is known as registration.

Observations : Recording the values by observing a person/individual, with or without letting him know.

Telephone: Telephone can be used for collecting data in survey.

Internet: In this new era of information technology, internet is widely used in the collection of data.

1.13.1.3 Secondary data

Data collected in the past or from other parties is called secondary data.

1.13.1.4 Sources of secondary data

Secondary data is often readily available. Number of secondary sources are more than primary sources in the collection of data.

Published printed sources

There are varieties of published printed sources

Books: Books on any topic that you want to research are most authentic one in secondary sources.

Journals/Periodicals: Journals are most reliable secondary source. They provide up to date data, not available in books, regarding a field of study.

Magazines/Newspapers

Magazines are easily available source, used for collection of secondary data. It is not a reliable source.

Published electronic sources
Internet can be used as both primary and secondary source for collection of data. Data can be collected from general websites, weblogs and journals.

General websites: Generally websites are easily reachable secondary sources, but not reliable.

Weblogs: Weblogs are electronic diaries written by an individual. They are reliable to some extent they are like personal diaries..

E-journals: E-journals are more commonly available than printed journals. Some e-journals are free of cost.

Unpublished personal records
Unpublished personal records including **diaries** and **letters** are also reliable sources of secondary data. They may also be useful in some cases.

Official / Government records
Government records are very important and reliable secondary sources. They include census data/population statistics, health records and educational institutions records.

Private sector records
They includes banks, chamber of commerce and industries etc.

1.13.1.5 Cross sectional data

Data that comprise observations about a variable for different groups or places in same time interval. For example prices of a commodity for various cities of Pakistan in 2017.

City	Milk Price (2017) in Rs
Lahore	120
Faisalabad	90
Okara	85

1.13.1.6 Time series

Data that comprise observations about a variable for same groups or places in different regular time intervals. For example prices of a commodity for the years from 2015 to 2019 in a city.

Year	Milk price (Rs) in Faisalabad
2015	65
2016	70
2017	85
2018	90
2019	100

1.13.1.7 Pooled data

A data that is mixture of cross sectional data and time series is known as pooled data. For example enrollment of students in two different colleges from 2016 – 2018

Year	College 1	College 2
2016	1200	800
2017	1350	900
2018	1400	920

1.14 Error of measurements

Errors in reading, calculating or recording a numerical value. The difference between observed values of a variable recorded under similar conditions and some fixed true value. **[Cambridge Dictionary of Statistics 4th Edition]**

There are two types of error (i) biased error and (ii) unbiased error

1.14.1 Biased Error: An error that occurs due to faulty measuring device. It is also known as instrumental, cumulative, systematic error. It does affect the average.

1.14.2 Unbiased error: An error that occurs on the result of repeated measurements and will disappear in the long run. Error caused by a factor that randomly affects the measurement, such as change in environment, noise, tiredness or human mistake. It is also known as residual error, random error, compensating error or accidental error. It does not affect the average.

1.15 Significant digits

The digits in a number are called significant digits which give exact and important information. All the digits are significant other than zero. There are some rules for the zero to be significant. To understand the concept of significant digits the table is given below.

Number	27	632	4500	30	7.000	0.36	235.67	0.006	0.039	3.4000
Significant digits	2	3	2	1	4	2	5	1	2	5

1.16 Rounding off numbers

The procedure used for reporting numerical information to fewer decimal places that used during analysis. The rule generally adopted is that excess digits are simply discarded if the first of them is less than five, otherwise the last retained digit is increased by one (replace with zeros in case of whole numbers). The rules for rounding decimal numbers are given below

01) 127.2492341 to three decimal places gives 127.249

02) 25.64682 to two decimal places gives 25.65

If exactly digit 5 is to be dropped then, retained last digit will be increased by one if it is odd and will be remained unchanged if it is even for example

03) 6.465 to two decimal places gives 6.46

04) 6.435 to two decimal places gives 6.44

Multiple Choice Questions

1. Statistics is a science of:
(a) Sources (b) Decision making (c) Collection (d) All

2. Statistics can be divided into branches:
(a) 1 (b) 2 (c) 3 (d) 4

3. A quantity calculated from population:
(a) Frequency (b) Statistic (c) Parameter (d) Sample

4. Measurement provides:
(a) Qualitative data (b) Discrete data (c) Primary data (d) Continuous data

5. Statistics are always:
(a) Exact (b) Estimated values (c) Constant (d) Population

6. A constant can assume values
(a) Fixed (b) Not fixed (c) Variable (d) Grouped

7. In which sense statistics mean numerical data:
(a) Singular (b) Plural (c) Both (a) and (b) (d) None of these

8. Sum of random errors is equal to:
(a) 3 (b) 2 (c) 1 (d) 0

9. The data in their original form:
(a) Secondary data (b) Ordered data (c) Ungrouped data (d) unofficial data

10. The data which have already been collected:
(a) Secondary data (b) Primary data (c) Ungrouped data (d) grouped data

11. A representative part of the population:
(a) Sample (b) Parameter (c) Statistic (d) average

12. Weights of students in a class make:
(a) Discrete data (b) Continuous data (c) Qualitative data (d) Constant data

13. The branch of statistics that deals with analysis of data.
(a) Descriptive statistics (b) Inferential statistics
(c) Biostatistics (d) Biometry

14. The branch of statistics which deals with procedures of drawing inference about population on the basis of sample information:
(a) Descriptive statistics (b) Inferential statistics
(c) Applied statistics (d) Theoretical statistics

15. Statistics deals with:
(a) Qualitative facts only (b) Single facts (c) Aggregate of facts (d) None

16. Statistics are ---------- of the administration:
(a) Ears (b) Eyes (c) Mouth (d) Hands

17. A collection of all the elements in a group is called:
(a) Population (b) Sample (c) Data (d) Statistic

18. Technical and trade journals:
(a) Primary source (b) Secondary source
(c) Unpublished source (d) electronic source

19. Questionnaire is:
(a) Primary source (b) Secondary source
(c) Published source (d) Official source

20. The word "Statistics" is defined in:
(a) Singular sense (b) plural sense
(c) Plural of the word statistic (d) All of the above

21. The data collected by NADRA:
(a) Un-official data (b) Primary data (c) Secondary data (d) Qualitative data

22. Statistics is the backbone of :
(a) Mathematics (b) Computer (c) Economics (d) Research

23. Counting usually provide:
(a) Continuous data (b) Discrete data (c) Primary data (d) Qualitative data

24. A characteristic that cannot be expressed numerically:
(a) Continuous variable (b) Quantitative variable
(c) Attribute (d) discrete variable

25. A characteristic that changes from one individual to another:
(a) Statistic (b) parameter (c) Constant (d) variable

26. Village patwari collecting data about crops:
(a) Primary data (b) Secondary data (c) Qualitative data (d) Published data

27. Statistical laws are true for:
(a) Short run (b) Long run (c) Both (a) and (b) (d) None of these

28. A numerical value calculated from sample:
(a) Parameter (b) Statistic (c) Population (d) sample

29. A constant can assume ----- value(s):
(a) Different b) More than one c) Only one d) No value at all

30. Hourly temperature recorded by weather bureau:
(a) Discrete data (b) Qualitative data (c) Secondary data (d) Continuous data

31. Registration is the source of:
(a) Secondary data (b) Primary data (c) Published data (d) electronic data

32. The number of road accidents on M-1 during a month:
(a) Discrete variable (b) Continuous variable
(c) Attribute (d) categorical variable

33. Statistics has handicap dealing with:
(a) Qualitative data (b) Quantitative data (c) Discrete data (d) All of the above

34. A variable that makes measurable values:
(a) Constant (b) Discrete variable (c) Qualitative variable (d) Continuous variable

35. Identify the attribute:
(a) Ages of patients (b) Temperature of a room
(c) Weights of children (d) Hobbies of students

36. Which of the following is not the example of continuous variable:
(a) Smoking habits of college students (b) The hobbies of students
(c) The amount of rain fall in Muree (d) Both (a) and (b) but not (c)

37. Proportion becomes percentage when multiplied by:
(a) 1/10 (b)1/100 (c) 10 (d) 100

38. Inferential statistics deals with:
(a) Collection of the data (b) Analysis of the data
(c) Presentation of the data (d) Interpretation of the data

Key

Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans
1	b	2	b	3	c	4	d	5	b	6	a
7	b	8	d	9	c	10	a	11	a	12	b
13	a	14	b	15	c	16	b	17	a	18	b
19	a	20	d	21	b	22	d	23	b	24	c
25	d	26	a	27	b	28	b	29	c	30	d
31	b	32	a	33	a	34	d	35	d	36	d
37	d	38	d								

Exercise

Q No.1.1: Define statistics and its types descriptive and inferential statistics.

Q No.1.2: Explain the three different meanings of the word statistics.

Q No.1.3: Differentiate the followings terms.

(i) Qualitative and quantitative variable.
(ii) Discrete and continuous variable.
(iii) Primary and secondary data.
(iv) Cross sectional data and time series.
(v) Interval and ratio scale of measurements.

Q No. 1.4: Classify following variables as Attribute, Discrete and continuous, also name scale of measurement.

Variable	Type	Scale
Meal preference		
Political orientation		
Family size		
Income of a person		
Cost of a commodity		
Room number in a hotel		
Germination percentage		
Survival percentage		
Tree height		
Tree growth		
Stem form		
Biomass weight		
Crop yield		
Soil fertility		
Level of essential elements		

Q No.1.5: Define the followings.
Variable, Population, Sample, Parameter and Statistic.

Q No.1.6: Write the names of primary and secondary sources related to the collection of the data.

Q No.1.7: What is measurement scale? Identify type and measurement scale for the following variables.

Variable	Type	Scale
Age of person		
Height of a person		
Religion of a person		
Marks obtained by a student		
GPA of a student		
Milk produced by a cow		
Yield of a crop		
Weight of an animal		
No of petals in a flower		
Fertilizer used in a field		
Quality of a seed		
Fertility of a land		
Grade of a student		
No of prayers offered by a person		
Hair colour of a person		

Q No.1.8: From the following table of variables indicate the relationship.

Variable		Variable
Shoes size of a person		Age of a person
Heat		Height of a person
Yield of a crop		Age of a bird
		Quantity of fertilizer used
		Temperature

Solution

Q No. 1.4: Classify following variables as Attribute, Discrete and continuous, also name scale of measurement.

Variable	Type	Scale
Meal preference	Attribute	Nominal
Political orientation	Attribute	Nominal
Family size	Discrete	Ratio
Income of a person	Continuous	Ratio
Cost of a commodity	Continuous	Ratio
Room number in a hotel	Attribute	Nominal
Germination percentage	Continuous	Ratio
Survival percentage	Continuous	Ratio
Tree height	Continuous	Ratio
Tree growth	Continuous	Ratio
Stem form	Attribute	Nominal
Biomass weight	Continuous	Ratio
Crop yield	Continuous	Ratio
Soil fertility	Attribute	Ordinal
Level of essential elements	Attribute	Ordinal

Q No.1.7: What is measurement scale? Identify type and measurement scale for the following variables.

Variable	Type	Scale
Age of person	Continuous	Ratio
Height of a person	Continuous	Ratio
Religion of a person	Attribute	Nominal
Marks obtained by a student	Continuous	Ratio
GPA of a student	Continuous	Ratio
Milk produced by a cow	Continuous	Ratio
Yield of a crop	Continuous	Ratio
Weight of an animal	Continuous	Ratio

No of petals in a flower	Discrete	Ratio
Fertilizer used in a field	Continuous	Ratio
Quality of a seed	Attribute	Ordinal
Fertility of a land	Attribute	Ordinal
Grade of a student	Attribute	Ordinal
No of prayers offered by a person	Discrete	Ratio
Hair colour of a person	Attribute	Nominal

Q No.1.8: From the following table of variables indicate the relationship.

Variable	Variable
Shoes size of a person	Age of a person
Heat	Height of a person
Yield of a crop	Age of a bird
	Quantity of fertilizer used
	Temperature

Q No.1.9: From the following table of variables indicate the relationship.

Variable	Variable
Price of a commodity	Experience of a person
Quantity sold	Quantity sold
Income of a person	Income of a person
Profit	Profit
Investment	Investment
Quantity demanded	Quantity demanded

CHAPTER 2

PRESENTATION OF DATA

Drawing and interpreting the graphs are the first step in data analysis

Learning Goals

(i) To develop the skills for arranging the data.

(ii) To represent the data graphically.

2.1 Presentation of the data

Presentation of data simplifies the complex data to make it easily understandable. There are three methods of presentation of data.

(i) Classification

(ii) Tabulation

(iii) Graphical display

2.1.1 Ungrouped / Raw data

Data collected in original form is called raw data or ungrouped data.

2.1.2 Classification

Classification is the process of arranging data into classes on the basis of certain properties.

2.1.3 Distribution

Arrangement of data according to the values of a variable characteristic is called a distribution.

2.1.4 Frequency distribution

A frequency distribution is a table that divides observations in the data set into classes (groups or categories) along with their frequencies.

2.1.5 Grouped data

A data in the form of frequency distribution is known as grouped data.

2.1.6 Discrete frequency distribution

Arrangement of discrete data in a table, such that, against each observation its frequency is given is called discrete frequency distribution, for example

X	1	2	3	4	5
f	6	8	4	12	10

2.1.7 Bivariate frequency distribution

A frequency distribution in which we have observations in pairs and each pair is obtained by observing two variables at a time is called bi-variate frequency distribution, for example

Age(Years)	Height(inches)		
	21-30	31-40	41-50
1-10	14	22	-
11-20	28	35	11
21-30	15	22	28

2.1.8 Constructing a frequency distribution

The important steps involved in construction of frequency distribution are given below:

(i) Find the range of data as
Range = maximum value - minimum value

(ii) Decide on the number of classes into which the data are to be grouped.

(iii) Determine the suitable class width (Class interval size) as

$$Class\ interval = \frac{Range}{No.of\ classes}$$

(iv) Write the class limits in such a way that the smallest observation is absorbed in the first class and highest value in the last class.

(v) All the observations are put into respective classes (It may be done by using tally marks) and determine the frequency of each class.

(vi) Finally, total the frequency column to see that all the data have been accounted for.

Example 2.1: Marks of 50 students of BS Mathematics in a test are given below make a frequency distribution.
51, 56, 91, 96, 54, 53, 58, 57, 57, 81, 68, 75, 89, 88, 84, 87, 82, 78, 75, 74, 75, 62, 82, 62, 95, 76, 68, 95, 60, 53, 90, 78, 79, 74, 62, 63, 83, 86, 88, 72, 71, 76, 66, 79, 93, 78, 62, 65, 61, 57

Solution: 51, 56, 91, 96, 54, 53, 58, 57, 57, 81, 68, 75, 89, 88, 84, 87, 82, 78, 75, 74, 75, 62, 82, 62, 95, 76, 68, 95, 60, 53, 90, 78, 79, 74, 62, 63, 83, 86, 88, 72, 71, 76, 66, 79, 93, 78, 62, 65, 61, 57

No. of values = n =50, Range = $X_m - X_0 = 96 - 51 = 45$

No. of classes = 10, $Class\ interval = \frac{Range}{No.of\ classes} = \frac{45}{10} = 4.5 = 5$

Frequency distribution

Classes	Tally	Frequency
50 – 54	IIII	4
55 – 59	⧸⧸⧸⧸	5
60 – 64	⧸⧸⧸⧸ II	7
65 – 69	IIII	4
70 – 74	IIII	4
75 – 79	⧸⧸⧸⧸ ⧸⧸⧸⧸	10
80 – 84	⧸⧸⧸⧸	5
85 – 89	⧸⧸⧸⧸	5
90 – 94	III	3
95 – 99	III	3
		$\sum f_i = 50$

$n = \sum f_i$

2.1.9 Frequency

The number of observations in each class is referred to as frequency denoted as f_i OR The number of times a certain value or class of values occurs.

2.1.10 Relative frequency

The frequency of a class divided by the total frequency of all the classes is called the relative frequency.

$$Relative\ Frequency = \frac{f_i}{\sum f_i};$$ The total of relative frequency is unity.

2.1.11 Relative frequency distribution

A table showing relative frequency is called relative frequency distribution.

2.1.12 Percentage relative frequency

If relative frequency is multiplied by 100, we obtain percentage relative frequency. A table showing percentage frequencies is called as percentage frequency distribution.

2.1.13 Cumulative frequency

The total frequency of all the classes less than the upper class boundary of a given class is called the cumulative frequency of that class.

2.1.14 Cumulative frequency distribution

An arrangement of different class boundaries with their respective cumulative frequencies is called cumulative frequency distribution. If the cumulative frequency distribution is made below than the class boundaries it is called less than type cumulative frequency distribution. If the cumulative is made above than the class boundaries it is called more than type cumulative frequency distribution.

Note:

(i) Before forming a cumulative frequency distribution class limits are converted into class boundaries.

(ii) If it is not mentioned about less than or more than type cumulative frequency distribution than we always mean less than type cumulative frequency distribution.

(iii) It is used to obtain the partition values such as median, quartiles, percentiles etc.

Example 2.2: Convert the frequency distribution made in above example 2.1 into relative frequency distribution and cumulative frequency distribution.

Solution:

Relative frequency distribution

Classes	f	rf
50 – 54	4	$\frac{4}{50} = 0.08$
55 – 59	5	0.10
60 – 64	7	0.14
65 – 69	4	0.08
70 – 74	4	0.08
75 – 79	10	0.20
80 – 84	5	0.10
85 – 89	5	0.10
90 – 94	3	0.06
95 – 99	3	0.06
	$\Sigma f = 50$	

Cumulative frequency distribution

Classes	f_i	CF
50 – 54	4	4
55 – 59	5	4+5=9
60 – 64	7	9+7=16
65 – 69	4	20
70 – 74	4	24
75 – 79	10	34
80 – 84	5	39
85 – 89	5	44
90 – 94	3	47
95 – 99	3	50
	$\sum f_i = 50$	

Example 2.3: Make a frequency distribution of GPA obtained by BS students of GGC Samundri in 1st semester 3.6, 2.6, 2.7, 3.7, 4.0, 2.5, 2.7, 2.2, 3.2, 3.8, 3.5, 3.5, 2.2, 3.9, 4.0, 3.6, 3.8, 3.5, 3.7, 3.9, 3.4, 2.9, 3.9, 3.2, 2.1, 3.6, 3.0, 4.0, 3.5, 2.5, 2.2, 3.4, 3.6, 2.6, 2.4, 2.8, 3.4, 3.0, 2.1, 3.2

Solution:

No. of values = n =40, Range = $X_m - X_0 = 4.0 - 2.1 = 1.9$

No. of classes = 5, Class interval = $\frac{Range}{No.of\,classes} = \frac{1.9}{5} \approx 0.4$

Frequency distribution

Classes	Tally	Frequency
2.1 – 2.4	𝍸 I	6
2.5 – 2.8	𝍸 II	7
2.9 – 3.2	𝍸 I	6
3.3 – 3.6	𝍸 𝍸 I	11
3.7 – 4.0	𝍸 𝍸	10
		$\sum f_i = 40$

$n = \sum f_i$

Example 2.4: Auto sale by 40 sale persons over month period is given below. Make a frequency distribution for this discrete set of data.

7, 2, 8, 5, 8, 1, 2, 6, 5, 6, 8, 9, 1, 5, 8, 7, 9, 1, 10, 4, 10, 1, 5, 8, 5, 3, 7, 7, 2, 5, 6, 5, 8, 9, 8, 5, 6, 3, 8, 4.

Solution: No. of values = 40, Range = $X_m - X_0 = 10 - 1 = 9$

Frequency distribution

NO. Auto Sale	Tally	Frequency
1	IIII	4
2	III	3
3	II	2
4	II	2
5	⧻ III	8
6	IIII	4
7	IIII	4
8	⧻ III	8
9	III	3
10	II	2
		$\sum f_i = 40$

$n = \sum f_i$

2.1.15 Open end frequency distribution

When the lower limit of the first class or upper limit of the last class or both is unknown then the frequency distribution is called open end frequency distribution.

2.2 Tabulation of data

Systematic presentation of data into rows and columns is called tabulation.

TITLE

Prefatory Notes

Row Captions	Box Head Column Captions		
Stub	Body		

Foot Note

Source Note

2.2.1 The Title

A title is the main heading written in capital letters and shown at the top of the table. It must explain the contents of the table and throw light on the table as whole different parts of the heading can be separated by commas there are no full stop be used in the title.

2.2.2 The Box Head (column captions)

The vertical heading and subheading of the column are called columns captions. The spaces where these column headings are written is called box head. Only the first letter of the box head is in capital letters and the remaining words must be written in small letters.

2.2.3 The Stub (row captions)

The horizontal headings and sub heading of the rows are called rows captions and the space where these rows headings are written is called stub.

2.2.4 The Body

It is the main part of the table which contains the numerical information classified with respect to rows and columns captions.

2.2.5 Prefatory Notes

A statement given below the title and enclosed in brackets, usually describe the units of measurement is called prefatory notes.

2.2.6 Foot Notes

It appears immediately below the body of the table providing the further additional explanation.

2.2.7 Source Notes

The source notes is given at the end of the table indicating the source from where information has been taken. It includes the information about compiling agency, publisher etc.

2.2.8 General rules of tabulation

01) A table should be simple and attractive. There should be no need of further explanations.
02) Proper and clear headings for columns and rows should be need.
03) Suitable approximation may be adopted and figures may be rounded off.
04) The unit of measurement should be well defined.
05) Thick lines should be used to separate the data under big classes and thin lines to separate the sub classes of data.

Example 2.5: There are total 185 universities providing their services in both public and private sectors of education in Pakistan. Out of these universities, 110 (59%) are working under umbrella of public sector, whereas 75 (41%) are working under the supervision of private sector. The total enrolment in the universities, i.e., at post graduate stage, is 1.463 million. Out of this enrolment 1.192 million (81%) students are enrolled in public universities, whereas, 0.270 million (19%) students are studying in private universities. The total teachers in the universities are 58,733 out of which 40,258 (69%) are in public and 18,475 (31%) are in private sector. Arrange the data mentioned in this paragraph in the form of a table.

[Pakistan education statistics 2016-17]

Solution:

UNIVERSITIES IN PAKISTAN
(Enrollments and Teachers)

University Type	Number of Universities	Enrollment (millions)	Teachers
Public	110 (59%)	1.192 (81%)	40258 (69%)
Private	75 (41%)	0.270 (19%)	18475 (31%)
Total	185	1.463	58733

[Pakistan education statistics 2016-17]

2.3 Diagrams

The important types of diagrams are:

(i) Simple bar chart
(ii) Multiple bar chart

(iii) Component bar chart
(iv) Pie chart

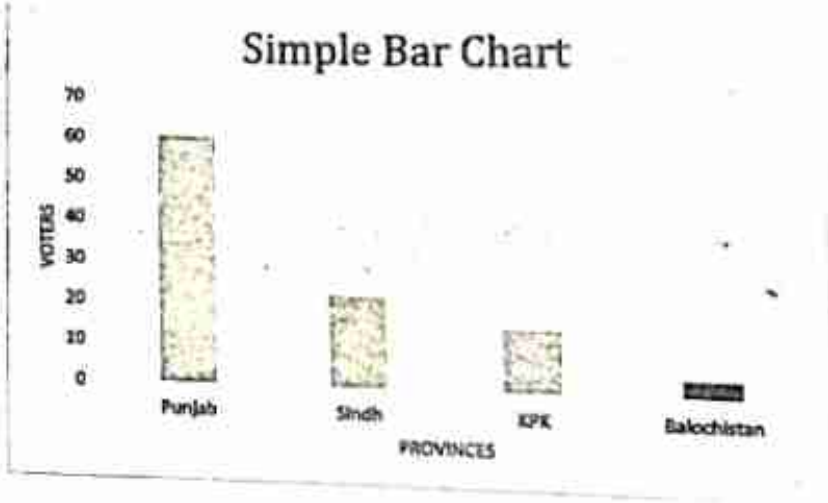
2.3.1 Simple bar chart

This chart consist of equally spaced vertical or horizontal bars of the same width, having heights proportional to the corresponding numerical values.

Example 2.6: Draw a simple bar chart of total number of voters (millions) in four provinces of Pakistan for general election 2018.

Province	Voters
Punjab	60.67
Sindh	22.39
KPK	15.32
Balochistan	4.30

Solution:



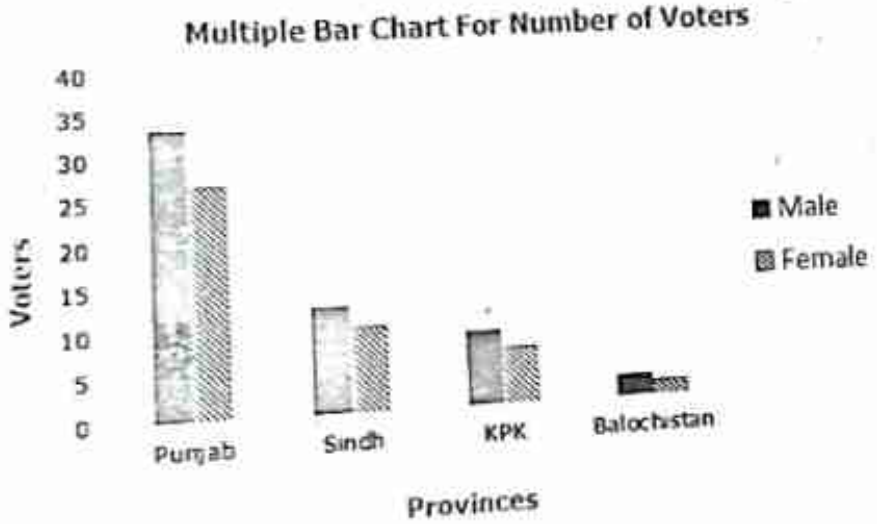
2.3.2 Multiple bar chart

This chart is use to represent two or more related sets of data having some common characteristics in variable values. It is an extension of simple bar chart.

Example 2.7: Draw a multiple bar chart of total number of voters (millions) in four provinces of Pakistan for general election 2018.

Province	Male	Female
Punjab	33.68	26.99
Sindh	12.44	9.95
KPK	8.71	6.61
Balochistan	2.49	1.81

Solution:



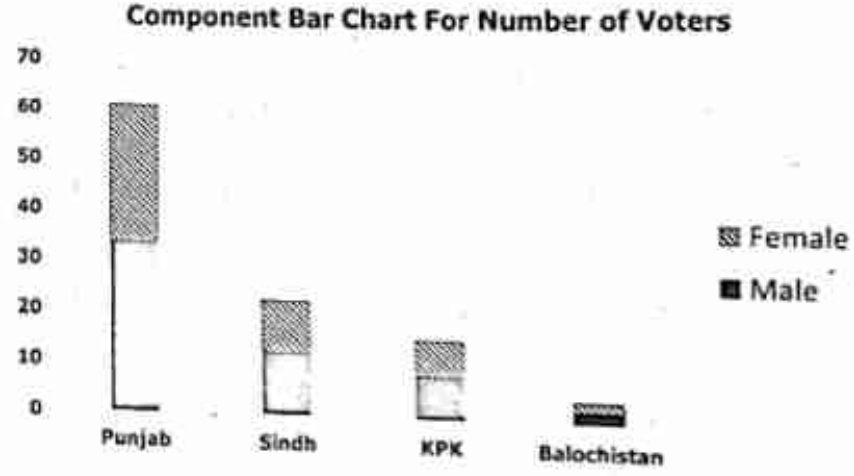
2.3.3 Component bar chart

A bar chart that shows the component parts of the aggregate represented by the total length of the bar. The component parts are shown as sectors of the bar with lengths in proportion to their relative size.

Example 2.8: Draw a component bar chart of total number of voters (millions) in four provinces of Pakistan for general election 2018.

Province	Male	Female
Punjab	33.68	26.99
Sindh	12.44	9.95
KPK	8.71	6.61
Balochistan	2.49	1.81

Solution:



2.3.4 Pie chart

Pie chart consists of a circle and is divided into a number of sectors. The area of a sector of circle is proportional to the angle of the sector. The angle can be calculated as:

$$\text{Angle of sector} = \frac{\text{Component part}}{\text{Total}} \times 360^0$$

A pie diagram is also known as **sector diagram.** It is used to visualize the proportions in excellent way. It is suitable for small number of categories.

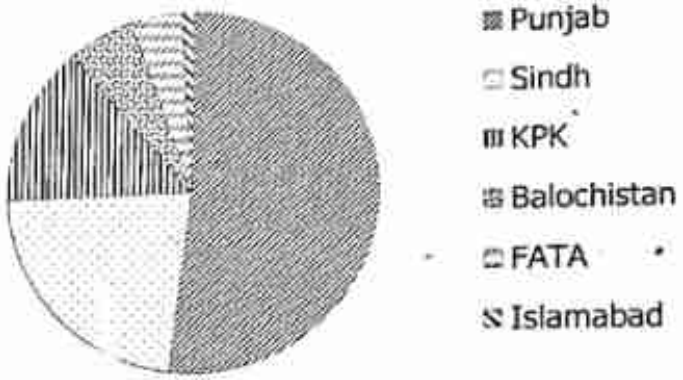
Example 2.9: Draw a pie chart of the data about number of general seats in National Assembly of Pakistan for provinces, FATA and Islamabad in general election 2018.

Area	Number of general seats
Punjab	141
Sindh	61
KPK	39
Balochistan	16
FATA	12
Islamabad	3
Total	272

Solution:

Area	Number of general seats	Angle of sector	Cumulative Angle of sector
Punjab	141	187^0	187^0
Sindh	61	80^0	267^0
KPK	39	52^0	319^0
Balochistan	16	21^0	340^0
FATA	12	16^0	356^0
Islamabad	3	4^0	360^0
Total	272		

Pie Chart



2.3.6 Class limits

Class limits are the smallest and largest observations in each class. Smallest and largest values are called lower and upper class limits respectively.

2.3.7 Class Intervals

Class interval is the difference between two consecutive lower or upper class limits/boundaries of any class.

2.3.8 Class-mark/ Mid-value/Mid-point

The class-mark or, mid-point of a class is arithmetic mean between lower and upper class limits of that class.

$$Mid-Point = \frac{Lower\ Class\ Limit + Upper\ Class\ Limit}{2}$$

2.3.9 Class boundaries

Class boundaries are the true-limits of a class. It is associated with grouped frequency distribution, class boundary is a mid-point between the upper class-limit and the lower class-limit of the next class.

2.3.10 Construction of a "Ogive" (Cumulative frequency polygon)

The construction of a "Ogive" involves the following steps:

01) Compute the cumulative frequencies less than the upper class boundaries.
02) Mark off the upper class boundaries along x-axis.
03) Using an appropriate scale, plot the cumulative frequencies against the upper class boundaries, join the resulting points by straight line.
04) As this graph does not meet the x-axis, add one class with zero frequencies at the lower end of the frequency distribution and extend this graph to meet the x-axis. Also join the point of the graph with the last upper class boundary to get a polygon.

Example 2.10 Insert midpoint, class boundaries and cumulative frequency in the following frequency distribution.

Classes	50 – 54	55 – 59	60 – 64	65 – 69	70 – 74
f_i	4	5	7	4	4
Classes	75 – 79	80 – 84	85 – 89	90 – 94	95 – 99
f_i	10	5	5	3	3

Solution:

Classes	f_i	Midpoint (X)	CB	cf
50 – 54	4	$\frac{50+54}{2}=52$	49.5-54.5	4
55 – 59	5	57	54.5-59.5	9
60 – 64	7	62	59.5-64.5	16
65 – 69	4	67	64.5-69.5	20
70 – 74	4	72	69.5-74.5	24
75 – 79	10	77	74.5-79.5	34
80 – 84	5	82	79.5-84.5	39
85 – 89	5	87	84.5-89.5	44
90 – 94	3	92	89.5-94.5	47
95 – 99	3	97	94.5-99.5	50
$\sum f_i = 50$				

2.3.11 Histogram

A graph that displays the data by using vertical bars of various heights to represent frequencies is called histogram. The horizontal axis can be the class boundaries.

Histograms can provide insights on skewness, behavior in the tails, presence of multi-modal behavior, and data outliers; histograms can be compared to the fundamental shapes associated with standard analytic distributions.

2.3.12 Histogram and Historigram

Graph of a frequency distribution is called histogram, whereas graph of a time series is called Historigram.

2.3.13 Frequency polygon

A line graph in which the frequency is placed along the vertical axis and the class midpoints are placed along the horizontal axis is called frequency polygon. These points are connected with lines. Two points are taken lower and upper sides with zero frequencies to close the figure.

Example 2.11: Make histogram, frequency polygon, frequency curve and Ogive from the following frequency distribution.

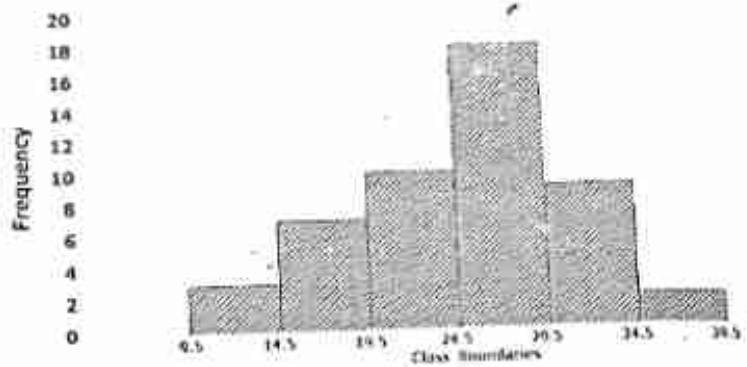
Classes	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39
Frequency	3	7	10	18	9	2

Solution:

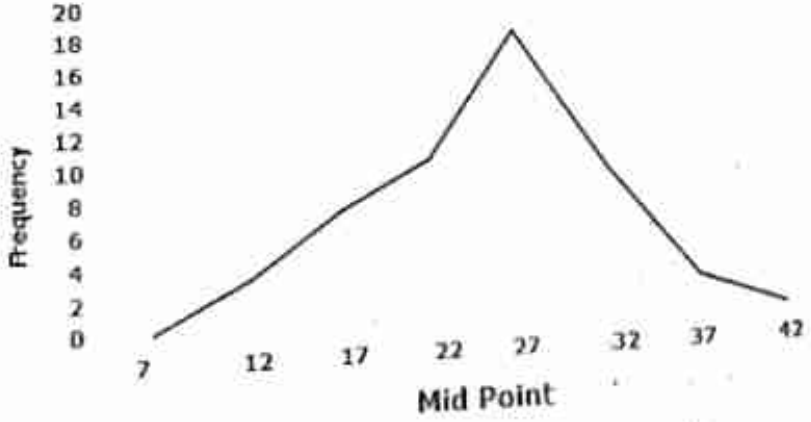
For histogram, frequency polygon and Ogive we make following table

Classes	f	C.B	Midpoint (X)	cf
10 – 14	3	9.5 – 14.5	12	3
15 – 19	7	14.5 – 19.5	17	10
20 – 24	10	19.5 – 24.5	22	20
25 – 29	18	24.5 – 29.5	27	38
30 – 34	9	29.5 – 34.5	32	47
35 – 39	2	34.5 – 39.5	37	49

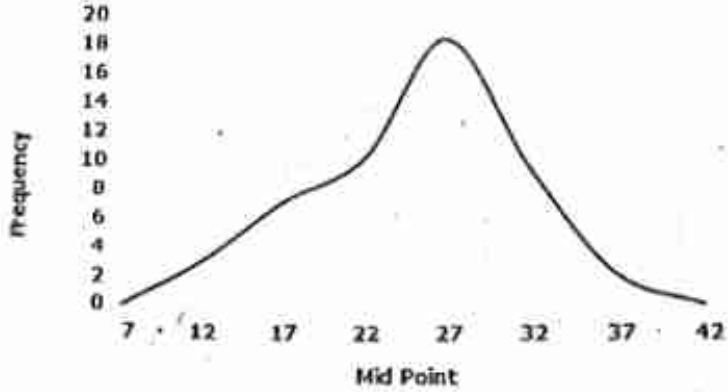
Histogram



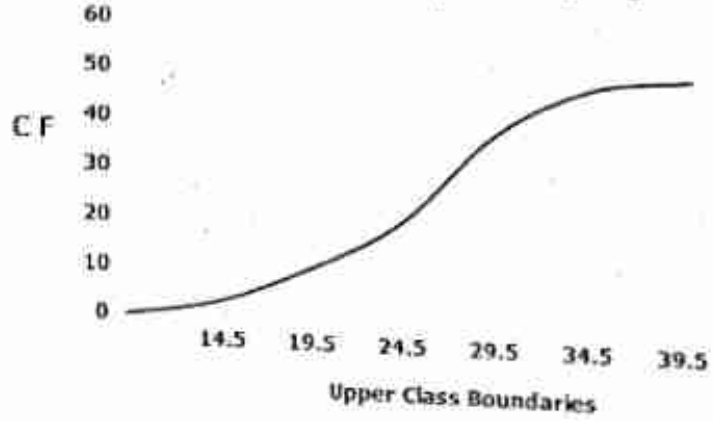
Frequency Polygon



Frequency Curve



Cumulative Frequency Curve / Ogive



2.4 Stem and leaf plot

A data plot which uses leading part of the data value as the stem and the rest of the data value (the leaf) to form groups or classes. This is very useful for sorting data quickly.

Example 2.12: make stem and leaf display for the data given below.
12.1, 11.7, 9.6, 13.1, 8.5, 8.7, 11.6, 12.3, 13.2, 9.0, 10.2, 11.4, 8.7, 9.1, 9.4, 11.5, 13.2, 10.7, 11.2, 10.9, 12.0, 12.6, 11.7

Solution: Stem and leaf display

Stem	Leaf
8	5, 7, 7
9	0, 1, 4, 6
10	2, 7, 9
11	2, 4, 5, 6, 7, 7
12	0, 1, 3, 6
13	1, 2, 2

Key: $8|5 = 8.5$

Example 2.13: Marks obtained by BSc III year of Govt. College Renala Khurd in BA/BSc A/2018 exam are given below. Make stem and leaf display. Convert them into a frequency distribution. 252, 204, 224, 243, 266, 280, 300, 215, 238, 272, 297, 294, 285, 271, 267, 257, 206, 211, 226, 235, 240, 202, 204, 221, 226, 243, 241, 266, 278, 216, 212, 235, 230, 256, 254, 207, 200, 202, 203, 213, 216, 223, 229, 230, 233, 234, 243, 245, 259, 251, 255, 203.

Solution:

Stem	Leaf	Stem	Leaf
20	0 2 2 3 3 4 4 6 7	28	0 5
21	1 2 3 5 6 6	29	4 7
22	1 3 4 6 6 9	30	0
23	0 0 3 4 5 5 8		
24	0 1 3 3 3 5		
25	1 2 4 5 6 7 9		
26	6 6 7		
27	1 2 8		

Key: 27 | 1 = 271

Frequency distribution

Classes	Tally	Frequency	
200 – 209	𝍸 IIII	9	
210 – 219	𝍸 I	6	
220 – 229	𝍸 I	6	
230 – 239	𝍸 II	7	
240 – 249	𝍸 I	6	
250 – 259	𝍸 II	7	
260 – 269	III	3	
280 – 289	II	2	
290 – 299	II	2	
300 – 309	I	1	$\sum f = 52$

Note:

(I) **Sigma (Σ):** It is a Greek letter used to denote the summation of a series of number for example the sum of $X_1, X_2, ..., X_n$ is written as

$$X_1 + X_2 + ... + X_n = \sum_{i=1}^{n} X_i$$

(II) **Pie (Π):** It is a Greek letter used to denote the product of a series of number for example the product of $X_1, X_2, ..., X_n$ is written as $X_1 \times X_2 \times ... \times X_n = \prod_{i=1}^{n} (X_i)$

2.5 Properties of summation (Σ)

01) $\sum_i (c) = nc$ where c is any constant

02) $\sum_i (cX_i) = c\sum_i X_i$

03) $\sum_i (X_i + Y_i) = \sum_i X_i + \sum_i Y_i$

04) $\sum_i (X_i - Y_i) = \sum_i X_i - \sum_i Y_i$

05) $\sum_i X_i^2 \neq \left(\sum_i X_i\right)^2$

06) $\left(\sum_i X_i\right)^2 = \sum_i X_i^2 + \sum_{i \neq j}\sum_j X_i X_j = \sum_i X_i^2 + 2\sum_{i<j}\sum_j X_i X_j$

07) $\sum(X_i Y_i) \neq \sum X_i \sum Y_i$

Example 2.14: Determine the value of the following expression in which $X_1 = 2, X_2 = 4, X_3 = 8, X_4 = 1, X_5 = 7, X_6 = 5$

(a). (i) $\sum_{i=1}^{6} X_i$ (ii) $\sum_{i=2}^{6} X_i$ (iii) $\sum_{i=1}^{5} X_i$ (iv) $\prod_{i=1}^{3} X_i$

(b). verify that $\sum_{i=1}^{6} X_i^2 \neq \left(\sum_{i=1}^{6} X_i\right)^2$

Solution:

(i) $\sum_{i=1}^{4} X_i = 2+4+8+1 = 15$

(ii) $\sum_{i=2}^{6} X_i = 4+8+1+7+5 = 25$

(iii) $\sum_{i=1}^{5} X_i = 8+1+7+5 = 21$

(iv) $\prod_{i=1}^{3} (X_i) = 2 \times 4 \times 8 = 64$

(b) $\left(\sum_{i=1}^{6} X_i\right)^2 = (2+4+8+1+7+5)^2 = (27)^2 = 729$

$\sum_{i=1}^{6} X_i^2 = 2^2 + 4^2 + 8^2 + 1^2 + 7^2 + 5^2 = 159$

Hence $\left(\sum_{i=1}^{6} X_i\right)^2 \neq \sum_{i=1}^{6} X_i^2$

Example 2.15: Express the following in summation notation

(i) $X_1 + X_2 + X_3 + X_4$

(ii) $a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$

(iii) $X_4^2 + X_5^2 + X_6^2 + X_7^2$

Solution:

(i) $X_1 + X_2 + X_3 + X_4 = \sum_{i=1}^{4} X_i$

(ii) $a_1 X_1 + a_2 X_2 + \ldots + a_n X_n = \sum_{i=1}^{n} a_i X_i$

(iii) $X_4^2 + X_5^2 + X_6^2 + X_7^2 = \sum_{i=4}^{7} X_i^2$

Example 2.16: Expand $\sum_{i=1}^{4} A$ and $\sum_{i=1}^{n} A$

Solution: $\sum_{i=1}^{4} A = A + A + A + A = 4A$ and $\sum_{i=1}^{n} A = A + A + \ldots + A = nA$

Multiple Choice Questions

1. Arranged in ascending or descending order is:
a) Grouped data b) Classification c) Array d) ungrouped data

2. Data not arranged in ascending or descending order is:
a) Raw data b) grouped data c) arrangement d) groups

3. Systematic arrangement of data into rows and columns:
a) Bar chart b) classification c) tabulation d) histogram

4. In a table, foot note and source notes are:
a) Same b) different c) identical d) main heading

5. Graph of cumulative frequency distribution:
a) Histogram b) bar chart c) Ogive d) polygon

6. Midpoint of rectangles in histogram are connected:
a) Ogive b) Historigram c) Bar Chart d) Frequency Polygon

7. The relation showing between whole and its components, chart is:
a) Frequency Polygon b) Multiple bar chart c) Pie chart d) Ogive

8. Value that divides a class into two equal parts:
a) Class interval b) mid-point c) Class limit d) frequency

9. Dividing range by numbers of classes:
a) Frequency b) midpoint c) class interval d) class limits

10. In histogram, along X-axis:
a) Frequency b) cf c) Class boundaries d) Class limits

11. Frequency table is an arrangement of data by classes together with their corresponding class ------.
a) Frequencies b) midpoints c) boundaries d) limits

12. Heading at the top of the table:
a) Footnote b) Head note c) Sub-note d) Title

13. Headings for various columns:
a) Source note b) column captions c) Stub d) body

14. Stub:
a) Row captions b) Column captions c) Footnote d) Prefatory note

15. Box head:
a) Row captions b) Column captions c) Footnote d) Prefatory note

16. Cumulative frequencies are:
a) Decreasing b) increasing c) non increasing d) fixed

17. Division of a circle into different sectors:
a) Pictogram b) histogram c) Ogive d) pie chart

18. Histograms, bar charts and frequency polygons are:
(a) One dimension diagrams (b) two dimension diagrams
(c) Cumulative diagrams (d) dispersion diagrams

19. Angles in a Pie diagram:
a) $\frac{Total\ part}{component\ part} \times 360^{\circ}$ b) $\frac{component\ part}{Total\ part} \times 360$
c) $\frac{component\ part}{Total\ part} \times 360^{\circ}$ d) $\frac{Total\ part}{component\ part} \times 360$

20. Circle with sectors representing various quantities:
a) Histogram b) Frequency Polygon c) Pie Chart d) Component Bar chart

21. A Histogram contains:
a) Adjacent rectangles b) Non Adjacent Rectangles
c) Adjacent squares d) Adjacent triangles

22. A frequency polygon is a figure of:
a) Two sides b) Three Sides c) Many sides d) Circles

23. Graph of a time series:
a) Histogram b) Ogive c) Historigram d) Polygon

24. Component bar charts are used when data is divided into:
a) Ratios b) groups c) circles d) intervals

25. Frequency curve is:
a) Asymptotic to y-axis b) Non-Asymptotic to y-axis
c) Asymptotic to x-axis d) Non-Asymptotic to X-axis

26. A frequency curve touches x-axis
a) Yes b) No c) Sometime d) always

27. Decumulative frequency is presented by:
a) More than Ogive b) Less than Ogive c) Equal to Ogive d) Curve

28. In a histogram the area of each rectangle is proportional to:
a) The class mark of the corresponding class
b) The class size of the corresponding class c) Frequency of the corresponding class
d) Cumulative frequency of the corresponding class

29. In frequency polygon along X-axis:
a) upper limit of the class b) lower limit of the class
c) mid value of the class d) frequency

30. A sector diagram:
a) Pie chart b) Bar diagram c) Scatter diagram d) Histogram

key

Sr.	Ans	Sr.	Ans	Sr.	Ans
1	c	2	a	3	c
4	b	5	c	6	d
7	c	8	b	9	c
10	c	11	a	12	d
13	b	14	a	15	b
16	b	17	d	18	b
19	c	20	c	21	a
22	c	23	c	24	b
25	c	26	b	27	a
28	c	29	c	30	a

Exercise

Q. NO. 2.1: Define the following terms

a) Classification and tabulation
b) Frequency and frequency distribution
c) Class limits and class boundaries
d) Ungrouped and grouped data
e) Cumulative and relative frequencies

Q NO. 2.2: Draw simple bar chart and pie diagram for the following data

Continent	Number of countries
Africa	54
Asia	47
Europe	43
North America	23
Australia/Oceania	14
South America	12

Q NO. 2.3: Draw multiple and component bar chart for the following data

Country	Crude Birth Rate	Crude Death Rate
Pakistan	26.88	6.80
India	17.23	7.34
Bangladesh	17.22	5.52
Sri Lanka	14.87	6.96
Afghanistan	30.53	6.96

Q NO. 2.4: Make histogram, frequency polygon, frequency curve and Ogive.

Classes	f	Classes	f
11 – 20	1	51 – 60	30
21 – 30	11	61 – 70	20
31 – 40	26	71 – 80	16
41 – 50	40	81 – 90	3

Q NO. 2.5: Make histogram, frequency polygon, frequency curve and Ogive, after making frequency distribution from the following data.

11, 20, 21, 30, 31, 36, 45, 50, 27 ,22 , 18, 13, 33, 35, 37, 39, 41, 43, 47, 45, 42, 44, 44, 12, 14, 13, 18, 20, 21, 23, 24, 25, 27, 26, 31, 32, 33, 37, 40, 39, 44, 23, 24, 27, 28, 26, 33, 32, 39, 22, 24, 27

Q NO. 2.6: Make histogram, frequency polygon, frequency curve and Ogive.

Classes	f	Classes	f
15 – 29.9	8	75 – 89.9	37
30 – 44.9	15	90 – 104.9	13
45 – 59.9	32	105 – 119.9	5
60 – 74.9	45		

Q NO. 2.7: Make multiple and component bar diagram, of the data about schools in district Okara.

School Type	Number of schools	
	Male	Female
SS	109	77
Elementary	106	78
Primary	542	487

Source: School education department

Q NO. 2.8: Following are marks obtained by BSc Final students of Govt. College Renala Khurd in BA/BSc A/2018 exam.

472, 480, 492, 403, 414, 426, 499, 404, 419, 422, 438, 440, 452, 469, 477, 409, 414, 425, 431, 449, 452, 466, 463, 454, 442, 428, 415, 403, 491, 496, 489, 493, 411, 421, 448, 459, 462, 457, 463, 447, 429, 448, 459, 469, 497, 419, 414, 423, 431, 446, 451

Make stem and leaf display. Form a frequency distribution and present data by histogram, frequency curve, frequency polygon and Ogive.

Q NO. 2.9: Make multiple and component bar diagram, of the data about number of degree awarding institutes in provinces of Pakistan.

Province	Number of degree awarding institutes	
	Public	Private
Punjab	34	26
Sindh	23	31
KPK	25	10
Balochistan	7	1
AJ & K	5	2
GB	1	0
ICT	15	5

Source: School education department

Q NO. 2.10: Make multiple and component bar diagram, of the data about survival rate (retention rate) in provinces of Pakistan.

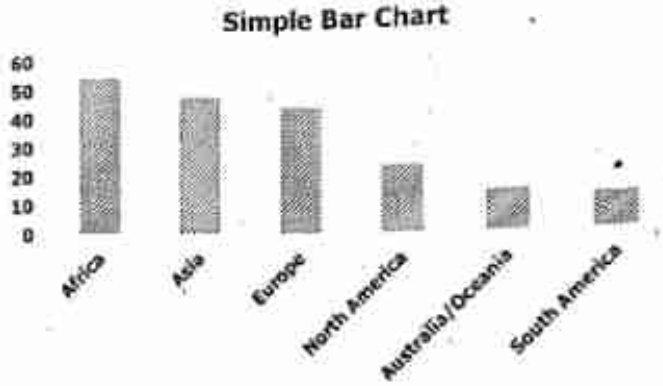
Province	Survival Rate	
	Male	Female
Punjab	71	74
Sindh	60	58
KPK	71	57
Balochistan	39	44
AJ & K	84	85
GB	100	100
FATA	37	26

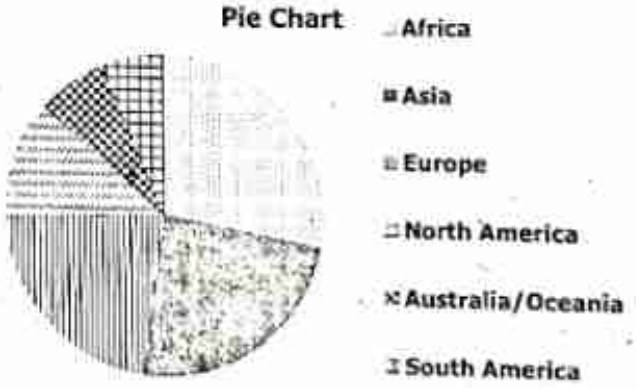
Source: School education department

Q NO. 2.11: Time in minutes taken by a group of 40 ten years old boys to do a series of abstract puzzles: 24, 83, 36, 22, 81, 39, 60, 62, 38, 66, 38, 36, 45, 20, 20, 67, 41, 87, 41, 82, 35, 82, 28, 80, 80, 68, 40, 27, 43, 80, 31, 89, 83, 24, 75, 78, 71, 30, 49, 54 Make frequency distribution taking interval size of 10 and histogram.

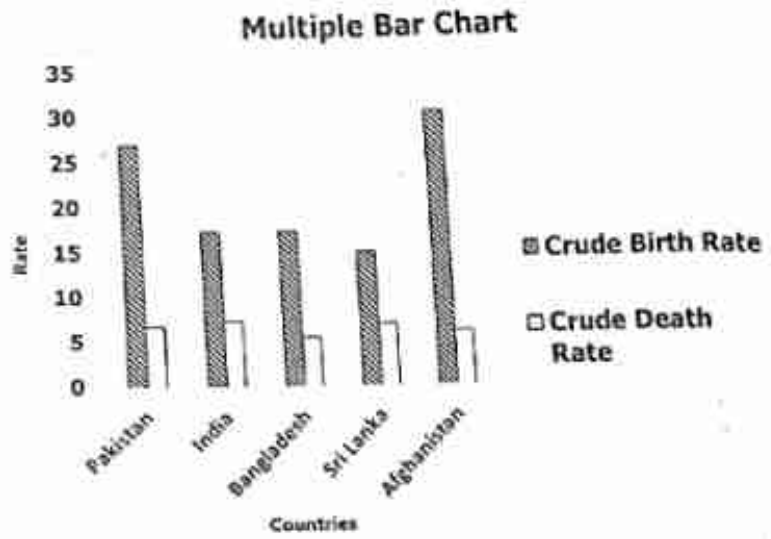
Solution

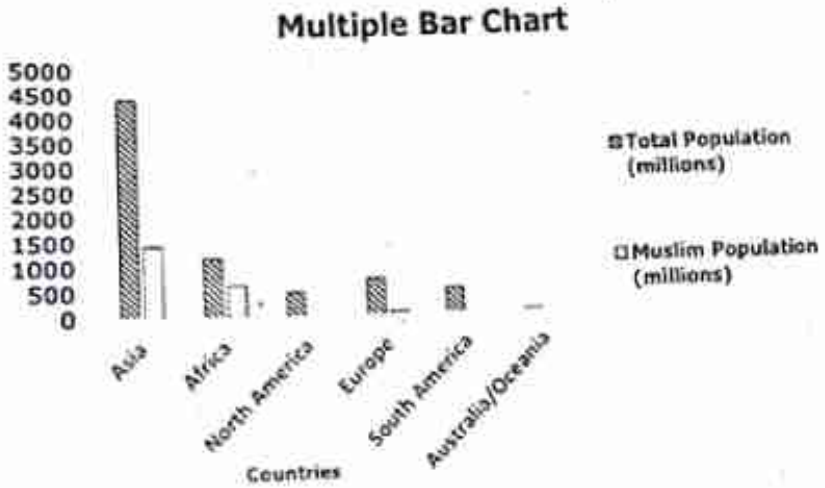
Q NO. 2.2:



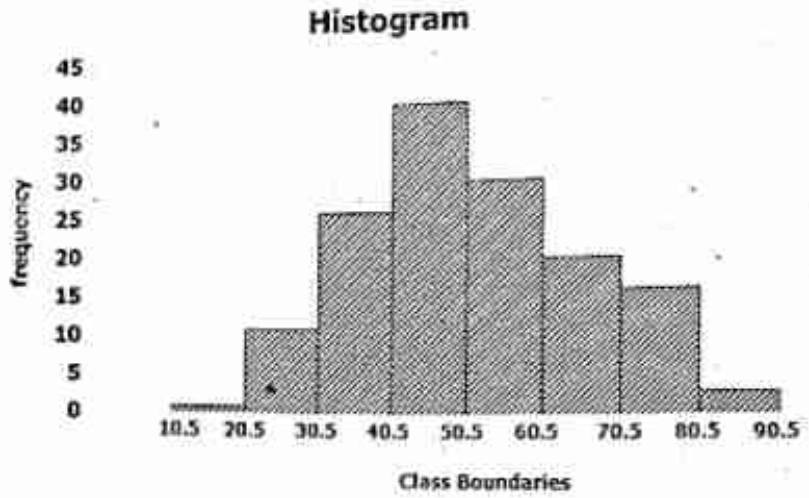


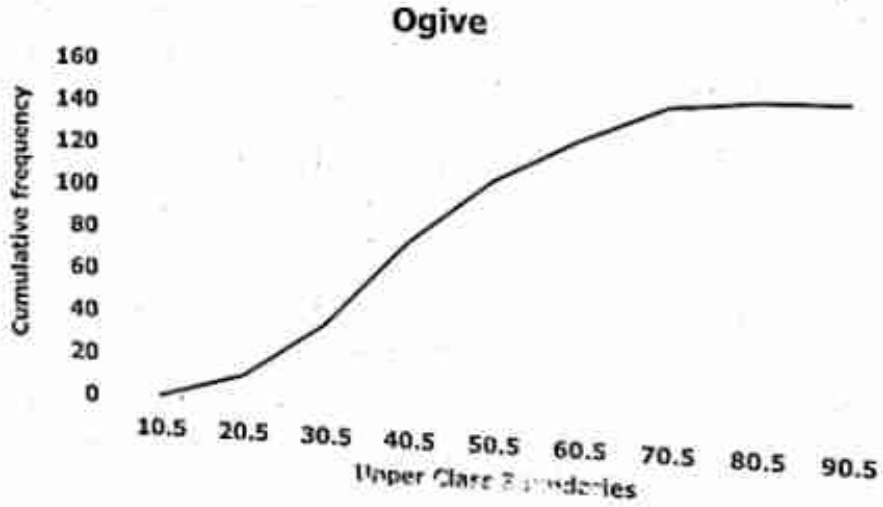
Q N0. 2.3:


Multiple Bar Chart
35
30
25
20
15
10
5
0
Rate
Pakistan
India
Bangladesh
Sri Lanka
Afghanistan
Countries
Crude Birth Rate
Crude Death Rate


Multiple Bar Chart
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Asia
Africa
North America
Europe
South America
Australia/Oceania
Countries
Total Population (millions)
Muslim Population (millions)

Q N0. 2.4:


Histogram
45
40
35
30
25
20
15
10
5
0
frequency
10.5
20.5
30.5
40.5
50.5
60.5
70.5
80.5
90.5
Class Boundaries

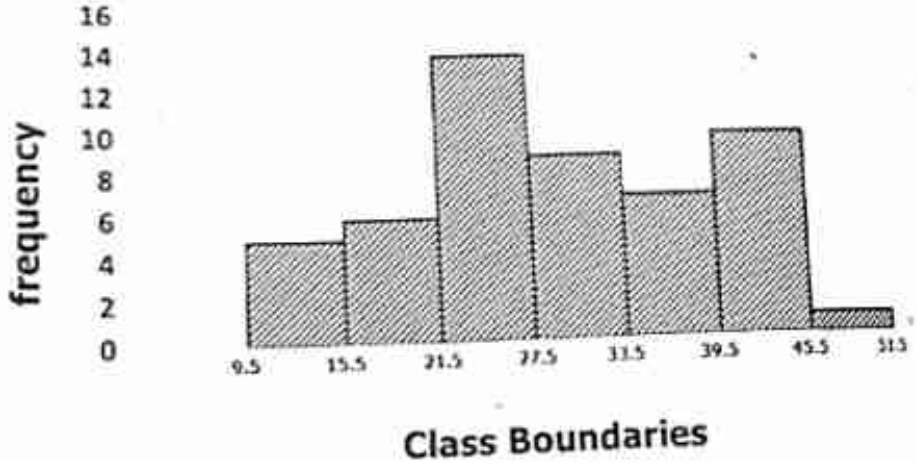

Ogive
160
140
120
100
80
60
40
20
0
Cumulative frequency
10.5
20.5
30.5
40.5
50.5
60.5
70.5
80.5
90.5

Q NO. 2.5:

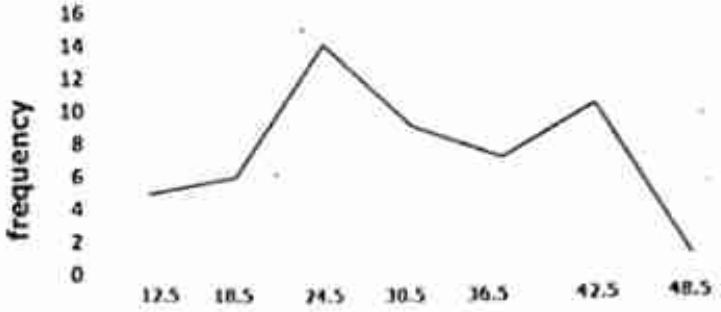
Frequency distribution

Classes	Tally	Frequency
10 – 15	𝍸	5
16 – 21	𝍸 I	6
22 – 27	𝍸 𝍸 IIII	14
28 – 33	𝍸 IIII	9
34 – 39	𝍸 II	7
40 – 45	𝍸 𝍸	9
46 – 51	I	2
		$\Sigma f = 52$

Histogram

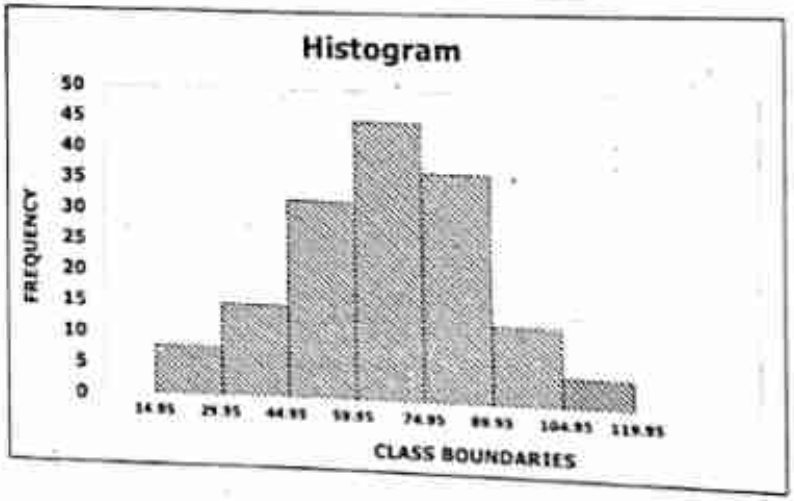


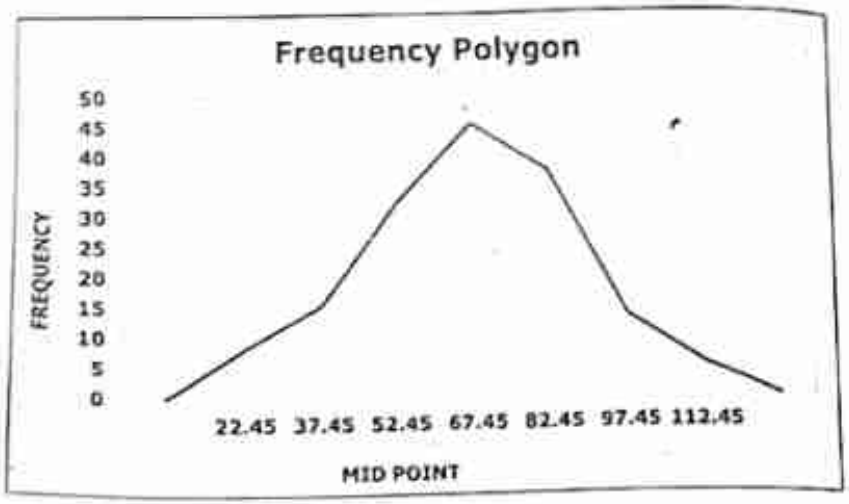
Frequency Polygon

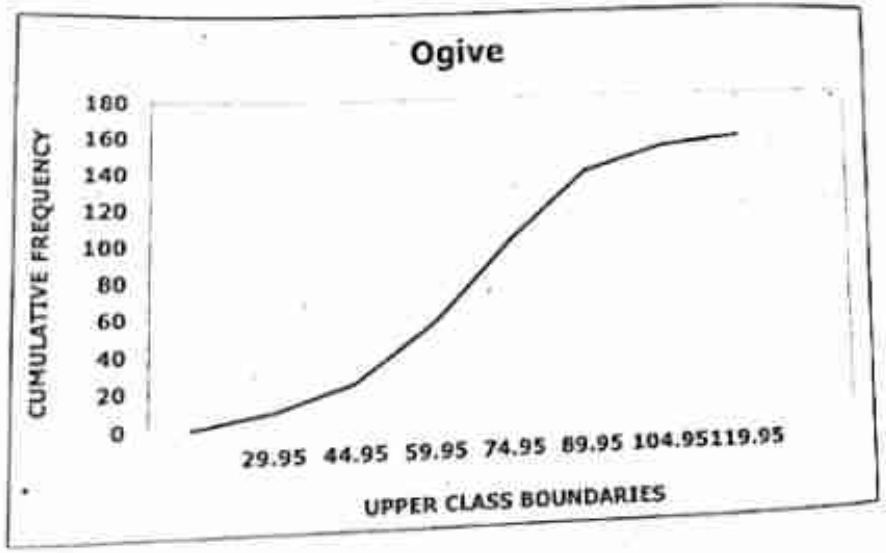


Q NO. 2.6: Make histogram, frequency polygon, frequency curve and Ogive.

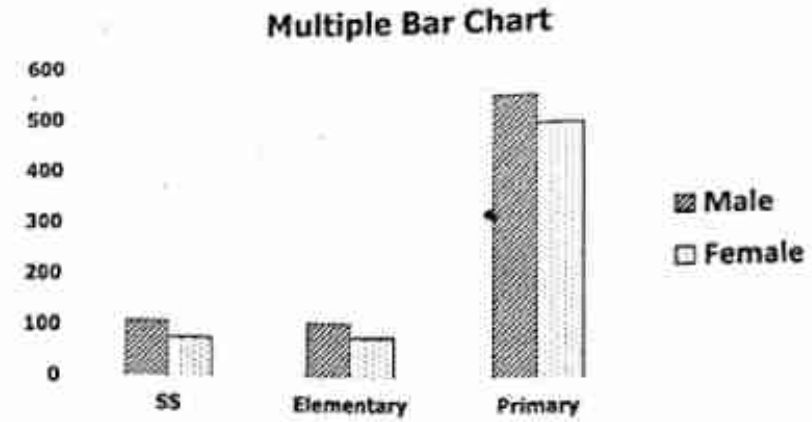
Classes	f	Classes	f
15 – 29.9	8	75 – 89.9	37
30 – 44.9	15	90 – 104.9	13
45 – 59.9	32	105 – 119.9	5
60 – 74.9	45		

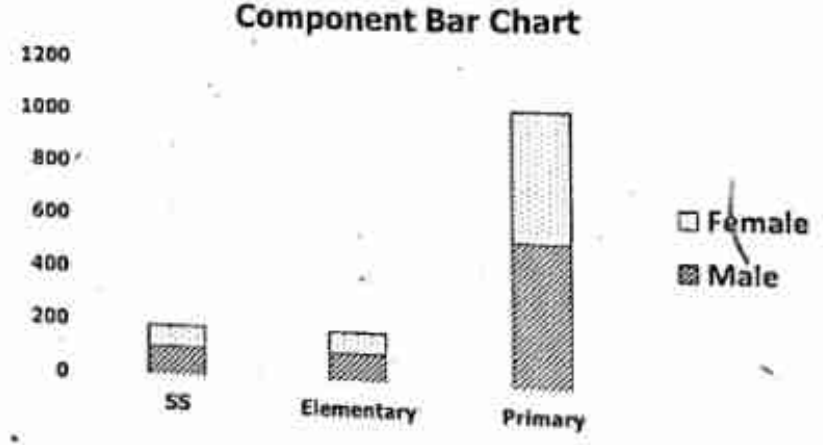



Frequency Polygon
FREQUENCY
50
45
40
35
30
25
20
15
10
5
0
22.45 37.45 52.45 67.45 82.45 97.45 112.45
MID POINT


Ogive
CUMULATIVE FREQUENCY
180
160
140
120
100
80
60
40
20
0
29.95 44.95 59.95 74.95 89.95 104.95 119.95
UPPER CLASS BOUNDARIES

Q NO. 2.7:


Multiple Bar Chart
600
500
400
300
200
100
0
SS
Elementary
Primary
Male
Female


Component Bar Chart
1200
1000
800
600
400
200
0
SS
Elementary
Primary
Female
Male

Q NO. 2.8

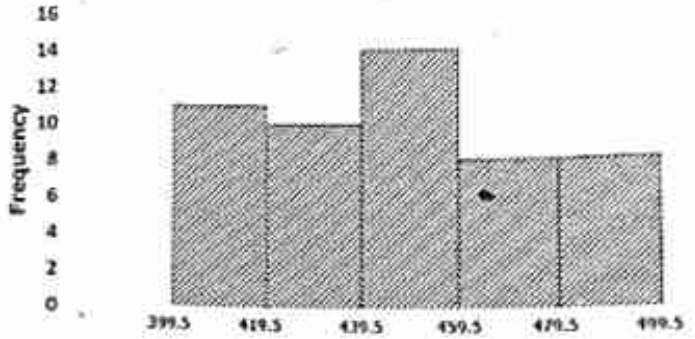
Stem	Leaf
40	3 3 4 9
41	1 4 4 4 5 9 9
42	1 2 3 5 6 8 9
43	1 1 8
44	0 2 6 7 8 8 9
45	1 2 2 4 7 9 9
46	2 3 3 6 9 9
47	2 7
48	0 9
49	1 2 3 6 7 9

$n = 51$ $X_0 = 403$, $X_{99} = 499$, $R = 499 - 403 = 96$, No. of classes = 5

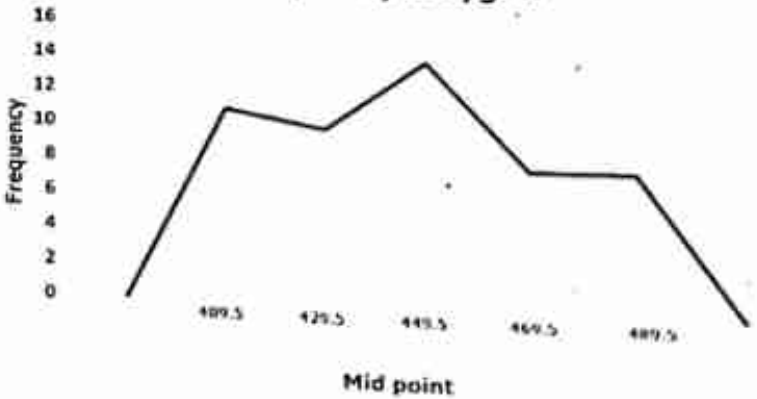
$$h = \frac{R}{C} = \frac{96}{5} = 19.2 = 20$$

C.I	f	C.B	Midpoint (X)	cf
400 – 419	11	399.5 – 419.5	409.5	11
420 – 439	10	419.5 – 439.5	429.5	21
440 – 459	14	439.5 – 459.5	449.5	35
460 – 479	8	459.5 – 479.5	469.5	43
480 – 499	8	479.5 – 499.5	489.5	51

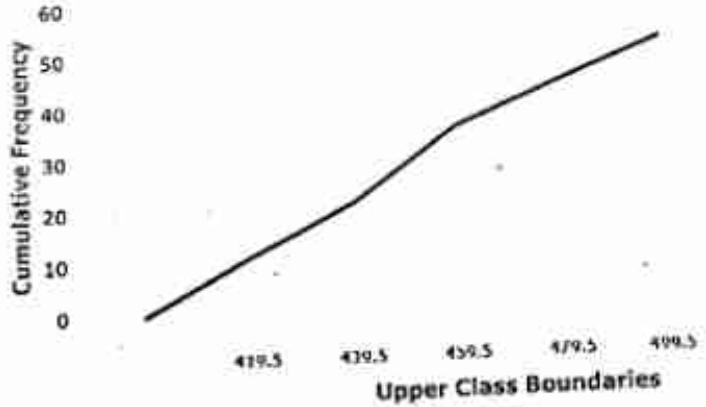
Histogram



Frequency Polygon

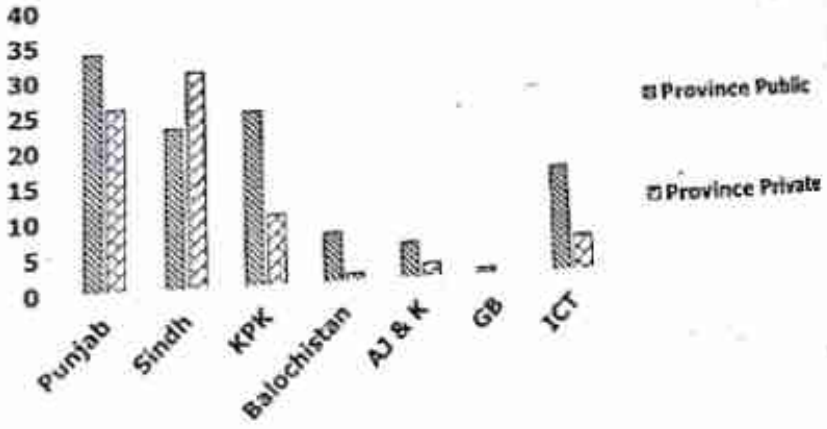


Ogive

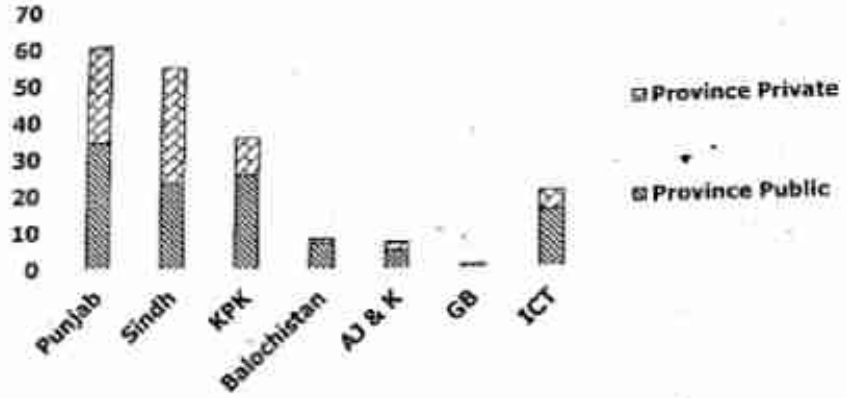


Q NO. 2.9:

Multiple Bar Chart

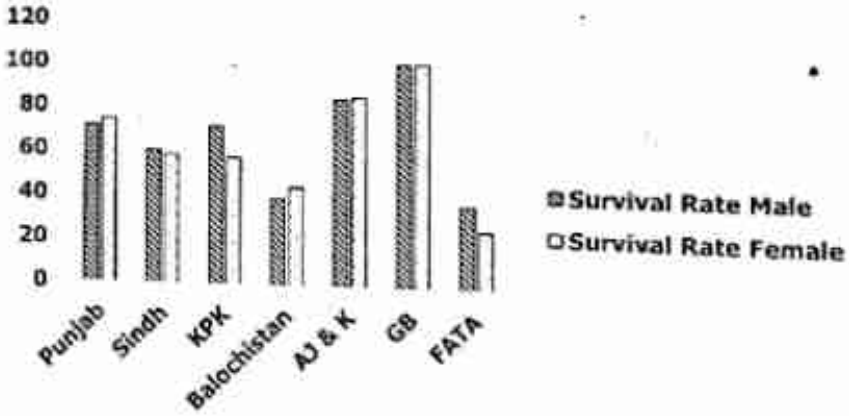


Component Bar Chart

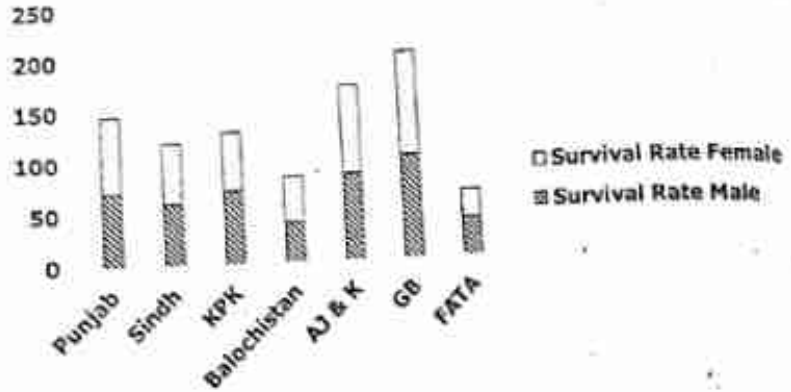


Q NO. 2.10:

Multiple Bar Chart



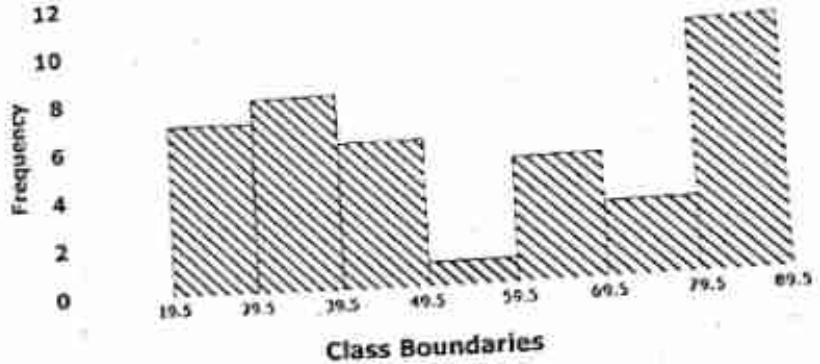
Component Bar Chart



Q N0. 2.11:

Classes	Tally	f	C.B
20 – 29	𝍸 II	7	19.5 – 29.5
30 – 39	𝍸 III	8	29.5 – 39.5
40 – 49	𝍸 I	6	39.5 – 49.5
50 – 59	I	1	49.5 – 59.5
60 – 69	𝍸	5	59.5 – 69.5
70 – 79	III	3	69.5 – 79.5
80 – 89	𝍸 𝍸	10	79.5 – 89.5

Histogram



Chapter 3

MEASURES OF CENTRAL TENDENCY

Basic Terms used in statistics are well known to a common person

Learning Goals

(i) To develop the skills to analyze the data by finding its representative value.
(ii) To understand and interpret the distribution of the data.
(iii) To analyze the data through measure of scatterings, skewness and kurtosis.
(iv) To compare individuals in a data set by Z-score.

3.1 Average

A single value that represents the whole set of the data is called an average.

3.1.1 Measures of central tendency

The measures which tend to lie in the Centre of the distribution are called measures of central tendency(also known as Measures of Location). Averages are usually called measures of central tendency because they lie in the center of a distribution.

3.1.2 Characteristics of a good average

A good average should be

01) Easy to understand.
02) Rigidly defined.
03) Based on all the observation.
04) Least affected by the extreme value.
05) Capable to further mathematical treatment.
06) Easy to calculate and simple to follow.
07) Least affected by fluctuation of sampling.
08) Found by arithmetic (calculation) as well as by graphic method.
09) Calculated in closed and open end frequency distribution.

3.1.3 Functions of an average

01) It presents simple and concise picture of large and complicated set of data.
02) It makes possible and easier to compare two or more groups of data.
03) It facilitates the interpretation of data.
04) It helps in taking decision.

3.1.4 Types of average

There are five common and well known averages

01) Arithmetic mean ($\bar{X}$)
02) Geometric mean (G)
03) Harmonic mean (H)
04) Median ($\tilde{X}$) and
05) Mode ($\hat{X}$)

3.2 Arithmetic mean

Arithmetic mean (AM) of a variable denoted by $\bar{X}$ (read as X bar) is defined as the ratio between sum of the values and number of the values. It is given as $\bar{X} = \frac{\sum X_i}{n}$.

Other Formulae

Un- grouped	Grouped	Termed as	
$\bar{X} = \frac{\sum X_i}{n}$	$\bar{X} = \frac{\sum f_i X_i}{\sum f_i}$	**Direct/ Long Method**	
$\bar{X} = A + \frac{\sum D_i}{n}$	$\bar{X} = A + \frac{\sum f_i D_i}{\sum f_i}$	**Indirect / short cut Method** $D_i = X_i - A$	A is assumed mean / Provisional mean And h is class interval or common divisor
$\bar{X} = A + \frac{\sum u_i}{n} \times h$	$\bar{X} = A + \frac{\sum f_i u_i}{\sum f_i} \times h$	**Step deviation / Coding method** $u_i = \frac{X_i - A}{h}$	

Example 3.1: calculate arithmetic mean from the following data by using direct and short cut method. 23, 12, 34, 27, 11, 19, 22, 25

Solution: X_i = 23, 12, 34, 27, 11, 19, 22, 25

ΣX_i = 23+12+34+27+11+19+22+25 = 173

$$\bar{X} = \frac{\sum X_i}{n} = \frac{173}{8} = 21.625$$

Short cut method

X_i = 23, 12, 34, 27, 11, 19, 22, 25

$D_i = X_i - A = X_i - 19$ Where A = 19

D_i = 4, -7, 15, 8, -6, 0, 3, 6; $\sum D_i = 21$

$$\bar{x} = A + \frac{\sum_i D_i}{n} = 19 + \frac{21}{8} = 19 + 2.625 = 21.625$$

Example 3.2: Compute arithmetic mean from the following data of marks in statistics of BS (H) mathematics by (a) Direct method (b) In-direct (Short cut) method and (c) Step deviation (coding) method

Marks	0 – 19	20 – 39	40 – 59	60 – 79	80 – 99
No. of students	6	10	18	12	4

Solution:

(a) Direct method

Marks	f_i	X_i	$f_i X_i$
0 – 19	6	9.5	57
20 – 39	10	29.5	295
40 – 59	18	49.5	891
60 – 79	12	69.5	834
80 – 99	4	89.5	358
Σ	**50**		**2435**

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i}$$

$$\bar{X} = \frac{2435}{50} = 48.7 \text{ marks}$$

(b) and (c)

Marks	f_i	X_i	D_i	$f_i D_i$	u_i	$f_i u_i$
0 – 19	6	9.5	-40	-240	-2	-12
20 – 39	10	29.5	-20	-200	-1	-10
40 – 59	18	49.5	0	0	0	0
60 – 79	12	69.5	20	240	1	12
80 – 99	4	89.5	40	160	2	8
Σ	**50**			**-40**		**-2**

Where

$D_i = X_i - 49.5$ and $u = \frac{X - 49.5}{20}$; A = 49.5 and h = 20

(b) Short cut/In-direct method

$$\bar{x} = A + \frac{\sum_i f_i D_i}{\sum_i f_i} = 49.5 + \frac{-40}{50} = 49.5 - 0.8 = 48.7 \text{ marks}$$

(c) Coding/Step deviation method

$$\bar{x} = A + \frac{\sum_i f_i u_i}{\sum_i f_i} \times h = 49.5 + \frac{-2}{50} \times 20 = 49.5 - 0.8 = 48.7 \text{ marks}$$

3.3.1 Properties of arithmetic mean

01) AM of a constant is constant itself i.e. If $X = C$ then $\bar{X} = C$

02) Sum of the deviations of observations from their AM is zero. i.e. $\sum(X - \bar{X}) = 0$

03) Sum of the squared deviations of the observations from their AM is least. i.e $\sum(X - \bar{X})^2 < \sum(X - A)^2$

04) AM is affected by both the change of origin and scale. If $Y = aX + b$ then $\bar{Y} = a\bar{X} + b$.

Example 3.3 (a) If $Y = X + 10$, find $\bar{Y}$ when $\bar{X} = 6$

(b) If $Y = 3X$, find $\bar{Y}$ when $\bar{X} = 4$ (c) If $Y = 2X + 5$, find $\bar{Y}$ when $\bar{X} = 7$

Solution:

(a) If $Y = X + 10$, then $\bar{Y} = \bar{X} + 10$, when $\bar{X} = 6$

$\bar{Y} = \bar{X} + 10 = 6 + 10 = 16$

(b) If $Y = 3X$, then $\bar{Y} = 3\bar{X}$ when $\bar{X} = 4$

$\bar{Y} = 3\bar{X} = 3(4) = 12$

(c) If $Y = 2X + 5$, then $\bar{Y} = 2\bar{X} + 5$ when $\bar{X} = 7$

$\bar{Y} = 2\bar{X} + 5 = 2(7) + 5 = 19$

05) Combined mean is given as $\bar{\bar{X}} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n_1 + n_2 + \ldots + n_k}$

3.3.2 Merits of arithmetic mean

01) It is well defined.
02) It is easy to calculate, understand and interpret.
03) It is based on all the observations, therefore it is best representative of the data.
04) For two observations it is closely related to standard deviation (most common measure of dispersion).
05) It is independent of order of the observations.
06) It is relatively stable measure.
07) It is determined for almost every kind of data.

3.3.3 Demerits of arithmetic mean

01) It is highly affected by the extreme values/outliers.
02) It is not appropriate for skewed distribution.
03) It cannot be calculated for qualitative (nominal and ordinal) data.
04) It cannot be computed for open-end classes without assuming open ends.

3.3.4 Weighted arithmetic mean

If we assign weights to the values depending on their relative importance then the arithmetic mean of such values is called weighted arithmetic mean or simply weighted mean. It is given as

$$\bar{X}_w = \frac{\sum_{i=1}^{n} X_i w_i}{\sum_{i=1}^{n} w_i}$$

Note: We use weighted arithmetic mean instead of arithmetic mean when the observations are not of equal importance.

Example 3.4: Calculate weighted arithmetic mean from the following marks obtained by a student in ICS examination

Subject	English	Urdu	Math	Stat	Computer
Marks	145	120	165	140	115
Weights	5	3	4	4	2

Solution:

Subject	English	Urdu	Math	Stat	Computer	Σ
X	145	120	165	140	115	
W	5	3	4	4	2	18
XW	725	360	660	560	230	2535

$$\bar{X}_w = \frac{\sum_{i=1}^{5} X_i W_i}{\sum_{i=1}^{5} W_i} = \frac{2535}{18} = 140.83 \ marks$$

3.3.5 Combined mean

If n_1 values have mean $\bar{X}_1$, n_2 values have mean $\bar{X}_2$ and so on n_k values have mean $\bar{X}_k$, then the mean of all values is called combined mean or overall mean. It is denoted by $\bar{\bar{X}}$ or $\bar{X}_c$ and is given as:

$$\bar{\bar{X}} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} n_i\bar{X}_i}{\sum_{i=1}^{k} n_i}$$

Example 3.5: Find combined mean for the following date.

$n_1 = 5,\ \bar{X}_1 = 8,\ n_2 = 6, \bar{X}_2 = 12,\ n_3 = 10,\ \bar{X}_3 = 14$

Solution $n_1 = 5,\ \bar{X}_1 = 8,\ n_2 = 6, \bar{X}_2 = 12,\ n_3 = 10,\ \bar{X}_3 = 14$

$$\bar{\bar{X}} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3}{n_1 + n_2 + n_3} = \frac{5\times 8 + 6\times 12 + 10\times 14}{5 + 6 + 10} = \frac{252}{21} = 12$$

3.4 Geometric mean

Geometric mean (G) of a set of n positive values $(x_1, x_2, \ldots, x_n)$ is defined as the positive nth root of their product, i.e.:

$$G = \sqrt[n]{(x_1 x_2 \ldots x_n)} = \left(\prod_{i=1}^{n}(x_i)\right)^{\frac{1}{n}} \quad \text{Where } x_i > 0$$

Other formula

$G = Anti\log\frac{1}{n}\sum_{i=1}^{n}\log x_i$ For ungrouped data

$G = Anti\log\frac{\sum_{i=1}^{n} f\log x_i}{\sum f}$ For grouped data

Example 3.6: Find geometric mean of the following data 10, 13, 17, 21, 15

Solution: First method (By Definition)

$X_i = 10, 13, 17, 21, 15 \qquad \prod_{i=1}^{5}(X_i) = 10\times 13\times 17\times 21\times 15 = 696150$

$$G = \left(\prod_{i=1}^{n}(X_i)\right)^{\frac{1}{n}} = (696150)^{\frac{1}{5}} = 14.74$$

Second Method:

X_i	10	13	17	21	15	Σ
$\log X_i$	1.0000	1.1139	1.2304	1.3222	1.1761	5.8426

$$G = Anti\log\frac{1}{n}\sum\log x_i \quad G = Anti\log\left(\frac{5.8426}{5}\right) = Anti\log(1.16852) = 14.74$$

Example 3.7: Compute GM from the following data

Income(000) (Rs)	10-20	20-30	30-40	40-50	50-60	60-70
No. of families	8	15	30	40	5	2

Solution:

Income (000) Rs	f_i	X_m	$f_i \log X_i$
10 – 20	8	15	9.4087
20 – 30	15	25	20.9691
30 – 40	30	35	46.3220
40 – 50	40	45	66.1285
50 – 60	5	55	8.7018
60 – 70	2	65	3.6258
Σ	**100**		**155.1559**

$$G = Anti\log\frac{\sum_{i=1}^{n} f_i \log X_i}{\sum_{i=1}^{n} f_i} = Anti\log\left(\frac{155.1559}{100}\right) = Anti\log(1.551559)$$

$$= 35.61(000)\ Rs$$

3.4.1 Merits of geometric mean

01) It is well defined by a mathematical formula.
02) It is based on all the observations.
03) It is least affected by extreme values.
04) It is useful to average percentage increase, ratios, Indices and growth rates.
05) It is capable of further algebraic manipulation.
06) It is independent of order of the observations.
07) It is an appropriate average for skewed distribution.

3.4.2 Demerits of geometric mean

01) It is not easy to understand and calculate.
02) It is not being calculated if any value is negative.
03) It is zero if any observation is zero.

3.5 Harmonic mean (H)

Harmonic mean is defined as the reciprocal of arithmetic mean of the reciprocals of values of a variable. Let $x_1, x_2, \ldots, x_n$ be "n" values, then harmonic mean is given as

$$H = \frac{n}{\sum_{i=1}^{n}(1/x_i)}$$ For ungrouped data $x_i \neq 0$

$$H = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n}(f_i/x_i)}$$ For grouped data $x_i \neq 0$

Example 3.8: Find Harmonic mean of the following speed (km/hr) data
46, 35, 55, 49, 60, 35

Solution: $X_i = 46, 35, 55, 49, 60, 35$ $\quad \frac{1}{X_i} = \frac{1}{46}, \frac{1}{35}, \frac{1}{55}, \frac{1}{49}, \frac{1}{60}, \frac{1}{35}$

$$\sum_{i=1}^{6}\left(\frac{1}{X_i}\right) = \frac{1}{46} + \frac{1}{35} + \frac{1}{55} + \frac{1}{49} + \frac{1}{60} + \frac{1}{35} = 0.1341$$

$$HM = \frac{n}{\sum(1/X_i)} = \frac{6}{0.1341} = 44.74\ km/hr$$

Example 3.9: Compute HM from the following data of rates of a commodity in Rs.

Rate(Rs)	5-9.9	10-14.9	15-19.9	20-24.9	25-29.9
f	9	15	18	7	4

Solution:

Rate (Rs)	f_i	X_i	$f_i(1/X_i)$
5 – 9.9	9	7.45	1.2080
10 – 14.9	15	12.45	1.2048
15 – 19.9	18	17.45	1.0315
20 – 24.9	7	22.45	0.3118
25 – 29.9	4	27.45	0.1457
Σ	**53**		**3.9018**

$$HM = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} f_i(1/X_i)} = \frac{53}{3.9018} = 13.5834\ Rs$$

3.5.1 Merits of harmonic mean

01) It is well defined.
02) It is based on all the observations.
03) It is amenable to mathematical treatment.
04) It is an appropriate average for averaging ratios rates and speed.
05) It is independent of order of the observations.

3.5.2 Demerits of harmonic mean

01) It is not easy to calculate and understand.
02) It cannot be calculated, if any one of the observation is zero.
03) It gives too much weights to the smaller observations.

Array

Arrangement of values in ascending or descending order of magnitude is called an array.

3.6 Median

Central most value of the arranged data is known as median.

OR

Value that divides the arranged data into two equal parts is called median. It is denoted by $\tilde{X}$ (read as X tilde(*til duh*)).

$Median = \frac{n+1}{2}th value$, for ungrouped data/discrete grouped data

$Median = l + \frac{h}{f}\left(\frac{n}{2} - c\right)$, for grouped data

Where, l = lower class boundary of median class.

h = class interval of median class.

f = frequency of median class.; $n = \sum f$ = Total frequency

c = Cumulative frequency before the median class.

3.6.1 Merits of median

01) It is easy to understand and calculate.
02) Its graphic location is possible, thorough Ogive.
03) It is not affected by extreme values.
04) It is suitable for highly skewed distribution.
05) It can be calculated in open end distribution.

3.6.2 Demerits of median

01) It is not based on all the observations.
02) If data is not arranged, median will not be correct.
03) It is not capable of further mathematical treatment.
04) It cannot give total when multiplied by the number of observations.

3.7 Quantiles

Dividing a set of data into some equal parts are called quantiles. Quartiles, deciles and percentiles etc. are known as quantiles. They are also called fractiles.

3.7.1 Quartiles

Values, that divide an arrayed set of data into four equal parts, are called quartiles. These are denoted by Q_1 (first or lower quartile), Q_2 (2^{nd} quartile), Q_3 (third or upper quartile)

In general

$Q_k = \frac{k(n+1)}{4}th value$, for ungrouped data/discrete grouped data

$Q_k = l + \frac{h}{f}\left(\frac{kn}{4} - c\right)$, for grouped data, where $k = 1, 2, 3$

3.7.2 Deciles

Values, that divide an arrayed set of data into ten equal parts, are called deciles .These are denoted by $D_1, D_2, \ldots, D_9$

$D_k = \frac{k(n+1)}{10}th value$, for ungrouped data/discrete grouped data

$D_k = l + \frac{h}{f}\left(\frac{kn}{10} - c\right)$, for grouped data where $k = 1, 2, \ldots, 9$

3.7.3 Percentiles

Values, that divide an arrayed set of data into hundred equal parts, are called percentiles. These are denoted by $P_1, P_2, \ldots, P_{99}$

$P_k = \frac{k(n+1)}{100}th value$, for ungrouped data/discrete grouped data

$P_k = l + \frac{h}{f}\left(\frac{kn}{100} - c\right)$, for grouped data, where $k = 1, 2, \ldots, 99$

Example 3.10: Find median, Q_1, D_7 and P_{40} from the following data set.
1.0, 4.1, 7.5, 2.9, 11.2, 9.6, 7.3, 8.7, 12.0, 12.5, 2.3, 3.7, 4.7, 12.2, 11.8, 9.1, 8.4, 7.2, 6.8, 10.3

Solution: Arrayed data:
1.0, 2.3, 2.9, 3.7, 4.1, 4.7, 6.8, 7.2, 7.3, 7.5, 8.4, 8.7, 9.1, 9.6, 10.3, 11.2, 11.8, 12.0, 12.2, 12.5

$$Median = \frac{n+1}{2}th\ value = \frac{20+1}{2}th\ value = 10.5th\ value$$

$$\tilde{X} = 10^{th}\ value + 0.5\left(11^{th}\ value - 10^{th}\ value\right)$$

$$\tilde{X} = 7.5 + 0.5(8.4 - 7.5) = 7.95$$

$$Q_1 = \frac{1(n+1)}{4}th\ value = \frac{1(20+1)}{4}th\ value = \frac{21}{4}th\ value$$

$$Q_1 = 5.25^{th}\ value$$

$$Q_1 = 5^{th}\ value + .25\left(6^{th}\ value - 5^{th}\ value\right)$$

$$Q_1 = 4.1 + .25(4.7 - 4.1) = 4.25$$

$$D_7 = \frac{7(n+1)}{10}th\ value = \frac{7(20+1)}{10}th\ value = 14.7th\ value$$

$$D_7 = 14^{th}\ value + 0.7\left(15^{th}\ value - 14^{th}\ value\right)$$

$$D_7 = 9.6 + 0.7(10.3 - 9.6) = 10.09$$

$$P_{40} = \frac{40(n+1)}{100}th\ value = \frac{40(20+1)}{100}th\ value = 8.4th\ value$$

$$P_{40} = 8^{th}\ value + 0.4\left(9^{th}\ value - 8^{th}\ value\right)$$

$$P_{40} = 7.2 + 0.4(7.3 - 7.2) = 7.24$$

Example 3.11: Calculate median, Q_3 and P_{65} from the following data

Classes	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59
Frequency	5	9	17	15	7	2

Solution:

Classes	f	C.B	cf	
0 – 9	5	-0.5 – 9.5	5	
10 – 19	9	9.5 – 19.5	14	
20 – 29	**17**	**19.5 – 29.5**	**31 →**	$n/2$
30 – 39	**15**	**29.5 – 39.5**	**46 →**	$3n/4$ and $65n/100$
40 – 49	7	39.5 – 49.5	53	
50 – 59	2	49.5 – 59.5	55	

$$\tilde{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right), \qquad \frac{n}{2} = \frac{55}{2} = 27.5$$

$$\tilde{X} = 19.5 + \frac{10}{17}(27.5 - 14) = 27.44$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right) \qquad \frac{3n}{4} = \frac{3\times 55}{4} = 41.25$$

$$Q_3 = 29.5 + \frac{10}{15}(41.25 - 31) = 36.33$$

$$P_{65} = l + \frac{h}{f}\left(\frac{65n}{100} - c\right) \qquad \frac{65n}{100} = \frac{65\times 55}{100} = 35.75$$

$$P_{65} = 29.5 + \frac{10}{15}(35.75 - 31) = 32.67$$

Example 3.12: In examination of GAT (general) conducted by NTS, a student obtained the marks 85 with percentile as 90. What it means.
Answer: If percentile of a student in an examination is 90^{th}, it means 90% students perform below this student and only 10% perform well than him/her.

3.8 Mode

Mode is a value that occurs the maximum number of times in the data. It is denoted by $\hat{X}$ (read as X hat).

Note: A set of data may have more than one mode or no mode at all when each observation occur the same number of times.

3.8.1 Methods of calculation of mode

01) Mode(for ungrouped data)
02) In ungrouped data mode is found by inspection .For example the mode of 4, 6, 8, 10, 6 and 3 is 6.
03) Mode for frequency distribution (Discrete) is the value corresponding to the maximum frequency.
04) Mode for frequency distribution (continuous) is, $\hat{X} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$,

Where

l = Lower class boundary of modal class
f_m = The frequency of model class
f_1 = The frequency preceding the model class
f_2 = The frequency following the model class
h = Size of class interval of model class

Example 3.13: Compute mode for the data given previous example 3.11.

Solution:

$$\hat{X} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h = 19.5 + \frac{17-9}{(17-9)+(17-15)} \times 10 = 27.5$$

Example 3.14: Calculate mode

Classes	50 -60	60 – 70	70 – 80	80 – 90	90 – 100
f	2	4	11	3	1

Solution:

Classes	f	
50 – 60	2	
60 – 70	4	f_1
70 – 80	11	f_m
80 – 90	3	f_2
90 – 100	1	

$$\hat{X} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h = 70 + \frac{11-4}{(11-4)+(11-3)} \times 10 = 74.67$$

3.8.2 Merits of mode

01) It is simply defined and easily calculated
02) It is not affected by extreme values
03) It can be calculated in open end frequency distribution.
04) Its value can be located graphically, using histogram.
05) It is the most suitable average to be used for qualitative data.

3.8.3 Demerits of mode

01) It is not well defined.
02) It is not based on all observations
03) It is the not capable of further mathematical treatment.
04) Sometimes a distribution may have no mode or sometimes more than one mode.
05) If there are small number of values in a data set mode may not exist.

Example for Discrete Grouped data

Example 3.15: A survey is conducted for number of children in 20 families, results are given below, calculate median, Q_3 , D_6 , P_{80} and mode.

No. of Children	1	2	3	4	5	6
No. of families	3	2	6	3	4	2

Solution:

X	f	cf	
1	3	3	
2	2	5	
3	6 f_m	11	------- $\tilde{X}$
4	3	14	------- D_6
5	4	18	------- Q_3 / P_{80}
6	2	20	

$\tilde{X} = \frac{n+1}{2}$ th value $= \frac{20+1}{2}$ th value $= 10.5^{th}$ value So, $\tilde{X} = 3$

$Q_1 = \frac{3(n+1)}{4}th\ value = \frac{3(20+1)}{4}th\ value = 15.75^{th}\ value$, So, $Q_1 = 5$

$D_4 = \frac{6(n+1)}{10}th\ value = \frac{6(20+1)}{10}th\ value = 12.6^{th}\ value$, So, $D_4 = 4$

$P_{80} = \frac{80(n+1)}{100}th\ value = \frac{80(20+1)}{100}th\ value = 16.8^{th}\ value$, So, $P_{80} = 5$

Mode: $\hat{X} = 3$ (6 is maximum frequency)

3.9 Choice of suitable average

AM ($\bar{X}$): It is general purpose average and suitable when there is no extreme values.

Median ($\tilde{X}$): When extreme value/s exist in the data / the distribution is skewed and middle most value is required

Mode ($\hat{X}$): It is used to find average in case of qualitative data, and when most common value is required e.g. dress or shoe size etc.

G.M: G.M is used for averaging the rate of change / ratio e.g. to average the rates of increase in population per year

H.M: H.M is used to average certain kinds of ratios and rates of change it is used in averaging speeds for various distances covered.

Note: For averaging rate and speed of moving item.

Condition	Average
Time is same, distance not same	Arithmetic mean
Time is not same, distance same	Harmonic mean
Time is not same, distance is not same	Weighted harmonic mean

Example 3.16: If a car ran for 2 hours at the speed of 70 Km per hour, for next 2 hours its speed was 75 Km/hr and next 2 hours it ran at 100 Km/hr. Find its average speed.

Solution: In this problem times are same but distances are different.

So AM is used for average speed

X_i	70	75	100	$\sum X = 245$

Average speed: $AM = \frac{245}{3} = 81.67 Km/hr$

Example 3.17: If a bus travelled at the speed of 50, 70, 55 and 60 Km/hr for four stages of equal distance, find average speed of that bus.

Solution: In this problem distances are same, but times are different, so for average speed HM is used.

X_i	50	70	55	60	Total
$1/X_i$	0.02	0.0143	0.0182	0.0167	0.0692

Average speed: $HM = \frac{4}{0.0692} = 57.8 Km/hr$

Example 3.18: A train's speed between three different cities was 100, 145 and 130 Km/hr, if distance between these three cities are 100, 60 and 80 Km, find average speed of the train.

Solution: In this problem both times and speeds are not same for three stages, therefore weighted harmonic mean is used for average speed.

W = distance; X = speed

	W_i	X_i	W_i / X_i
	100	100	1
	60	145	0.4138
	80	130	0.6154
Σ	240		2.0292

Average speed:

$$H = \frac{\sum W_i}{\sum (W_i / X_i)} = \frac{240}{2.0292} = 118.27$$

3.10 Empirical relation between mean, median and mode

In a moderately skewed distribution, median remains closer to mean than mode. Median divides the distance between mean and mode in the ratio 1:2.

Hence Mode = 3 Median – 2 Mean i.e. $\hat{X} = 3\tilde{X} - 2\bar{X}$

Example 3.19 Calculate mode by using empirical relation when mean = 10 and median = 12

Solution: As empirical relationship is
Mode = 3Median – 2Mean
Mode = 3(12) – 2(10) = 36 – 20 = 16

Note: Relationship between AM, GM and HM
AM > GM > HM

AM = GM = HM if values of a distribution are same

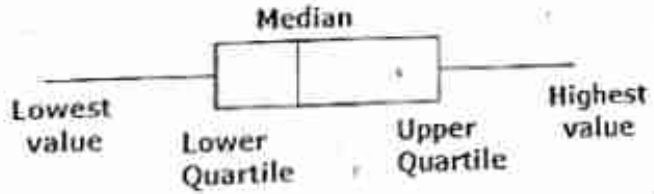
3.11 Five point summary

Five point summary includes

01) Minimum value (X_0)
02) Maximum value (X_m)
03) Lower quartile (Q_1)
04) Upper quartile (Q_3)
05) Median ($\tilde{X}$)

3.11.1 Box and whisker plot

A box-and-whisker plot or box plot is a diagram based on the five-point summary of a data set. Where five-point summary includes minimum value, maximum value, lower quartile, median and upper quartile. It gives a sense of data's distribution.
Draw a box with ends (hinges) through the points for the first and third quartiles. Then draw a vertical line through the box at the median point. Now, draw the whiskers (or lines) from each end of the box to minimum and maximum values.



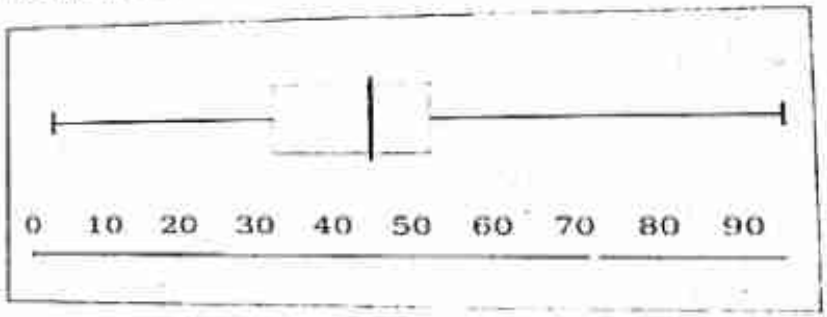
3.11.2 Detailed box plot

An observation of a variable that lies outside the range of other observations is called an outlier or extreme value. If the outlier is included in the whisker then this box plot is called detailed box plot.

Example 3.20: for the following data make boxplot.
3, 30, 35, 45, 50, 55, 40, 50, 95
Solution:
Five point summary:
$X_0 = 3$, $Q_1 = 32.5$, Median = 45, $Q_3 = 52.5$, $X_m = 95$

IQR = 52.5 – 32.5 = 30; 1.5 (IQR) = 1.5(30) = 45
Q_1 – 1.5 IQR = 32.5 – 45 = -12.5
Q_3 + 1.5 IQR = 52.5 + 45 = 97.5; As all values lies within this limit -12.5 to 97.5, so there are no extreme values

Boxplot:



Multiple Choice Questions

1. Median divides, arranged data into:
(a) 2 parts (b) 4 parts (c) 10 parts (d) 100 parts

2. The arithmetic mean of "n" values $Y_1, Y_2, \ldots, Y_n$:
(a) $\frac{Y_1 + Y_n}{2}$ (b) $\frac{Y_1 + Y_n}{2}$ (c) $\frac{1}{n}\sum Y$ (d) $\left[\prod(Y)\right]$

3. The sum of values divided by their numbers:
(a) Mode (b) Median (c) Mean (d) G.M

4. If $Y_i = aX_i + b$, then Mode (Y_i)
(a) aMode (X_i) + b (b) Mode (X_i) + b (c) Mode (X_i) (d) aMode (X_i)

5. The observation, occurs maximum numbers of times:
(a) Mean (b) Median (c) Mode (d) H.M

6. The mean is affected by the change of:
(a) Origin (b) Scale (c) Both (a) and (b) (d) Method

7. The suitable average for averaging, shoes size:
(a) Mean (b) Mode (c) Geometric mean (d) Median

8. For a symmetrical distribution, mean = 100, then which is correct?
(a) Median = 50, Mode = 50 (b) Median = 0, Mode = 100
(c) Median = 50, Mode =150 (d) Median = 100, Mode = 100

9. Have more than one value? (a) AM (b) HM (c) GM (d) Mode

10. Suitable average for averaging speed of a journey:
(a) AM (b) GM (c) HM (d) Median

11. Based on the idea of fifty-50:
(a) AM (b) GM (c) Median (d) Mode

12. For a symmetrical distribution:
(a) Mean = Median = Mode (b) Mode = 3Median-2Mean
(c) Median = 3Mean - Mode (d) Mode = 2Median-3Mean

13. If 10 % is added to each value of a variable, then AM is increased by
(a) 10% (b) 110 (c) 100 (d) 10

14. The mean is based on:
(a) All the values (b) Small values (c) Large values (d) Extreme values

15. In a symmetrical distribution, mean, median and mode are:
(a) Zero (b) coincide (c) Not equal (d) Not identical

16. If $Z = X - Y$, then $\bar{Z}$ is: (a) $\bar{X} - \bar{Y}$ (b) $X - Y$ (c) $\bar{X} - Y$ (d) $X - \bar{Y}$

17. The mean of a constant: (a) 0 (b) constant (c) not possible (d) 1

18. Sum of deviations of observations is zero, deviations are taken from:
(a) Mean (b) Median (c) Mode (d) G.M

19. In a symmetrical distribution $Q_1 = 4$, $Q_3 = 12$, then median:
(a) 4 (b) 8 (c) 12 (d) 16

20. If $\bar{X}$, $\tilde{X}$ and $\hat{X}$ are identical, distribution is:
(a) Positively Skewed (b) Negatively Skewed (c) Symmetrical (d) asymmetrical

21. If $\bar{X} = 10$ and $Y = 5 - 2X$, then $\bar{Y}$ is: (a) 3 (b) 5 (c) 15 (d) -15

22. The sum of the deviations of observations from mean is:
(a) Zero (b) least (c) Maximum (d) Positive

23. If "a" and "b" are two positive numbers, geometric mean:
(a) $\sqrt{a-b}$ (b) $\sqrt{a-b}$ (c) $\sqrt{ab}$ (d) $a\sqrt{b}$

24. The modal letter of the "STATISTICS" is:
(a) S (b) T (c) [illegible] (d) [illegible] T

25. Arithmetic mean for X_1 and X_2:

(a) $\sqrt{X_1X_2}$ (b) $\frac{X_1+X_2}{2}$ (c) $\frac{2}{X_1+X_2}$ (d) $\frac{X_1+X_2}{n}$

26. Sum of the absolute deviations of the values from ---- is least.

(a) Mean (b) Median (c) G.M (d) Mode

27. In a moderately skewed distribution mean = 120, median = 110, mode is:

(a) 50 (b) 90 (c) 140 (d) 235

28. Distribution has two modes:

(a) Uni-model (b) Bi-model (c) Tri-model (d) Multi-model

29. If all items are not of equal importance you will prefer

(a) AM (b) Median (c) Mode (d) Weighted mean

30. If any value in the data is zero, it is impossible:

(a) A.M (b) G.M (c) Mode (d) HM

31. If any value in the data is zero, then average vanishes:

(a) A.M (b) G.M (c) Mode (d) HM

32. Data is 2, 3, 7, 0 and 8, GM:

(a) Negative (b) Positive (c) zero (d) Undefined

33. Step deviation or coding method is for:

(a) Median (b) GM (c) AM (d) HM

34. Not based upon all the observations:

(a) A.M (b) G.M (c) H.M (d) Mode

35. Reciprocal of AM of reciprocal of the observations:

(a) Mean (b) Median (c) H.M (d) Mode

36. Must arrange the data for:

(a) Mode (b) Median (c) Mean (d) G.M

37. For symmetrical distribution; Mean.....Median.....Mode:

(a) = (b) < (c) > (d) ≠

38. In a symmetrical distribution Q_1=20, Median=30, then Q_3 =

(a) 20 (b) 30 (c) 40 (d) 50

39. Upper quartile Q_3 = (a) P_{33} (b) D_3 (c) P_{75} (d) Median

40. For open end distribution, not possible to find:

(a) AM (b) Mode (c) Mode (d) Quantiles

41. AM of ten numbers is 9.2, then sum of observations:

(a) 72 (b) 82 (c) 92 (d) 102

42. The most central value of arranged data:

(a) AM (b) Mode (c) Median (d) G.M

43. The mode of 2, 3, 3, 3, 4, 4, 5 and 6: (a) 2 (b) 3 (c) 4 (d) 5

44. The mean of two observations is 10.5 then median:

(a) 10 (b) 10.5 (c) 11 (d) 21

45. The mean of 10, 10, 10, 10, 10 and 10:

(a) 0 (b) 1 (c) 10 (d) 60

46. Geometric mean of 2, 4, 8:

(a) 2 (b) 4 (c) 8 (d) 64

47. In stem and leaf display diagrams, stems are:

(a) Central digits (b) trailing digits

(c) Leading digits (d) dispersed digits

key

Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans
1	a	2	c	3	c	4	a	5	c	6	c
7	b	8	d	9	d	10	c	11	c	12	a
13	a	14	a	15	b	16	a	17	b	18	a
19	b	20	c	21	d	22	a	23	c	24	d
25	b	26	b	27	b	28	b	29	d	30	d
31	b	32	c	33	c	34	d	35	c	36	b
37	a	38	c	39	c	40	a	41	c	42	c
43	b	44	b	45	c	46	b	47	c		

Exercise

Q No. 3.1: (a) Define measure of central tendency, average, arithmetic mean, weighted mean, geometric mean, harmonic mean, median and mode.
(b) Describe the properties of arithmetic mean.

Q No. 3.2: Find mean, median and mode: 1, 0, 4, 6, 4, 2, 3, 7, 6, 9, 4, 8, 0, 1

Q No. 3.3: Find AM from the following calculations.

(i) $n = 50, \sum D = 90$ and $A = 65$

(ii) $n = 20, \sum D = 14$ and $D = X - 50$

(iii) $n = 100, \sum D = -25.9$ and $A = 15.5$

(iv) $n = 40, \sum u = 10.5$ and $A = 15$, h = 10

(v) $\sum f = 60; \sum fD = 9.5$ and $D = X - 80$

(vi) $\sum f = 100; \sum fu = -15$ and $u = \frac{X - 50}{5}$

Q No. 3.4: For a certain frequency distribution, the mean was 51 and median 42. Find mode approximately using the formula connecting these three measurements.

Q No. 3.5: Find AM and median.

Weight (kg)	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59
f	4	17	36	48	27	6

Q No. 3.6: Find median and mode.

Income (000) in Rs.	0 – 4.9	5 – 9.9	10 – 14.9	15 – 19.9	20 – 24.9
No of employees	10	18	22	14	5

Q No. 3.7: Find mean, median, and mode.

Classes	0 – 9.9	10 – 19.9	20 – 29.9	30 – 39.9	40 – 49.9
f	5	7	15	10	6

Q No. 3.8: Find Arithmetic mean by (i) Direct (ii) Indirect and (iii) Coding method.

Classes	1 – 10	11 – 20	21 – 30	31 – 40	41 – 50
f	5	10	18	6	2

Q No. 3.9: Find Arithmetic mean by all three techniques.

Weights(Kg)	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49
f	6	10	19	12	4

Q No. 3.10: The number of contaminating particles on a silicon wafer prior to a certain rinsing process was determined for each wafer in a sample, resulting in the following frequencies:

Number of particles	0	1	2	3	4	5
Frequency	1	2	3	12	11	15

Find mean, median and mode.

Q No. 3.11: Following are the observations on weights of 20 week old mallard ducks. 3.61, 3.75, 3.79, 3.82, 3.85, 3.87, 3.90, 3.94, 3.96, 3.99, 3.99, 4.00, 4.03, 4.04, 4.05, 4.06, 4.09, 3.97, 3.98 and 3.77 kilograms.

(i) Calculate mean, median, mode, lower and upper quartile (ii) make its box and whisker plot.

Comment on the distribution of weights of these ducks.

Q No. 3.12: Find mean, median and mode of the amount of phosphorus in leaves

Phosphorus (mg/g)	8.15 – 8.35	8.35 – 8.55	8.55 – 8.75	8.75 – 8.95	8.95 – 9.15
Frequency	8	19	34	25	11

Q No. 3.13: Consider the following frequency tabulation of leaf weight (grams), calculate AM, median and mode.

Weights (Grams)	1.85 – 2.04	2.05 – 2.24	2.25 – 2.44	2.45 – 2.64	2.65 – 2.84
Frequency	3	5	11	7	1

Q No. 3.14: Find mode.

Classes	1 – 10	11 – 20	21 – 30	31 – 40	41 – 50
f	5	10	18	6	2

Q No. 3.15: stress measurement, on a scale of 0 to 10, with 0 being not stressed and 10 being as stressed as possible. Scores from 30 students of a certain college are: 8, 7, 4, 10, 8, 6, 8, 9, 9, 7, 3, 7, 6, 5, 0, 9, 10, 7, 7, 3, 6, 7, 5, 2, 1, 6, 7, 10, 8, 8. Find arithmetic mean, median and mode.

Q No. 3.16: Six participants in a particular MRI (brain scan) study are measured for the increase in activation of their amygdala while they are viewing pictures of violent scenes. The activation increases are 0.43, 0.32, 0.64, 0.21, 0.29, and 0.51. Find mean for these six activation increases.

Q No. 3.17: Number of words that seven infants have learned at a particular age. The numbers are 11, 10, 8, 2, 3, 15, and 9. Find mean, median, and mode for the number of words learned by these seven infants.

Q No. 3.18: Duration (days) governments for prime ministers of Pakistan are as follows 1525, 549, 848, 398, 401, 61, 296, 14, 1422, 1163, 613, 986, 1113, 967, 582, 57, 1174, 1495, 275, 1514 and 303. Compute AM, Median and Mode and select the most representative answer.

Q No. 3.19: Find AM, median and mode from the following distribution of marks obtained by students of BS Level in the course of statistics.

Marks	31 – 40	41 – 50	51 – 60	61 – 70	71 – 80
Number of students	3	7	12	6	2

Solution

Q No. 3.2:

Mean	$\sum X = 55, n = 14$	$\bar{X} = \frac{\sum X}{n} = \frac{55}{14} = 3.93$	Median	4	Mode	4

Q No. 3.3:

Sr.	Calculation of mean($\bar{X}$)	Sr.	Calculation of mean($\bar{X}$)
(i)	$A + \frac{\sum D}{n} = 65 + \frac{90}{50} = 66.8$	(ii)	$A + \frac{\sum D}{n} = 50 + \frac{14}{20} = 50.7$
(iii)	$A + \frac{\sum D}{n} = 15.5 + \frac{-25.9}{100} = 15.24$	(iv)	$A + \frac{\sum u}{n} \times h = 15 + \frac{10.5}{40} \times 10 = 17.63$
(v)	$A + \frac{\sum fD}{\sum f} = 80 + \frac{9.5}{60} = 80.16$	(vi)	$A + \frac{\sum fu}{\sum f} \times h = 50 + \frac{-15}{100} \times 5 = 49.25$

Q No. 3.4: Mode = 3 median − 2mean = 3(42) − 2(51) =126 − 102 = 24

Q No. 3.5:

Weights(kg)	f	X	fX	C.B.	Cf
0 – 9	4	4.5	18	-0.5 – 9.5	4
10 – 19	17	14.5	246.5	9.5 – 19.5	21
20 – 29	36	24.5	882	19.5 – 29.5	57
30 – 39	48	34.5	1656	29.5 – 39.5	105
40 – 49	27	44.5	1201.5	39.5 – 49.5	132
50 – 59	6	54.5	327	49.5 – 59.5	138
Sum(Σ)	**138**		**4331**		

Computation: $\bar{X} = \frac{\sum fX}{\sum f} = \frac{4331}{138} = 31.8$ kg, $\tilde{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right); \frac{n}{2} = \frac{138}{2} = 69$ Kg, $\hat{X} = 29.5 + \frac{10}{38}(69 - 57) = 32$ kg

Q No. 3.6:

Income(000) (Rs)	f	C.B.	Cf
0 – 4.9	10	-0.05 – 4.95	10
5 – 9.9	18	4.95 – 9.95	28

10 – 14.9	22	9.95 – 14.95	50
15 – 19.9	14	14.95 – 19.95	64
20 – 24.9	5	19.95 – 24.95	69
Sum(Σ)	**69**		

Computation: $\tilde{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right); \frac{n}{2} = \frac{69}{2} = 34.5$ *Rs*,

$\tilde{X} = 9.95 + \frac{5}{22}(34.5 - 28) = 11.43$ *Rs*,

$\hat{X} = l + \frac{f_m - f_1}{f_m - f_1 + f_m - f_2} \times h = 9.95 + \frac{22-18}{22-18+22-14} \times 5 = 11.62$ Rs

Q No. 3.7:

Classes	f	X	fX	C.B.	Cf
0 – 9.9	5	4.95	24.75	-0.05 –	5
10 – 19.9	7	14.95	104.65	9.95 –	12
20 – 29.9	15	24.95	374.25	19.95 –	27
30 – 39.9	10	34.95	349.50	29.95 –	37
40 – 49.9	6	44.95	269.70	39.95 –	43
Sum(Σ)	**43**		**1122.85**		

Computation

$\bar{X} = \frac{\sum fX}{\sum f} = \frac{1122.85}{43} = 26.11$

$\tilde{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right); \frac{n}{2} = \frac{43}{2} = 21.5$

$= 19.95 + \frac{10}{15}(21.5 - 12) = 26.28$

$\hat{X} = l + \frac{f_m - f_1}{f_m - f_1 + f_m - f_2} \times h = 19.95 + \frac{15-7}{15-7+15-10} \times 10 = 26.10$

Q No. 3.8:

Classes	f	X	fX	D	fD	u	fu
1 – 10	5	5.5	27.5	-20	-100	-2	-10
11 – 20	10	15.5	155	-10	-100	-1	-10
21 – 30	18	25.5	459	0	0	0	0
31 – 40	6	35.5	213	10	60	1	6
41 – 50	2	45.5	91	20	40	2	4
Sum(Σ)	**41**		**945.5**		**-100**		**-10**

$\bar{X} = \frac{\sum fX}{\sum f} = \frac{945.55}{41} = 23.06$

$$\bar{X} = A + \frac{\sum fD}{\sum f} = 25.5 + \frac{-100}{41} = 23.06; where\ D = X - 25.5$$

$$\bar{X} = A + \frac{\sum fu}{\sum f} \times h = 25.5 + \frac{-10}{41} \times 10 = 23.06$$

Q No. 3.9:

Weights(kg)	f	X	fX	D	fD	u	fu
0 – 9	6	4.5	27	-20	-120	-2	-12
10 – 19	10	14.5	145	-10	-100	-1	-10
20 – 29	19	24.5	465.5	0	0	0	0
30 – 39	12	34.5	414	10	120	1	12
40 – 49	4	44.5	178	20	80	2	8
Sum(Σ)	**51**		**1229.5**		**-20**		**-2**

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{1229.5}{51} = 24.11\ Kg$$

$$\bar{X} = A + \frac{\sum fD}{\sum f} = 24.5 + \frac{-20}{51} = 24.11\ Kg$$

$$\bar{X} = A + \frac{\sum fu}{\sum f} \times h = 24.5 + \frac{-2}{51} \times 10 = 24.11\ Kg$$

Q No. 3.10:

X	0	1	2	3	4	5
f	1	2	3	12	11	15
fX	0	2	6	36	44	75
cf	1	3	6	18	29	44

n/2

Mean	$\sum fX = 163, \sum f = 44$	$\bar{X} = \frac{\sum fX}{\sum f} = \frac{163}{44} = 3.70$	Median	4	Mode	5

Q No. 3.11:

Arranged data: 3.61, 3.75, 3.77, 3.79, 3.82, 3.85, 3.87, 3.90, 3.94, 3.96, 3.97, 3.99, 3.99, 3.99, 4.00, 4.03, 4.04, 4.05, 4.06, 4.09,

Mean: $\bar{X} = \frac{\sum fx}{\sum f} = \frac{78.46}{20} = 3.923$

median: $= median = \frac{n+1}{2}th\ value = \frac{21}{2}th\ value = 10.5th\ value$

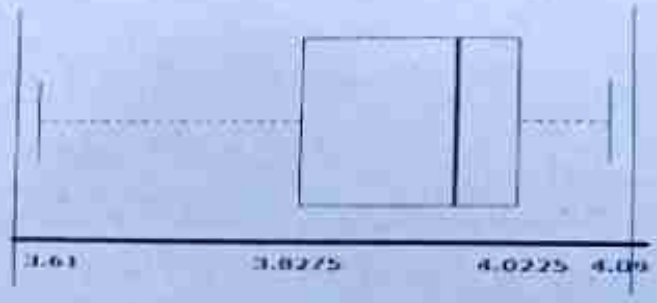
$median = 10th + 0.5(11th - 10th) = 3.96 + 0.5(3.97 - 3.96) = 3.965$

Lower quartile: $= Q_1 = \frac{n+1}{4}th\ value = \frac{21}{4}th\ value = 5.25th\ value$

$Q_1 = 5th + 0.25(6th - 5th) = 3.82 + 0.25(3.85 - 3.82) = 3.8275$

Upper quartile: $= Q_3 = \frac{3(n+1)}{4}th\ value = \frac{63}{4}th\ value = 15.75th\ value$

$Q_3 = 15th + 0.75(16th - 15th) = 4.00 + 0.75(4.03 - 4.00) = 4.0225$



Q No. 3.12:

Phosphorus(mg/g)	f	X	u	fu	cf
8.15 – 8.35	8	8.25	-2	-16	8
8.35 – 8.55	19	8.45	-1	-19	27
8.55 – 8.75	34	8.65	0	0	61
8.75 – 8.95	25	8.85	1	25	86
8.95 – 9.15	11	9.05	2	22	97
Total	**97**			**12**	

n/2=48.5

$$\bar{X} = A + \frac{\sum fu}{\sum f} \times h = 8.65 + \frac{12}{97} \times 0.2 = 8.67\frac{mg}{g}; where\ u = \frac{X - 8.65}{0.20}$$

$$\bar{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right) = 8.55 + \frac{0.2}{34}(48.5 - 27) = 8.68\ mg/g$$

$$\bar{X} = l + \frac{f_m - f_1}{f_m - f_1 + f_m - f_2} \times h = 8.55 + \frac{34 - 19}{34 - 19 + 34 - 25} \times 0.2 = 8.675\ mg/g$$

Q No. 3.13:

Weights (grams)	f	X	u	fu	C.B.	cf	
1.85 – 2.04	3	1.945	-2	-6	1.845 –	3	
2.05 – 2.24	5	2.145	-1	-5	2.045 –	8	
2.25 – 2.44	11	2.345	0	0	2.245 –	19	n/2=13.5
2.45 – 2.64	7	2.545	1	7	2.445 –	26	
2.65 – 2.84	1	2.745	2	2	2.645 –	27	
Total	**27**			**-2**			

$$\bar{X} = A + \frac{\sum fu}{\sum f} \times h = 2.345 + \frac{-2}{27} \times 0.2 = 2.33 \; grams$$

$$\bar{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right) = 2.245 + \frac{0.2}{11}(13.5 - 8) = 2.345 \; grams$$

$$\bar{X} = l + \frac{f_m - f_1}{f_m - f_1 + f_m - f_2} \times h = 2.245 + \frac{11 - 5}{11 - 5 + 11 - 7} \times 0.2 = 2.365 \; grams$$

Q No. 3.14:

Classes	f	C.B.	
1 – 10	5	0.5 –	$\bar{X} = l + \frac{f_m - f_1}{f_m - f_1 + f_m - f_2} \times h$
11 – 20	10	10.5 –	$= 20.5 + \frac{18 - 10}{18 - 10 + 18 - 6} \times 10$
21 – 30	18	20.5 –	$= 24.5 \; grams$
31 – 40	6	30.5 –	
41 – 50	2	40.5 –	
Total	**41**		

Q No. 3.15:

X	0	1	2	3	4	5	6	7	8	9	10	**Total**
f	1	1	1	2	1	2	4	7	5	3	3	**30**
fX	0	1	2	6	4	10	24	49	40	27	30	**193**
cf	1	2	3	5	6	8	12	19	24	27	30	
								15.5th value				

$$\bar{X} = \frac{\sum fx}{\sum f} = \frac{193}{30} = 6.43$$

$$median = \frac{n+1}{2} th \; value = \frac{31}{2} th \; value = 15.5 th \; value; \; \bar{X} = 7$$

$$mode: \bar{X} = 7$$

Q No. 3.16: $\bar{X} = \frac{\sum X}{n} = \frac{24}{6} = 0.4$

Q No. 3.17: $\bar{X} = \frac{\sum X}{n} = \frac{58}{7} = 8.29$

$median = \frac{n+1}{2} th \; value = \frac{8}{2} th \; value = 4th \; value; \; \bar{X} = 9$, $mode: No \; answer$

Q No. 3.18: 14, 57, 61, 275, 296, 303, 398, 401, 549, 582, 613, 848, 967, 986, 1113, 1163, 1174, 1422, 1495, 1514, 1525

$$\bar{X} = \frac{\sum X}{n} = \frac{15756}{21} = 750.29$$

$$median = \frac{n+1}{2} th \; value = \frac{22}{2} th \; value = 11th \; value; \; \bar{X} = 613$$

$$mode: No \; answer$$

Q No. 3.19:

Marks	f	X	u	fu	C.B.	cf	
31 – 40	3	35.5	-2	-6	30.5 –	3	
41 – 50	7	45.5	-1	-7	40.5 –	10	
51 – 60	12	55.5	0	0	50.5 –	22	n/2=15
61 – 70	6	65.5	1	6	60.5 –	28	
71 – 80	2	75.5	2	4	70.5 –	30	
Total	**30**			**-3**			

$$\bar{X} = A + \frac{\sum fu}{\sum f} \times h = 55.5 + \frac{-3}{30} \times 10 = 54.5 \; marks$$

$$\bar{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right) = 50.5 + \frac{10}{12}(15 - 10) = 54.67 \; marks$$

$$\bar{X} = l + \frac{f_m - f_1}{f_m - f_1 + f_m - f_2} \times h = 50.5 + \frac{12 - 7}{12 - 7 + 12 - 6} \times 10 = 55.05 \; marks$$

MEASURES OF DISPERSION

Basic Terms used in statistics are well known to a common person

Learning Goals

(i) To develop the skills to analyze the data by finding its representative value.
(ii) To understand and interpret the distribution of the data.
(iii) To analyze the data through measure of scatterings, skewness and kurtosis.
(iv) To compare individuals in a data set by Z-score.

4.1 Dispersion

The scattering of the values of a distribution of the data from an average is called dispersion.

4.2 Measure of dispersion

A measure of dispersion expresses quantitatively the degree of variation or dispersion of values of a variable about any average. OR any measure that indicates how the observations are spread out from an average is called a measure of dispersion.

4.2.1 Types of measure of dispersion

There are two types of dispersion
(i) Absolute measure of dispersion and (ii) Relative measure of dispersion

4.2.2 Absolute and relative measure of dispersion

The measures of dispersion which are expressed in terms of original units of a data are termed as absolute measures of dispersion. Relative measures of dispersion also known as coefficients of dispersion, are obtained as ratios or percentages. These are pure numbers independent of the units of measurement and used to compare two or more sets of data values.

4.2.3 Methods of measure of dispersion

01) Range (R)
02) Semi inter quartile range/Quartile deviation (SIQR/QD)
03) Mean deviation/Average deviation (MD)
04) Variance (S^2) and Standard deviation (S)

4.2.3.1 Range

Difference between maximum and minimum value of the data is called range. *It is* given as $R = X_m - X_0$

Coefficient of range $= \frac{X_m - X_0}{X_m + X_0}$

Example 4.1 Following are the values of a variable X
5, 3, 7, 9, 11, 15, 18, 19, 16, and 20. Find range and coefficient of range.

Solution: X: 5, 3, 7, 9, 11, 15, 18, 19, 16, 20

$X_m = 20$ and $X_0 = 3$

$Range = X_m - X_0 = 20 - 3 = 17$

Coefficient of range: $Co-R = \frac{X_m - X_0}{X_m + X_0} = \frac{20-3}{20+3} = \frac{17}{23} = 0.739$

Example 4.2 Find range and coefficient of range from the following frequency distribution.

Classes	10 – 20	20 – 30	30 – 40	40 – 50
Frequency	7	12	15	2

Solution:

Classes	10 – 20	20 – 30	30 – 40	40 – 50
Frequency	7	12	15	2

X_m = upper class boundary of the highest class, $X_m = 50$

X_0 = lower class boundary of the lowest class, $X_0 = 10$

Range:

$R = X_m - X_0 = 50 - 10 = 40$

Coefficient of range: $Co-R = \frac{X_m - X_0}{X_m + X_0} = \frac{50-10}{50+10} = \frac{40}{60} = 0.667$

Range can also be calculated by using two extreme mid points.

4.2.3.2 Quartile Deviation (QD) / Semi inter quartile range (SIQR)

Half of the difference between upper and lower quartiles (Q_3 and Q_1) is called semi inter quartile range or simply quartile deviation (QD). It is given as $QD = \frac{Q_3 - Q_1}{2}$.

Coefficient of QD $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Example 4.3: Calculate QD and its coefficient from the data 6, 4, 2, 8, 10, 14, 12

Solution: Arranging the data: 2,4,6,8,10,12,14 $n = 7$

$Q_1 = \frac{n+1}{4}th\ value = \frac{7+1}{4}th\ value = 2^{nd}\ value\ \ Q_1 = 4$

$Q_3 = \frac{3(n+1)}{4}th\ value = \frac{3(7+1)}{4}th\ value = 6^{th}\ value\ \ Q_3 = 12$

$QD = \frac{Q_3 - Q_1}{2} = \frac{12-4}{2} = 4$

$Co-QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{12-4}{12+4} = 0.5$

Example 4.4: Calculate QD and its coefficient from the following data

Classes	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59
Frequency	5	9	17	15	7	2

Solution:

Classes	f	C.B	cf	
0 – 9	5	-0.05 – 9.5	5	
10 – 19	**9**	**9.5 – 19.5**	**14** →	$n/4$
20 – 29	17	19.5 – 29.5	31	
30 – 39	**15**	**29.5 – 39.5**	**46** →	$3n/4$
40 – 49	7	39.5 – 49.5	53	
50 – 59	2	49.5 – 59.5	55	

$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right)$; $\frac{n}{4} = 13.75$

$Q_1 = 9.5 + \frac{10}{9}(13.75 - 5) = 19.22$

$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$; $\frac{3n}{4} = \frac{3 \times 55}{4} = 41.25$

$Q_3 = 29.5 + \frac{10}{15}(41.25 - 31) = 36.33$

$QD = \frac{Q_3 - Q_1}{2} = \frac{36.33 - 19.22}{2} = 8.55$

$CO-QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{36.33 - 19.22}{36.33 + 19.22} = 0.31$

4.2.3.3 Mean deviation (Average deviation)

Mean deviation is defined as the AM of the absolute deviations of the observation from their any average (*mean, median or mode*). It is also known as average deviation and is given as $MD_{m} = \frac{\sum f|X - m|}{\sum f}$ and Coefficient of MD $= \frac{MD}{m}$

Example 4.5: calculate mean deviation from mean, median and mode and their coefficients 3, 4, 7, 7, 8, 3, 5, 3.

Solution:

X	$\lvert X-\bar{X}\rvert$	$\lvert X-\tilde{X}\rvert$	$\lvert X-\hat{X}\rvert$
3	2	1.5	0
4	1	0.5	1
7	2	2.5	4
7	2	2.5	4
8	3	3.5	5
3	2	1.5	0
5	0	0.5	2
3	2	1.5	0
$\sum X_i = 40$	$\sum\lvert X-\bar{X}\rvert = 14$	$\sum\lvert X-\tilde{X}\rvert = 14$	$\sum\lvert X-\hat{X}\rvert = 16$

$$\bar{X} = \frac{\sum X_i}{n} = \frac{40}{8} = 5\,marks$$

Median:

3, 3, 3, 4, 5, 7, 7, 8 n = 8

$\tilde{X} = 4.5$

Mode: $\hat{X} = 3$

Mean Deviations and their coefficients:

$$MD_{(mean)} = \frac{\sum|X-\bar{X}|}{n} = \frac{14}{8} = 1.75\,marks$$

$$Co-MD_{(mean)} = \frac{MD}{mean} = \frac{1.75}{5} = 0.35$$

$$MD_{(median)} = \frac{\sum|X-\tilde{X}|}{n} = \frac{14}{8} = 1.75\,marks$$

$$Co-MD_{(median)} = \frac{MD}{median} = \frac{1.75}{4.5} = 0.38$$

$$MD_{(mode)} = \frac{\sum|X-\hat{X}|}{n} = \frac{16}{8} = 2\,marks$$

$$Co-MD_{(mode)} = \frac{MD}{mode} = \frac{2}{3} = 0.67$$

Example 4.6: Compute mean deviation from mean, median and mode from the following data of marks in statistics of BS (H) mathematics. Also calculate coefficients for each.

Marks	0 – 19	20 – 39	40 – 59	60 – 79	80 – 99
No. of students	6	10	18	12	4

Solution:

Marks	f_i	X_i	$f_i\lvert X_i-\bar{X}\rvert$	$f_i\lvert X_i-\tilde{X}\rvert$	$f_i\lvert X_i-\hat{X}\rvert$
0 – 19	6	9.5	235.2	240	248.58
20 – 39	10	29.5	192	200	214.3
40 – 59	18	49.5	14.4	0	25.74
60 – 79	12	69.5	249.6	240	222.84
80 – 99	4	89.5	163.2	160	154.28
Σ	**50**		**854.4**	**840**	**865.74**

Calculations of mean, median and mode discussed in the previous chapter are

$\bar{X} = 48.7, \tilde{X} = 49.5$ and $\hat{X} = 50.93$

Mean Deviations

$$MD_{mean} = \frac{\sum f_i|X_i-\bar{X}|}{\sum f_i} = \frac{854.4}{50} = 17.08\,marks$$

$$Co-MD_{mean} = \frac{MD}{mean} = \frac{17.08}{48.7} = 0.35$$

$$MD_{median} = \frac{\sum f_i |X_i - \tilde{X}|}{\sum f_i} = \frac{840}{50} = 16.8\,marks$$

$$Co-MD_{median} = \frac{MD}{\text{median}} = \frac{16.8}{49.5} = 0.34$$

$$MD_{mode} = \frac{\sum f_i |X_i - \hat{X}|}{\sum f_i} = \frac{865.74}{50} = 17.31\,marks$$

$$Co-MD_{mode} = \frac{MD}{\text{mode}} = \frac{17.31}{50.93} = 0.33$$

4.2.3.4 Variance

Variance is defined as the AM of the squared deviations of the observation from their mean. It is given as

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n}, \qquad \text{For ungrouped data}$$

$$S^2 = \frac{\sum f_i (X_i - \bar{X})^2}{\sum f_i}, \qquad \text{For grouped data}$$

Direct method

$$S^2 = \frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2; \qquad S^2 = \frac{\sum fX^2}{\sum f} - \left(\frac{\sum fX}{\sum f}\right)^2$$

Short cut method

$$S^2 = \frac{\sum D^2}{n} - \left(\frac{\sum D}{n}\right)^2; \qquad S^2 = \frac{\sum fD^2}{\sum f} - \left(\frac{\sum fD}{\sum f}\right)^2$$

Coding / Step deviation method

$$S^2 = \left\{\frac{\sum u^2}{n} - \left(\frac{\sum u}{n}\right)^2\right\} h^2; \qquad S^2 = \left\{\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2\right\} h^2$$

4.2.3.5 Properties of variance

(i) Variance of a constant is zero. Var(c) = 0

(ii) Variance is not affected by change of origin, that is if

$Y = X + b$ Then Var(Y)=Var(X)

(iii) Variance is affected by change of scale, that is if $Y = aX$ then

$$\text{Var}(Y) = a^2\,\text{Var}(X)$$

(iv) Var(X±Y)=Var(X)+Var(Y) if X and Y are independent

(v) Combined variance:

If a distribution consists of k components with $n_1, n_2, \ldots, n_k$ observations with $\sum n_j = n$ having means $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ and variances $S_1^2, S_2^2, \ldots, S_k^2$, then the combined variance S^2 of all n observations is given by $S_c^2 = \frac{\sum_{j=1}^{k} n_j \left[S_j^2 + (\bar{X}_j - \bar{X}_c)^2\right]}{\sum_{j=1}^{k} n_i}$

OR

$S_c^2 = \frac{n_1[S_1^2 + (\bar{X}_1 - \bar{X}_c)^2] + n_2[S_2^2 + (\bar{X}_2 - \bar{X}_c)^2] + \ldots + n_k[S_k^2 + (\bar{X}_k - \bar{X}_c)^2]}{n_1 + n_2 + \ldots + n_k}$ Where

$$\bar{X}_c = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij} = \frac{1}{n}\sum_{j=1}^{k} n_j \bar{X}_j$$

4.2.3.6 Standard deviation

Standard deviation (SD) is defined as the positive square root of AM of the squared deviations of the observation from their mean. It is given as

$S = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n}}$, for ungrouped data

$S = \sqrt{\frac{\Sigma f(X - \bar{X})^2}{\Sigma f}}$, for grouped data

4.2.3.7 Properties of SD:

(i) $SD(C) = 0$, Where C is a constant.

(ii) $SD(aX_i) = |a| SD(X_i)$, Where a is a constant.

(iii) $SD(X_i + b) = SD(X_i)$

(iv) $SD(X_i + Y_i) = \sqrt{Var(X_i) + Var(Y_i)}$

(v) $SD(X_i - Y_i) = \sqrt{Var(X_i) + Var(Y_i)}$;

Provided X_i and Y_i are independent variables.

4.2.3.8 Coefficient of variation

Coefficient of variation (CV) is defined as ratio of SD of a variable to its mean ($\bar{X}$), expressed in percentage. i.e. $CV = \frac{S}{\bar{X}} \times 100$

4.2.3.9 Coefficient of standard deviation

Coefficient of standard deviation is defined as ratio of SD of a variable to its mean ($\bar{X}$) i.e. $Co-efficient\ SD = \frac{S}{\bar{X}}$

4.2.3.10 Uses of Dispersion

It tells about the reliability of a measure of central value.

It makes possible to compare two sets of data about their variability.

4.3 Z-scores / Standard scores

The value of a normally distributed random variable that has been standardized to have a mean of zero and a SD of one by the transformation $Z = \frac{X_i - \mu}{\sigma}$ is called Z score, for

sample data $Z = \frac{X_i - \bar{X}}{s}$. A "Z" score is the number of standard deviations an observation is above or below the mean. It gives measure of dispersion of an individual observation in standard deviation unit. Z-scores indicate the direction (+/-) and time of standard deviations away from the mean that a particular datum lies assuming X is normally distributed.

It helps to indicate outlier. A value with Z-score greater than 3 or less than -3 is considered as an outlier/extreme value. A value with Z-score greater than 2 and less than 3 or less than -2 and greater than -3 is considered as influential value.

Example 4.7: Mean number of times a child speak is 12 and the standard deviation is 4, find Z score of a child who spoke to other children 8 times in an hour.

Solution: $\mu = 12$, $\sigma = 4$ and X = 8 then

$Z = \frac{X_i - \mu}{\sigma} = \frac{8 - 12}{4} = -1.0$ It means child is normal

Example 4.8: Student's Score, Class Means, Class Standard Deviations, and z Scores on four different courses are given as under

Subject	Raw Score	Mean	SD	Z-Score
English	85	70	14	+1.07
Statistics	57	63	12	-0.50
Mathematics	65	72	16	-0.44
Economics	80	50	15	+2.00

As above table shows that Z-Score in economics is highest i.e. +2.00, it is concluded that level of the knowledge in economics of this particular student is high related to his other class fellows.

Z-Score in statistics is lowest i.e. -0.50, it is concluded that level of the knowledge in statistics of this particular student is low related to his other class fellows.

Example 4.9: Calculate variance, SD and CV from the following data: 5, 10, 20, 20, 25, 29 and 31

Solution:

X_i	X_i^2
5	25
10	100
20	400
20	400
25	625
29	841
31	961
$\sum X_i = 140$	$\sum X_i^2 = 3352$

$$\bar{X} = \frac{\sum X_i}{n} = \frac{140}{7} = 20$$

$$S^2 = \frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2 = \frac{3352}{7} - \left(\frac{140}{7}\right)^2 = 78.857$$

SD: $S = \sqrt{78.857} = 8.88$

Coefficient of variation: $CV = \frac{S}{\bar{X}} \times 100 = \frac{8.88}{20} \times 100 = 44.4\%$

Example 4.10: Calculate AM, variance, SD using direct and coding method and CV. 10, 20, 30, 40 and 50

Solution:

X_i	X_i^2	u_i	u_i^2
10	100	-1	1
20	400	0	0
30	900	1	1
40	1600	2	4
50	2500	3	9
$\sum X_i = 150$	$\sum X_i^2 = 5500$	$\sum u_i = 5$	$\sum u_i^2 = 15$

Where $u_i = \frac{X_i - 20}{10}$

Direct Method:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{150}{5} = 30$$

$$S^2 = \frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2 = \frac{5500}{5} - \left(\frac{150}{5}\right)^2 = 1100 - 900 = 200$$

Coding method:

$$\bar{X} = A + \frac{\sum u_i}{n} \times h = 20 + \frac{5}{5} \times 10 = 30$$

$$S^2 = \left\{\frac{\sum u_i^2}{n} - \left(\frac{\sum u_i}{n}\right)^2\right\} h^2 = \left\{\frac{15}{5} - \left(\frac{5}{5}\right)^2\right\} (10)^2 = (3-1)100 = 200$$

SD:

$S = \sqrt{200} = 14.14$

Coefficient of variation:

$$CV = \frac{S}{\bar{X}} \times 100 = \frac{14.14}{30} \times 100 = 47.13\%$$

Example 4.11: Calculate AM, variance, standard deviation and CV from the followi data of weights of animals in Kg.

Weights	50 -150	150-250	250-350	350-450	450-550
f_i	35	70	105	70	35

Solution:

Weights	f_i	X_i	$f_i X_i$	$f_i X_i^2$
50-150	35	100	3500	350000
150-250	70	200	14000	2800000
250-350	105	300	31500	9450000
350-450	70	400	28000	11200000
450-550	35	500	17500	8750000
Σ	**315**		**94500**	**32550000**

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{94500}{315} = 300$$

$$S^2=\frac{\sum f_iX_i^2}{\sum f_i}-\left(\frac{\sum f_iX_i}{\sum f_i}\right)^2=\frac{32550000}{315}-\left(\frac{94500}{315}\right)^2$$

$$=103333.33-90000=13333.33Kg^2 \qquad S=\sqrt{13333.33Kg^2}=115.47Kg$$

Short cut method: $D = X - A = X - 300$

Weights	f_i	X_i	D_i	f_iD_i	$f_iD_i^2$
50-150	35	100	-200	-7000	1400000
150-250	70	200	-100	-7000	700000
250-350	105	300	0	0	0
350-450	70	400	100	7000	700000
450-550	35	500	200	7000	1400000
Σ	**315**			**0**	**4200000**

$$\bar{X}=A+\frac{\sum f_iD_i}{\sum f_i}=300+\frac{0}{315}=300$$

$$S^2=\frac{\sum f_iD_i^2}{\sum f_i}-\left(\frac{\sum f_iD_i}{\sum f_i}\right)^2=\frac{4200000}{315}-\left(\frac{0}{315}\right)^2$$

$$S^2=13333.33Kg^2;\ S=\sqrt{13333.33Kg^2}=115.47Kg$$

Coding method: $u=\frac{X-A}{h}=\frac{X-200}{100}$

Weights	f_i	X_i	u_i	f_iu_i	$f_iu_i^2$
50-150	35	100	-1	-35	35
150-250	70	200	0	0	0
250-350	105	300	1	105	105
350-450	70	400	2	140	280
450-550	35	500	3	105	315
Σ	**315**			**315**	**735**

$$\bar{X}=A+\frac{\sum f_iu_i}{\sum f_i}\times h=200+\frac{315}{315}\times100=300$$

$$S^2=\left\{\frac{\sum f_iu_i^2}{\sum f_i}-\left(\frac{\sum f_iu_i}{\sum f_i}\right)^2\right\}h^2$$

$$=\left\{\frac{735}{315}-\left(\frac{315}{315}\right)^2\right\}(100)^2=13333.33Kg^2$$

$$S=\sqrt{13333.33Kg^2}=115.47Kg$$

Coefficient of variation:

$$CV=\frac{S}{\bar{X}}\times100=\frac{115.47}{300}\times100=38.49\%$$

4.4 Moments

Moments are defined as the AM of various powers of the deviations of the observations from any value. These are denoted by μ_1,μ_2,μ_3,μ_4 etc. for population data and m_1,m_2,m_3,m_4 etc. for sample data. The basic purpose of moments is to observe the symmetry, skewness and kurtosis of a frequency distribution.

4.4.1 Moments about mean (central moments):

The r^{th} sample moment about mean is denoted by m_r is given as

$$m_r=\frac{\sum(X_i-\bar{X})^r}{n},$$ For ungrouped data.

$$m_r=\frac{\sum f_i(X_i-\bar{X})^r}{\sum f_i},$$ For grouped data.

$r=1,2,3,4,....$

These moments are also called the mean moments.

Thus $m_1' = 1, m_1 = 0, m_2 = S^2$ (variance)

4.4.2 Moments about any value "a" (Non central moments):

The r^{th} sample moment about arbitrary value "a" is denoted by m_r' is given as

$$m_r' = \frac{\sum_i (X_i - a)^r}{n} = \frac{\sum_i D_i^r}{n}, \quad \text{For ungrouped data}$$

$$m_r' = \frac{\sum_i f_i (X_i - a)^r}{\sum_i f_i} = \frac{\sum_i f_i D_i^r}{\sum_i f_i}, \quad \text{For grouped data}$$

$r = 1, 2, 3, 4, \ldots$

These moments are also called the raw moments, thus

$m_0' = 1, m_1' = \bar{X} - a \Rightarrow \bar{X} = a + m_1'$

4.4.3 Moments about origin or zero:

The r^{th} sample moment about origin or zero is denoted by m_r' is given as

$$m_r' = \frac{\sum X_i^r}{n}, \quad \text{For ungrouped data}$$

$$m_r' = \frac{\sum f_i X_i^r}{\sum f_i}, \quad \text{For grouped data;} \quad r = 1, 2, 3, 4, \ldots$$

4.4.4 Relations between moments

The following relations exist between moments about mean (central) and moments about any arbitrary value "a" / "zero" (non-central).

$m_1 = 0$

$m_2 = m_2' - m_1'^2$

$m_3 = m_3' - 3m_2'm_1' + 2m_1'^3$

$m_4 = m_4' - 4m_3'm_1' + 6m_2'm_1'^2 - 3m_1'^4$

4.4.5 Moment ratios

The ratios $(b_1 \text{ and } b_2)$ are called moment ratios when both numerator and denominator are moments. They are defined as $b_1 = \frac{m_3^2}{m_2^3}$, and $b_2 = \frac{m_4}{m_2^2}$

Example 4.12: From the given data X_i = 5, 6, 9, 8, 4, 10
(i) Calculate first four moments about origin and convert them into moments about mean (ii) Calculate first four moments X = 6 and convert them into moments about mean (iii) Find b_1 and b_2

Solution:

(i) Moments about origin

X_i	X_i^2	X_i^3	X_i^4
5	25	125	625
6	36	216	1296
9	81	729	6561
8	64	512	4096
4	16	64	256
10	100	1000	10000
42	**322**	**2646**	**22834**

$$m_r' = \frac{\sum X_i^r}{n}$$

$$m_1' = \frac{\sum X_i}{n} = \frac{42}{6} = 7; \quad m_2' = \frac{\sum X_i^2}{n} = \frac{322}{6} = 53.6667$$

$$m_3' = \frac{\sum X_i^3}{n} = \frac{2646}{6} = 441; \quad m_4' = \frac{\sum X_i^4}{n} = \frac{22834}{6} = 3805.667$$

Converting into moments about mean

$m_1 = 0$ and $m_2 = m_2' - m_1'^2 = 53.6667 - (7)^2 = 4.6667$

$m_3 = m_3' - 3m_2'm_1' + 2m_1'^3 = 441 - 3(53.6667)(7) + 2(7)^3 = 0$

$m_4 = m_4' - 4m_3'm_1' + 6m_2'm_1'^2 - 3m_1'^4 = 3805.667 - 4(441)7 + 6(53.6667)(7)^2 - 3(7)^4 = 32.667$

(ii) Moments about X = 6

X	X − 6	(X − 6)²	(X − 6)³	(X − 6)⁴
5	-1	1	-1	1
6	0	0	0	0
9	3	9	27	81
8	2	4	8	16
4	-2	4	-8	16
10	4	16	64	256
42	**6**	**34**	**90**	**370**

$$m_1' = \frac{\sum (X_i - 6)}{n}$$

$$m_1' = \frac{\sum (X_i - 6)}{n} = \frac{6}{6} = 1$$

$$m_2' = \frac{\sum (X_i - 6)^2}{n} = \frac{34}{6} = 5.6667$$

$$m_3' = \frac{\sum (X_i - 6)^3}{n} = \frac{90}{6} = 15$$

$$m_4' = \frac{\sum (X_i - 6)^4}{n} = \frac{370}{6} = 61.6667$$

Converting into moments about mean

$m_1 = 0$

$$m_2 = m_2' - m_1'^2 = 5.6667 - (1)^2 = 4.6667$$

$$m_3 = m_3' - 3m_2'm_1' + 2m_1'^3 = 15 - 3(5.6667)(1) + 2(1)^3 = 0$$

$$m_4 = m_4' - 4m_3'm_1' + 6m_2'm_1'^2 - 3m_1'^4 = 61.6667 - 4(15)(1) + 6(5.6667)(1)^2 - 3(1)^4 = 32.667$$

(iii) b_1 and b_2:

$$b_1 = \frac{m_3^2}{m_2^3} = \frac{0^2}{4.6667^3} = 0$$

$$b_2 = \frac{m_4}{m_2^2} = \frac{32.667}{4.6667^2} = 1.49$$

Example 4.13: The following distribution gives the weights of 45 cotton bales. Find first four moments about (i) mean and (ii) 40

Weights(kg)	35-37	37-39	39-41	41-43	43-45	45-47
Frequency	2	3	9	10	11	10

Solution: (i) moments about mean

Weights (kg)	f_i	X_i	f_iX_i
35-37	2	36	72
37-39	3	38	114
39-41	9	40	360
41-43	10	42	420
43-45	11	44	484
45-47	10	46	460
Σ	**45**		**1910**

$$\bar{X} = \frac{\sum f_iX_i}{\sum f_i} = \frac{1910}{45} = 42.4444$$

$f_i(X_i - \bar{X})$	$f_i(X_i - \bar{X})^2$	$f_i(X_i - \bar{X})^3$	$f_i(X_i - \bar{X})^4$
-12.8889	83.0605	-535.2756	3449.5302
-13.3332	59.2581	-263.3666	1170.5064
-21.9996	53.7758	-131.4496	321.3155
-4.4444	1.9749	-0.8776	0.3900
17.1116	26.6188	41.4082	64.4146
35.556	126.4229	449.5093	1598.2753
0	**351.111**	**-440.052**	**6604.432**

$m_1 = 0$;

$$m_2 = \frac{\sum f_i(X_i - \bar{X})^2}{\sum f_i} = \frac{351.111}{45} = 7.802$$

$$m_3 = \frac{\sum f_i(X_i - \bar{X})^3}{\sum f_i} = \frac{-440.052}{45} = -9.778$$

$$m_4 = \frac{\sum f_i(X_i - \bar{X})^4}{\sum f_i} = \frac{6604.432}{45} = 146.76$$

(ii) Moments about 40: $D_i = X_i - 40$

Weights (kg)	f_i	X_i	D_i	fD	fD^2	fD^3	fD^4
35-37	2	36	-4	-8	32	-128	512
37-39	3	38	-2	-6	12	-24	48
39-41	9	40	0	0	0	0	0
41-43	10	42	2	20	40	80	160
43-45	11	44	4	44	176	704	2816
45-47	10	46	6	60	360	2160	12960
Σ	**45**			**110**	**620**	**2792**	**16496**

$$m_1' = \frac{\sum f_i D_i}{\sum f_i} = \frac{110}{45} = 2.4444; \qquad m_2' = \frac{\sum f_i D_i^2}{\sum f_i} = \frac{620}{45} = 13.7778$$

$$m_3' = \frac{\sum f_i D_i^3}{\sum f_i} = \frac{2792}{45} = 62.0444; \qquad m_4' = \frac{\sum f_i D_i^4}{\sum f_i} = \frac{16496}{45} = 366.5778$$

Example 4.14: Calculate first four moments about origin and convert them into central moments.

Classes	1-5	6-10	11-15	16-20	21-25
Frequency	7	11	6	4	1

Solution:

Classes	f_i	X_i	fX	fX^2	fX^3	fX^4
1-5	7	3	21	63	189	567
6-10	11	8	88	704	5632	45056
11-15	6	13	78	1014	13182	171366
16-20	4	18	72	1296	23328	419904
21-25	1	23	23	529	12167	279841
	29		**282**	**3606**	**54498**	**916734**

Moments about origin:

$$m_r' = \frac{\sum fX^r}{\sum f}$$

$$m_1' = \frac{\sum fX}{\sum f} = \frac{282}{29} = 9.724; \qquad m_2' = \frac{\sum fX^2}{\sum f} = \frac{3606}{29} = 124.345$$

$$m_3' = \frac{\sum fX^3}{\sum f} = \frac{54498}{29} = 1879.241 \qquad m_4' = \frac{\sum fX^4}{\sum f} = \frac{916734}{29} = 31611.517$$

Central moments

$m_1 = 0$

$m_2 = m_2' - m_1'^2 = 124.345 - (9.724)^2 = 29.79$

$m_3 = m_3' - 3m_2'm_1' + 2m_1'^3 = 1879.241 - 3(124.345)(9.724) + 2(9.724)^3 = 90.78$

$m_4 = m_4' - 4m_3'm_1' + 6m_2'm_1'^2 - 3m_1'^4$

$= 31611.517 - 4(1879.241)(9.724) + 6(124.345)(9.724)^2 - 3(9.724)^4 = 2239.47$

4.5 Symmetrical distribution

A distribution is said to be symmetrical if the left tail and right tail of the frequency curve are equal. In a symmetrical distribution mean, median and mode are identical in other words its smaller values and larger values are equal in numbers.

4.6 Skewness

A lack of symmetry in a frequency distribution is called skewness.

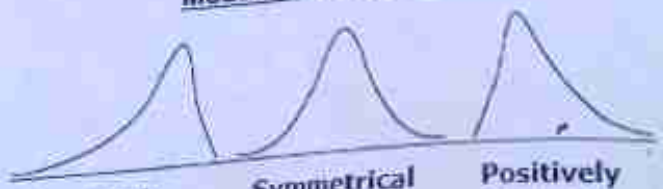
4.6.1 Positively skewed distribution

A distribution is called positively skewed distribution if right tail of the frequency curve is longer than left tail. In this distribution mean > median > mode.

4.6.2 Negatively skewed distribution

A distribution is called negatively skewed distribution if right tail of the frequency curve is smaller than left tail. In this distribution mean < median < mode.

Graphically skewness may be presented as follows.



Negatively Skewed	Symmetrical	Positively Skewed
$\bar{X} < \hat{X} < \tilde{X}$	$\bar{X} = \tilde{X} = \hat{X}$	$\bar{X} > \tilde{X} > \hat{X}$

4.6.3 Coefficient of skewness

(i) Karl Pearson coefficient of skewness:

$S_k = \frac{\bar{X} - \hat{X}}{SD}$ and $S_k = \frac{3(\bar{X} - \tilde{X})}{SD}$, $\quad -3 \le S_k \le 3$

(ii) Bowley's coefficient of skewness:

$S_k = \frac{Q_3 + Q_1 - 2\,median}{Q_3 - Q_1}$ $\quad -1 \le S_k \le 1$

(iii) Moments method:

$S_k = \sqrt{b_1} = \frac{m_3}{\sqrt{m_2^3}} = \frac{m_3}{S^3}$ $\quad -2 \le S_k \le 2$

If $S_k = 0$ then distribution is symmetrical.

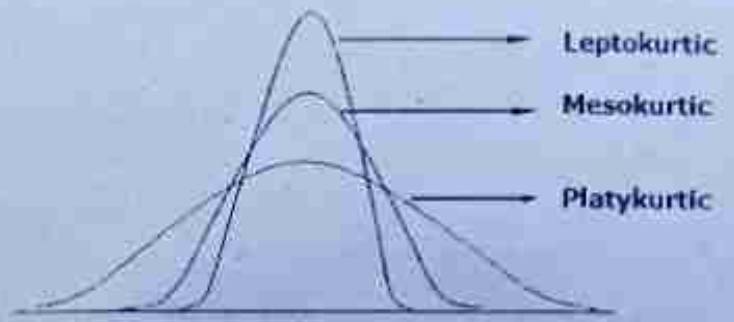
If $S_k < 0$ then distribution is negatively skewed.

If $S_k > 0$ then distribution is positively skewed.

4.7 Kurtosis

Kurtosis is the degree of peakedness of a distribution usually relative to a norm distribution. A distribution having a relatively high peak is called leptokurtic distribution which is flat-topped is called platykurtic. A normal distribution which neither very peaked nor very flat-topped is also called mesokurtic.

Following figure illustrate these types of kurtosis.



$\beta_2 = \frac{\mu_4}{\mu_2^2}$ (Population) and $b_2 = \frac{m_4}{m_2^2}$ (Sample)

If $\beta_2 = 3$ then distribution is mesokurtic.

If $\beta_2 > 3$ then distribution is leptokurtic.

If $\beta_2 < 3$ then distribution is platykurtic.

Example 4.15: Find skewness for the following data.

(i) $\bar{X} = 31.59, \tilde{X} = 33.17, S = 10.98$

(ii) $Q_1 = 5.06, \tilde{X} = 9.15, Q_3 = 16.57$

(iii) $m_2 = 39.91, m_3 = 120.75$

Solution:

(i) $S_k = \frac{3(\bar{X} - \tilde{X})}{S} = \frac{3(31.59 - 33.17)}{10.98} = -0.432$

Data is negatively skewed

(ii) $S_k = \frac{Q_3 + Q_1 - 2\tilde{X}}{Q_3 - Q_1} = \frac{16.57 + 5.06 - 2 \times 9.15}{16.57 - 5.06} = 0.289$

Data is positively skewed

(iii) $S_k = \sqrt{\beta_2} = \frac{m_4}{\sqrt{m_2^4}} = \frac{120.75}{\sqrt{30.91^4}} = 0.4789$ Data is positively skewed

Example 4.16: Find kurtosis for the following data.

$$m_2 = 6.53; m_3 = -10.91; m_4 = 478.29$$

Solution:

$b_2 = \frac{m_4}{m_2^2} = \frac{478.29}{(6.53)^2} = 11.22$; data is Lepto-kurtic

Grouping error: The difference between numerical values calculated from the original data and the data in the form of a frequency distribution, is called the grouping error.

Sheppard's Correction: Grouping error arises in the calculation of moments from a frequency distribution by the assumption that all values are equal to the mid-point of each class, W. F. Sheppard introduced Sheppard's correction to reduce grouping error which are given as follows:

$$m_2(corrected) = m_2(uncorrected) - \frac{h^2}{12}$$

$$m_3(corrected) = m_3(uncorrected)$$

$$m_4(corrected) = m_4(uncorrected) - \frac{h^2}{2} m_2(uncorrected) + \frac{7h^4}{240}$$

Sheppard's correction is applied only when

(i) Frequency distribution is continuous

(ii) Class interval is uniform

Multiple Choice Questions

1. Variance of a variable is always: (a) > 1 (b) < 1 (c) > 0 (d) < 0

2. Variance of a constant is
(a) constant (b) 1 (c) 0 (d) not possible

3. S.D is calculated from H.M:
(a) Always (b) Never (c) Often (d) None of these

4. Range based on:
(a) Upper and lower quartiles (b) Squared deviations
(c) Minimum and Maximum observations (d) Absolute deviations

5. AM of the squared deviations of values from their mean:
(a) Mean deviation (b) Standard deviation
(c) Variance (d) Quartile deviation

6. If S.D (X) = 5, then S.D (2X+5) is: (a) 5 (b) 10 (c) 15 (d) 20

7. A measure of dispersion is always:
(a) Positive (b) Zero (c) Small (d) One

8: If X and Y are independent, then S.D(X − Y) =
(a) S.D (X) − S.D(Y) (b) S.D(X) + S.D (Y)
(c) $\sqrt{\text{Var}(X) + \text{Var}(Y)}$ (d) $\sqrt{\text{Var}(X) - \text{Var}(Y)}$

9. For independent variables X and Y, if S.D(X) = 8, S.D(Y) = 6, then S.D(X-Y)
(a) 10 (b) 2 (c) 14 (d) 28

10. If mode is less than the mean, distribution is:
(a) Symmetrical (b) Normal (c) Positively skewed (d) negatively skewed

11. Right tail is longer than the left tail, distribution is:
(a) Negatively skewed (b) Positively skewed (c) Symmetrical (d) None

12. If mean = 40, mode= 42, distribution is:
(a) Negatively skewed (b) Positively skewed (c) Symmetrical (d) None

13. If Y = 5X + 10, then mean deviation of Y is:
(a) $MD(X)$ (b) $5MD(X)$ c) $5MD(X)+10$ d) $MD(X)+10$

14. Variance (or S.D) remains unchanged by change of:
(a) Origin (b) Scale (c) Both (a) and (b) (d) unit

15. The lowest value of variance is: (a) 1 (b) 0 (c) -2 (d) -1

16. A normal distribution has 68.26% of the observations:
(a) $\bar{X} \pm S$ (b) $\bar{X} \pm 2S$ (c) $\bar{X} \pm 3S$ (d) $\bar{X} \pm 4S$

17. Distribution is symmetrical, if $\sqrt{\beta_1}$:
(a) Negative (b) Positive (c) Zero (d) 3

18. For normal (mesokurtic) distribution, between $\bar{x}$-2S and $\bar{x}$+2S:
(a) 95.44% (b) 50% (c) 68% (d) 99.73%

19. Observations lying within limits $(\bar{X} \pm 3S)$ in the normal distribution is:
(a) 68.27% (b) 95.44% (c) 70% (d) 99.73%

20. To compare the variations of two or more than two series:
(a) Mean (b) Standard deviation (c) Variance (d) CV

21. Quartile deviation: (a) $\frac{2}{3}\sigma$ (b) $\frac{4}{5}\sigma$ (c) $\frac{5}{6}\sigma$ (d) $\frac{6}{5}\sigma$

22. Mean deviation: (a) $\frac{2}{3}\sigma$ (b) $\frac{4}{5}\sigma$ (c) $\frac{5}{6}\sigma$ (d) $\frac{6}{5}\sigma$

23. Bowely's co-efficient of skewness lies between:
(a) 0 and 1 (b) -1 and 0 (c) -1 and +1 (d) -∞ to ∞

24. If co-efficient of skewness is -0.58, distribution is:
(a) Positively skewed (b) Symmetrical
(c) Negatively skewed (d) asymmetrical

25. The types of dispersion are: (a) 2 (b) 3 (c) 4 (d) 5

26. The square root of second central moment:
(a) Variance (b) Standard deviation
(c) Quartile deviation (d) Mean deviation

27. For positively skewed distribution:
(a) Mean < Median < Mode (b) Mean < Mode < Median
(c) Mean > Median > Mode (d) Mean = Median = Mode

28. For negatively skewed distribution; Mean.....Median.....Mode:
(a) = (b) < (c) > (d) >,<

29. In grouped data, the range is the difference between:
(a) Two extremes class frequencies (b) Two extremes mid points
(c) Two extremes class boundaries (d) Both b & c

30. Which is poor measure of dispersion in open-end distribution:
a) Range b) Quartile deviation c) Semi-inter quartile range d) AM

31. The range of constant "A": a) Zero b) A c) 1 d) A^2

32. For relative dispersion, unit of measurement:
a) Changed b) Vanishes c) Does not vanish d) unit

33. The range of a series of -2, -3, -5 and -10 is:
a) -12 b) 8 c) -8 d) 12

34. The variance of 5, 5, 5, 5 and 5 is: a) 5 b) 25 c) 125 d) 0

35. If A.M=25 and $S^2 = 25$, then co-efficient of variation (C.V) is:
a) 100% b) 25% c) 20% d) 1%

36. Mean deviation is always:
a) Less than S.D b) Equal than S.D c) More than S.D d) Negative

37. In symmetrical distribution, the co-efficient of skewness is:
a) -1 b) +1 c) 0 d) 0.5

38. First moment about mean is always:
a) One b) Zero c) mean d) SD

39. First moment about origin is equal to:
a) One b) Zero c) mean d) SD

40. In a skewed distribution mean, median and mode are always:
a) Identical b) Different c) Zero d) Same

41. Mean deviation is associated with:
a) A.M b) HM c) QD d) GM

42. Third moment about mean (m_3) is zero, distribution is:
a) Positively skewed b) negatively skewed c) Symmetrical d) asymmet

43. Sum of absolute deviations of values are least if measured from:
a) Mean b) Mode c) Median d) GM

44. Sum of squares of deviations is least from:
a) Mean b) Median c) Mode d) HM

45. The second moment about mean is:
a) Variance b) Mean c) SD d) Zero

46. The variance (S.D) of constant is:
a) Constant b) Unity c) Zero d) ∞

47. Standard deviation of 2, 4, 6, 8 and 10 is 2.83, then standard deviation of 102, 104, 106, 108 and 110 is:
a) 283 b) 102.83 c) 2.83 d) 28.3

48. Standard deviation changes by the change of:
a) Origin b) Scale c) Algebraic d) Both (a) and (b)

49. Y = X±3, then range of Y is:
a) 3 b) Range(X+3) c) Range (X) d) Range (X-3)

50. If Y=3X±5, then S.D of Y is:
a) 9SD(X) b) 3S.D(X) c) 3S.D(X)+5 d) 3S.D(X) ±5

51. If $b_2(\beta_2) = 3$, distribution is:
a) Leptokurtic b) Platykurtic c) Mesokurtic d) Skewed

52. Var(2X±3) is:
a) 4Var (X) b) 2Var (X) c) 4Var (X)+3 d) 2Var (X)+3

53. Variance (standard deviation) is calculated from:
a) Mean b) Median c) Mode d) GM

54. If X and Y are independent than Var (X-Y) is equal to:
a) Var(X) - Var(Y) b) Var(X) + Var(Y) c) $\sqrt{var(X)+var(Y)}$ d) $\sqrt{var(X)-var(Y)}$

55. b_2 measures:
a) Symmetry b) Dispersion c) kurtosis d) Skewness

56. The standard deviation of 3, 3, 3,..... 3 is:
a) 3 b) 8 c) Zero d) 16

57. Symmetrical distribution is:
a) t-shaped b) J-shaped c) Bell-shaped d) long tailed

58. A.M=136.75, Median=148.37 and Mode=152.80:
a) Positive skewed b) Negatively skewed
c) Symmetrical d) asymmetrical

59. Addition of extreme value in a data set affects more
a) Q.D b) Mode c) Median d) Variance

60. For nominal data we use a) Mean b) median c) mode d) GM

61. For ordinal data set we use
a) Mean b) median c) mode d) GM

62. For a data set $n = 20$, $\bar{X} = 20$ and $S^2 = 16$ then $\sum(X - \bar{X}) =$
a) 1 b) 320 c) 2000 d) 0

63. Z score for X = 25 is 1.9 find SD if mean of X is 18
a) 7 b) 3.68 c) 1.39 d) 0

key

Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans
1	c	2	c	3	b	4	c	5	c	6	b
7	a	8	c	9	a	10	c	11	b	12	a
13	b	14	a	15	b	16	a	17	c	18	a
19	d	20	d	21	a	22	b	23	c	24	c
25	a	26	b	27	c	28	b	29	d	30	a
31	a	32	b	33	b	34	d	35	c	36	a
37	c	38	b	39	c	40	b	41	a	42	c
43	c	44	a	45	a	46	c	47	c	48	b
49	c	50	b	51	c	52	a	53	a	54	b
55	c	56	c	57	c	58	b	59	d	60	c
61	b	62	d	63	b						

Exercise

Q No. 4.1: (a) Define measure of dispersion and its types.
(b) Write different methods for measuring the dispersion.
(c) Define the Range, Quartile deviation, mean deviation, variance, standard deviation, moments, skewness and kurtosis.
(d) Describe the properties of variance and standard deviation.

Q No. 4.2 (a): Find variance and Standard deviation from the following calculations.

(i) $n=65, \sum (X-\bar{X})^2 = 842.4$

(ii) $n=80, \sum X = 6824, \sum X^2 = 584595$

(iii) $\sum f = 200, \sum fu = 9, \sum fu^2 = 150$; and $u = \dfrac{X-130}{20}$

(iv) $n=10, \sum D = 52, \sum D^2 = 952$, and $A = 570$

(v) $n=10, \sum D = -128, \sum D^2 = 3472$; and $D = X - 110$

(b) (i) Find CV if S.D. = 10 and $\bar{X} = 25$.
(ii) Find CV if variance = 16 and $\bar{X} = 50$.
(iii) Find SD if CV = 30% and $\bar{X} = 15$.
(iv) Find variance if CV = 40% and $\bar{X} = 25$.
(v) Find mean if CV = 25% and $SD = 15$.
(vi) Find mean if CV = 5% and variance = 9.

Q No. 4.3: Find Variance, SD and CV.

Weight (kg)	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59
f	3	17	36	58	27	6

Q No. 4.4: Find QD, SD and MD about median.

Income (000) in Ps.	0 – 4.9	5 – 9.9	10 – 14.9	15 – 19.9	20 – 24.9
No of employeer	8	28	60	30	5

Q No. 4.5: Find Moments about mean by calculating moments about zero.

Classes	0 – 9.9	10 – 19.9	20 – 29.9	30 – 39.9	40 – 49.9
f	5	7	15	12	6

Q No.4.6: The marks X scored by a sample of 56 students in an examination are summarized by $n=56, \sum X = 1026, \sum X^2 = 20889$ Calculate the mean and standard deviation of the marks.

Q No. 4.7: Find variance by (i) Direct (ii) Indirect and (iii) Coding method.

Classes	1 – 10	11 – 20	21 – 30	31 – 40	41 – 50
f	5	10	18	6	2

Also find coefficient of variation.

Q No. 4.8: Find mean deviation about mean, median and mode also find their coefficients.

Marks	1 – 20	21 – 40	41 – 60	61 – 80	81 – 100
f	5	20	50	18	8

Q No. 4.9: Find mean deviation about median.

Classes	1 – 15	16 – 30	31 – 45	46 – 60	61 – 75	76 – 90
f	6	10	25	20	8	2

Q No. 4.10: The mean stress rate of a group of students for a particular course is 6.40 and standard deviation is 2.5. Suppose a student's stress raw score is 12, find his Z score.

Solution

Q No. 4.2:

Sr.	Variance	Standard deviation
(i)	$S^2 = \frac{\sum(X-\bar{X})^2}{n} = \frac{842.5}{65} = 12.96$	$S = 3.60$
(ii)	$S^2 = \frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2 = \frac{584595}{80} - \left(\frac{6824}{80}\right)^2 = 31.35$	$S = 5.60$
(iii)	$S^2 = \left\{\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2\right\}h^2 = \left\{\frac{150}{200} - \left(\frac{9}{200}\right)^2\right\}20^2 = 299.19$	$S = 17.29$
(iv)	$S^2 = \frac{\sum D^2}{n} - \left(\frac{\sum D}{n}\right)^2 = \frac{952}{10} - \left(\frac{52}{10}\right)^2 = 68.16$	$S = 8.26$
(v)	$S^2 = \frac{\sum D^2}{n} - \left(\frac{\sum D}{n}\right)^2 = \frac{3472}{10} - \left(\frac{-128}{10}\right)^2 = 183.36$	$S = 13.54$

(b) $CV = \frac{S}{\bar{X}} \times 100$

Sr.	Computation	Sr.	Computation	Sr.	Computation
(i)	$CV = \frac{10}{25} \times 100 = 40\%$	(ii)	$CV = \frac{4}{50} \times 100 = 8\%$	(iii)	$30 = \frac{S}{15} \times 100 \Rightarrow S = 45$
(iv)	$40 = \frac{S}{25} \times 100$ $S = 10 \Rightarrow S^2 = 100$	(v)	$25 = \frac{15}{\bar{X}} \times 100$ $\bar{X} = 60$	(vi)	$S = \frac{3}{\bar{X}} \times 100$ $\bar{X} = 60$

Q No. 4.3:

Weights(Kg)	f	X	u	fu	fu^2
0 – 9	3	4.5	-2	-6	12
10 – 19	17	14.5	-1	-17	17
20 – 29	36	24.5	0	0	0
30 – 39	58	34.5	1	58	58
40 – 49	27	44.5	2	54	108
50 – 59	6	54.5	3	18	54
Total	**147**			**107**	**249**

$$S^2 = \left\{\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2\right\}h^2 = \left\{\frac{249}{147} - \left(\frac{107}{147}\right)^2\right\}10^2 = 116.41\ Kg^2$$

$$S = 10.79\ Kg$$

$$\bar{X} = A + \frac{\sum fu}{\sum f} \times h = 24.5 + \frac{107}{147} \times 10 = 31.78\ Kg$$

$$CV = \frac{10.79}{31.78} \times 100 = 33.95\%$$

Q No. 4.4:

Income (000) (Rs)	f	C.B.	Cf	X	u	fu	fu^2	$f\lvert X-\bar{X}\rvert$
0 – 4.9	8	-0.05 –	8	2.45	-2	-16	32	79.68
5 – 9.9	28	4.95 –	36	7.45	-1	-28	28	138.88
10 – 14.9	60	9.95 –	96	12.45	0	0	0	2.4
15 – 19.9	30	14.95 –	126	17.45	1	30	30	151.2
20 – 24.9	5	19.95 –	131	22.45	2	10	20	50.2
Sum(Σ)	**131**					**-4**	**110**	**422.36**

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right); \frac{n}{4} = \frac{131}{4} = 32.75 \Rightarrow Q_1 = 4.95 + \frac{5}{28}(32.75 - 8) = 9.37\ (000)Rs$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right); \frac{3n}{4} = \frac{393}{4} = 98.25 \Rightarrow Q_3 = 14.95 + \frac{5}{30}(98.25 - 96)$$

$$= 15.33\ (000)Rs$$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{15.33 - 9.37}{2} = 2.98(000)Rs$$

$$S^2 = \left\{\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2\right\}h^2 = \left\{\frac{110}{131} - \left(\frac{-4}{131}\right)^2\right\}5^2 = 20.97(000)\ Rs^2$$

$$S = 4.58\ (000)Rs$$

$$\bar{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right); \frac{n}{2} = \frac{131}{2} = 65.5 \Rightarrow \bar{X} = 9.95 + \frac{5}{60}(65.5 - 36) = 12.41\ (000)Rs$$

$$MD = \frac{\sum f|X - \bar{X}|}{\sum f} = \frac{422.36}{131} = 3.22\ (000)Rs$$

Q No. 4.5:

Classes	f	X	fX	fX^2	fX^3	fX^4
0 – 9.9	5	4.95	24.75	122.51	606.44	3001.86
10 – 19.9	7	14.95	104.65	1564.52	23389.54	349673.57
20 – 29.9	15	24.95	374.25	9337.54	232971.56	5812640.44
30 – 39.9	12	34.95	419.4	14658.03	512298.15	17904820.29
40 – 49.9	6	44.95	269.7	12123.02	544929.52	24494582.12
ΣfX^r	**45**		**1192.75**	**37805.61**	**1314195.21**	**48564718.28**
m'_r			26.50	840.12	29204.34	1079215.96

Moments about mean:

$$m_2 = m'_2 - (m'_1)^2 = 840.12 - (26.50)^2 = 137.87$$

$$m_3 = m'_3 - 3m'_2m'_1 + 2(m'_1)^3 = 29204.34 - 3(840.12)(26.5) + 2(26.5)^3 = -365.95$$

$$m_4 = m'_4 - 4m'_3m'_1 + 6m'_2(m'_1)^2 - 3(m'_1)^4$$

$$= 1079215.96 - 4(29204.34)(26.5) + 6(840.12)(26.5)^2 - 3(26.5)^4 = 439363525$$

Q No. 4.6:

$$\bar{X} = \frac{\Sigma X}{n} = \frac{1026}{56} = 18.32\ marks$$

$$S = \sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2} = \sqrt{\frac{20889}{56} - \left(\frac{1026}{56}\right)^2} = 6.11\ marks$$

Q No. 4.7:

Classes	f	X	fX	fX^2	D	fD	fD^2	u	fu	fu^2
1 – 10	5	5.5	27.5	151.25	-20	-100	2000	-2	-10	20
11 – 20	10	15.5	155	2402.5	-10	-100	1000	-1	-10	10
21 – 30	18	25.5	459	11704.5	0	0	0	0	0	0
31 – 40	6	35.5	213	7561.5	10	60	600	1	6	6
41 – 50	2	45.5	91	4140.5	20	40	800	2	4	8
Total	**41**		**945.5**	**25960.25**		**-100**	**4400**		**-10**	**44**

$$S^2 = \frac{\Sigma fX^2}{\Sigma f} - \left(\frac{\Sigma fX}{\Sigma f}\right)^2 = \frac{25960.25}{41} - \left(\frac{945.5}{41}\right)^2 = 101.37$$

$$S^2 = \frac{\Sigma fD^2}{\Sigma f} - \left(\frac{\Sigma fD}{\Sigma f}\right)^2 = \frac{4400}{41} - \left(\frac{-100}{41}\right)^2 = 101.37$$

$$S^2 = \left\{\frac{\Sigma fu^2}{\Sigma f} - \left(\frac{\Sigma fu}{\Sigma f}\right)^2\right\}h^2 = \left\{\frac{44}{41} - \left(\frac{-10}{41}\right)^2\right\}10^2 = 101.37$$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{945.5}{41} = 23.06$$

$$S = 10.07;\ CV = \frac{S}{\bar{X}} \times 100 = \frac{10.07}{23.06} \times 100 = 43.67\%$$

Q No. 4.8:

Marks	f	X	u	fu	CB	cf	$f\lvert X - \bar{X}\rvert$	$f\lvert X - \tilde{X}\rvert$	$f\lvert X - \hat{X}\rvert$
1 – 20	5	10.5	-2	-10	0.5-	5	204	201	198.5
21 – 40	20	30.5	-1	-20	20.5-	25	416	404	394
41 – 60	50	50.5	0	0	40.5-	75	40	10	15
61 – 80	18	70.5	1	18	60.5-	93	345.6	356.4	365.4
81 – 100	8	90.5	2	16	80.5-	101	313.6	318.4	322.4
Total	**101**			**4**			**1319.2**	**1289.8**	**1295.3**

Mean deviation about mean: $\bar{X} = A + \frac{\Sigma fu}{\Sigma f} \times h = 50.5 + \frac{4}{101} \times 20 = 51.3\ marks$

$$MD = \frac{\Sigma f|X - \bar{X}|}{\Sigma f} = \frac{1319.2}{101} = 13.06\ marks$$

Mean deviation about median: $\tilde{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right) = 40.5 + \frac{20}{50}(50.5 - 25) =$ 50.7 marks

$$MD = \frac{\Sigma f|X - \tilde{X}|}{\Sigma f} = \frac{1289.8}{101} = 12.77\ marks$$

Mean deviation about mean:

$$\hat{X} = l + \frac{f_m - f_1}{f_m - f_1 + f_m - f_2} \times h = 40.5 + \frac{50 - 20}{50 - 20 + 50 - 18} \times 20 = 50.2\ marks$$

$$MD = \frac{\Sigma f|X - \hat{X}|}{\Sigma f} = \frac{1295.3}{101} = 12.82\ marks$$

Q No. 4.9:

Marks	f	X	CB	cf	$f\lvert X-\bar{X}\rvert$
1 – 15	6	8	0.5-	6	205.2
16 – 30	10	23	15.5-	16	192
31 – 45	25	38	30.5-	41	105
46 – 60	20	53	45.5-	61	216
61 – 75	8	68	60.5-	69	206.4
76 – 90	2	83	75.5-	71	81.6
Total	**71**				**1006.2**

Mean deviation about median: $\bar{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right) = 30.5 + \frac{15}{25}(35.5 - 16) = 42.2$

$$MD = \frac{\sum f|X - \bar{X}|}{\sum f} = \frac{1006.2}{71} = 14.17$$

Q No. 4.10:

$$Z = \frac{X - \mu}{\sigma} = \frac{12 - 6.4}{2.5} = 2.24$$

Chapter 5

PROBABILITY

Statistics quantify uncertainty

Learning Goals:

(i) To measure the uncertainty of an event.

5.1 Some important terms:

5.1.1 Factorial

Multiplication of first "n" natural number is termed as "n factorial" denoted by n! or $\lfloor n$

Example 5.1: factorial of 6 denoted by 6! is $6! = 6\times5\times4\times3\times2\times1 = 720$

5.1.2 Combination

All possible selection of "r" objects out of "n" distinct objects ignoring their order is called combination, it is denoted by ${}^{n}C_{r}$ and is given as ${}^{n}C_{r} = \frac{n!}{r!(n-r)!}$, where n ≥ r

Other forms of writing combination are ${}_{n}C_{r}$, $\binom{n}{r}$ or $C(n,r)$

Example 5.2: Solve (i) ${}^{15}C_{7}$ (ii) ${}^{10}C_{8}$ (iii) $\binom{16}{9}$ and $C(14,6)$

Solution:

(i) ${}^{15}C_{7} = \frac{15!}{(15-7)!7!} = \frac{15!}{8!7!} = \frac{15.14.13.12.11.10.9.8!}{7.6.5.4.3.2.1.8!} = 6435$

(ii) ${}^{10}C_{8} = \frac{10!}{(10-8)!8!} = \frac{10!}{2!8!} = \frac{10.9.8!}{2.1.8!} = 45$

(iii) $\binom{16}{9} = \frac{16!}{(16-9)!9!} = \frac{16!}{7!9!} = \frac{16.15.14.13.12.11.10.9!}{7.6.5.4.3.2.1.9!} = 11440$

(iv) $C(14,6) = \frac{14!}{(14-6)!6!} = \frac{14!}{8!6!} = \frac{14.13.12.11.10.9.8!}{8!6.5.4.3.2.1} = 3003$

5.1.3 Permutation

All possible distinct arrangements of "r" objects out of "n" distinct objects considering their order is called permutation, denoted by ${}^nP_r$ and is given as ${}^nP_r = \frac{n!}{(n-r)!}$.

Example 5.3: expand (i) ${}^{10}P_8$ and (ii) ${}^{15}P_7$

Solution:

$$(i)\ {}^{10}P_8 = \frac{10!}{(10-8)!} = \frac{10!}{2!} = \frac{10.9.8.7.6.5.4.3.2!}{2!} = 1814400$$

$$(ii)\ {}^{15}P_7 = \frac{15!}{(15-7)!} = \frac{15!}{8!} = \frac{15.14.13.12.11.10.9.8!}{8!} = 32432400$$

5.1.4 Trial

A trial is a single performance of an experiment.

5.2 Random experiment

An experiment which may produce different results for each trial under similar condition is called random experiment.

Tossing a coin, rolling a die and drawing a card from a pack of playing cards, interviewing a person for a specific purpose, recording the amount of rain fall in a city are examples of a random experiment. Experiments conducted in the laboratories of physical and life sciences, observing various characteristics of a plant or an animal are also examples of random experiment.

5.3 Sample space

A set of all possible outcomes that can be generated by the performance of a random experiment is called sample space. Sample space provides a mathematical model of real life situation.

A member of sample space is called sample point.

Example 5.4: Make sample space for

1. Tossing a coin
2. Tossing two coins (a coin two times)
3. Rolling a die

4. A family having a child
5. A family having two children
6. Observing life of an elephant.
7. Conducting an experiment of titration with 1 ml volume of a liquid
8. Counting number of flowers on a plant.
9. Marks of students in a test with maximum marks as 50.

Answer:

1. $S = \{H, T\}$
2. $S = \{HH, HT, TH, TT\}$
3. $S = \{1,2,3,4,5,6\}$
4. $S = \{B, G\}$
5. $S = \{BB, BG, GB, GG\}$

In above example from 6 to 9 we cannot produce sample space in set form. For these certain model/distribution exists such as normal, Poisson or any other distribution. Limits of sample space are

6. $S = \{x | 0 \le x \le 100, x \in R\}$
7. $S = \{x | 1.5 < x < 2.5, x \in R\}$
8. $S = \{x | x \ge 0, x \in R\}$
9. $S = \{x | 0 \le x \le 50, x \in R\}$

5.4 Event

Any sub set of a sample space may called an event.

5.5 Probability

Probability is an area of study which involves predicting the relative likeli-hood of various outcomes. Probability can be described as mathematical theory of uncertainty.

5.6 Different approaches for calculation of probability

5.6.1 Classical approach:

If n(S) are the number of equally likely, mutually exclusive and exhaustive outcomes of a random experiment out of which n(E) outcomes are favorable to the occurrence of an event E, then the probability that event "E" occurs, denoted by $P(E)$, is given by

$$P(E)=\frac{n(E)}{n(S)}.$$

It is a simple form of probability. It is also known as priori definition of probability.

5.6.2 Relative frequency approach

In this approach probability of an event is determined on the basis of experimentation or historical data.

5.6.3 Axiomatic definition of probability

This definition of probability based on certain axioms. Let sample space S has the sample points $A_1, A_2, \ldots, A_n$ and their probabilities are $P(A_1), P(A_2), \ldots, P(A_n)$. The probability of a sample point A_i must satisfy the following properties.

(i) $0 \le P(A_i) \le 1$

(ii) $\sum P(A_i) = 1$

(iii) If A_i and A_j are mutually exclusive events then $P(A_i \cup A_j) = P(A_i) + P(A_j)$

5.6.4 Subjective approach

Subjective approach of probability is simply a guess of a person about outcome of an event. This may be different for each individual therefore it is also known as personalistic approach.

5.7 Probability of an event

Suppose an event E can happen in "r" ways out of a total of n possible equally likely ways. Then the probability of occurrence of the event (also called its likelihood) is given by, $P(E)=\frac{n(E)}{n(S)}=\frac{r}{n}$

Example 5.5: Pakistan and New Zealand cricket teams are going to play a T20 match. Past record shows that they played 21 matches from 2007-2017, Pakistan and New Zealand won 13 and 8 matches respectively. Find probability that Pakistan will win this match, using classical and relative frequency approach.

Solution:

Classical Approach

There are two teams, sample space for winner of this match will be

$S=\{\text{Pakistan, New Zealand}\}$, $n(S) = 2$

Let A = winning team of this match is Pakistan

$A=\{\text{Pakistan}\}$, $n(A) = 1$

$$P(A)=\frac{n(A)}{n(S)}=\frac{1}{2}=0.50$$

Relative frequency approach

According past record

Total matches = 21

Matches won by Pakistan = 13

$$P(A)=\frac{13}{21}=0.619$$

Example 5.6: (a) A coin is tossed. What is probability of (i) head (ii) tail appears? (b) A coin is tossed twice find the probability of (i) one head (ii) two heads (iii) at least one head (iv) at most one head

(c) Three coins are tossed, find the probability of (i) one head (ii) two heads (iii) at least two heads (iv) at most one head

Solution:

(a) $S=\{H,T\}$; $n(S)=2$

(i) Let A = head appears

$A=\{H\}$; $n(A)=1$; $P(A)=\frac{n(A)}{n(S)}=\frac{1}{2}$

(ii) Let B = tail appears $P(B)=\frac{n(B)}{n(S)}=\frac{1}{2}$

$S=\{T\};\ n(B)=1$;

(b) $S=\{HH,HT,TH,TT\};n(S)=4$

(i) Let A = one head

$A=\{HT,TH\};n(A)=2;\quad P(A)=\frac{n(A)}{n(S)}=\frac{2}{4}=\frac{1}{2}$

(ii) Let B = two heads

$B=\{HH\};n(B)=1$; $P(B)=\frac{n(B)}{n(S)}=\frac{1}{4}$

(iii) Let C = at least one head

$C=\{HH,HT,TH\};n(C)=3;\ P(C)=\frac{n(C)}{n(S)}=\frac{3}{4}$

(iv) Let D = at most one head

$D=\{HT,TH,TT\};n(D)=3;\ P(D)=\frac{n(D)}{n(S)}=\frac{3}{4}$

(c) $S=\{HHH,HHT,HTH,HTT,THH,THT,TTH,TTT\};n(S)=8$

(i) Let A = one head

$A=\{HTT,THT,TTH\};n(A)=3;\quad P(A)=\frac{n(A)}{n(S)}=\frac{3}{8}$

(ii) Let B = two heads

$B=\{HHT,HTH,THH\};n(B)=3$; $P(B)=\frac{n(B)}{n(S)}=\frac{3}{8}$

(iii) Let C = at least two heads

$C=\{HHT,HTH,THH,HHH\};n(C)=4;P(C)=\frac{n(C)}{n(S)}=\frac{4}{8}=\frac{1}{2}$

(iv) Let D = at most one head

$D=\{HTT,THT,TTH,TTT\};n(D)=4;\ P(D)=\frac{n(D)}{n(S)}=\frac{4}{8}=\frac{1}{2}$

Example 5.7: (a) A die is rolled. What is probability of (i) 6 appears (ii) even number appears (iii) prime number appears?
(b) A die is rolled twice find the probability of (i) same numbers turn up (ii) sum of dots is 9 (iii) sum of dots is greater than 9 (iv) sum of dots is less than 4.

Solution: (a) $S=\{1,2,3,4,5,6\};\quad n(S)=6$

(i) Let $A=6;\ A=\{6\};n(A)=1;\ P(A)=\frac{n(A)}{n(S)}=\frac{1}{6}$

(ii) Let B = even number

$B=\{2,4,6\};\ n(B)=3;\ P(B)=\frac{n(B)}{n(S)}=\frac{3}{6}=\frac{1}{2}$

(iii) Let C = prime number

(iv) $C=\{2,3,5\};\ n(C)=3;\ P(C)=\frac{n(C)}{n(S)}=\frac{3}{6}=\frac{1}{2}$

(b) $S=\begin{Bmatrix}(1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6)\\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6)\\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6)\\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6)\\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6)\\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6)\end{Bmatrix};\quad n(S)=36$

Let A = same number appears

(i) $A=\{(1,1),\ (2,2),\ (3,3),\ (4,4),\ (5,5),\ (6,6)\};n(A)=6;\ P(A)=\frac{n(A)}{n(S)}=\frac{6}{36}=\frac{1}{6}$

(ii) Let B = sum of dots is 9

$B=\{(3,6),\ (4,5),\ (5,4),\ (6,3)\};\ n(B)=4;\ P(B)=\frac{n(B)}{n(S)}=\frac{4}{36}=\frac{1}{9}$

(iii) Let C = sum of dots is greater than 9

$C=\{(4,6),\ (5,5),\ (6,4),\ (6,5),\ (5,6),\ (6,6)\};\ n(C)=6;$

$$P(C)=\frac{n(C)}{n(S)}=\frac{6}{36}=\frac{1}{6}$$

(iv) D = sum of dots is less than 4

$$D=\{(1,2),\ (2,1),\ (1,1)\};\ n(D)=3;\ P(D)=\frac{n(B)}{n(S)}=\frac{3}{36}=\frac{1}{12}$$

Example 5.8: What is probability that there are two girls in a family having two children?

Solution: $S=\{BB,BG,GB,GG\};\ n(S)=4$

Let E = there are two girls in the family

$$E=\{GG\};\ n(E)=1;\ P(E)=\frac{n(E)}{n(S)}=\frac{1}{4}$$

Distribution of playing cards

Red Cards		Black Cards		Card Name
Diamond ♦	Heart ♥	Spade ♠	Club ♣	
A	A	A	A	Ace
2	2	2	2	Numbered
3	3	3	3	
4	4	4	4	
5	5	5	5	
6	6	6	6	
7	7	7	7	
8	8	8	8	
9	9	9	9	
10	10	10	10	
J	J	J	J	Jack
Q	Q	Q	Q	Queen
K	K	K	K	King

Example 5.9: A card is drawn from a well shuffled pack of playing cards. What is the probability that it is (i) red (ii) black (iii) spade (iv) club (v) diamond(Tiles) (vi) heart (vii) face (viii) pictured (ix) king (x) queen (xi) jack (knive) (xii) numbered card and (xiii) ace card.

Solution: There are 52 cards in a deck of playing cards.

(i) $P(Red)=\frac{n(Red)}{n(Total)}=\frac{26}{52}=\frac{1}{2}$

(ii) $P(Black)=\frac{n(Black)}{n(Total)}=\frac{26}{52}=\frac{1}{2}$

(iii) $P(Spade)=\frac{n(Spade)}{n(Total)}=\frac{13}{52}=\frac{1}{4}$

(iv) $P(Club)=\frac{n(Club)}{n(Total)}=\frac{13}{52}=\frac{1}{4}$

(v) $P(Diamond)=\frac{n(Diamond)}{n(Total)}=\frac{13}{52}=\frac{1}{4}$

(vi) $P(Heart)=\frac{n(Heart)}{n(Total)}=\frac{13}{52}=\frac{1}{4}$

(vii) $P(Face)=\frac{n(Face)}{n(Total)}=\frac{16}{52}=\frac{4}{13}$

Face card are 16 including 12 pictures and 4 ace cards.

(viii) $P(Picture)=\frac{n(Picture)}{n(Total)}=\frac{12}{52}=\frac{3}{13}$

(ix) $P(King)=\frac{n(King)}{n(Total)}=\frac{4}{52}=\frac{1}{13}$

(x) $P(Queen)=\frac{n(Queen)}{n(Total)}=\frac{4}{52}=\frac{1}{13}$

(xi) $P(Jack)=\frac{n(Jack)}{n(Total)}=\frac{4}{52}=\frac{1}{13}$

(xii) $P(Numberal) = \frac{n(Numberal)}{n(Total)} = \frac{36}{52} = \frac{9}{13}$

(xiii) $P(Ace) = \frac{n(Ace)}{n(Total)} = \frac{4}{52} = \frac{1}{13}$

When more than one "n" items are selected/chosen from total "N" items, then number of elements in sample space "S" i.e. $n(S)$ are $n(S) = {}^{N}C_{n}$.

5.7 Types of events

5.7.1 Impossible event: An event whose chance of occurrence is zero is called impossible or null event. For example obtaining a number 0 in rolling a die.

5.7.2 Sure event: An event whose chance of occurrence is one (100%) is called sure or certain event. For example obtaining a number from 1 to 6 in rolling a die..

5.7.3 Independent events: Events are called independent events if the occurrence of one event does not affect the occurrence of other event. For example appearing a head and even number when a coin and a die are tossed simultaneously.

5.7.4 Dependent events: Events are called dependent events if the occurrence of one event affects the occurrence of other event. Mathematically $A \cap B \neq \phi$. For example appearing an even number and a number less than 5 when a die is rolled.

5.7.5 Mutually exclusive events: Events are called mutually exclusive events if they cannot occur together at the same time then. Mathematically if $A \cap B = \phi$, for example appearing an even number and an odd number when a die is rolled.

5.7.6 Not mutually exclusive events: Events are called not mutually exclusive events if they can occur together at the same time. For example appearing an even number and a number multiple of three when a die is rolled.

5.7.7 Equally likely events: If probabilities of occurrence of different events are same then these events are called equally likely events. For example probability of appearing

an even number (0.50) and an odd number (0.50) in rolling a die, therefore these are equally likely events.

5.7.8 Collectively exhaustive events: If there are "k" mutually exclusive events $A_1, A_2, \ldots, A_k$ and if their union is sample space i.e. $\bigcup_{i=1}^{k} A_i = S$; then these are called collectively exhaustive events. For example if a card is drawn from an ordinary deck of playing cards and if $A_1 = card\ of\ diamond$, $A_2 = card\ of\ heart$, $A_3 = card\ of\ spade$ and $A_4 = card\ of\ club$ as they are collectively exclusive but their union is the whole sample space (complete deck of playing cards) then they are collectively exhaustive events.

5.7.9 Simple (elementary) event: An event consisting of only one sample point is called a simple event. For example when a die is rolled let $A = \{6\}$

5.7.10 Compound event: An event consisting of more than one sample point is called a compound event. For example in rolling a die $A = \{2,4,6\}$

5.7.11 Complementary event: An event denoted by $\bar{A}$, A' or A^c is said to be complementary to an event A in S if $\bar{A}$ consists of all those points of sample space which are not in A. for example in rolling a die let $A = \{2,4,6\}$ then $\bar{A} = \{1,3,5\}$

Example 5.10: Table below shows data about gender and education level.

Gender	Un-educated	School Level	College Level	University Level	Total
Male	90	45	30	20	185
Female	160	50	28	12	250
Total	250	95	58	32	435

(i) What is probability of un-educated female?
(ii) What is probability of university level educated male?
(iii) What is probability of college level educated person?
(iv) What is probability of university level or college level educated male?
(v) What is probability of university level educated person or female?
(vi) What is probability of university level educated male and uneducated female?

Solution:

Gender	Un-educated	School Level	College Level	University Level	Total
Male	90	45	30	20	185
Female	160	50	28	12	250
Total	250	95	58	32	435

(i) What is probability of un-educated female?

Solution: Let A = uneducated female

$$P(A)=\frac{n(A)}{n(S)}=\frac{160}{435}=0.368$$

(ii) What is probability of university level educated male?

Solution: Let B = university level educated male.

$$P(B)=\frac{n(B)}{n(S)}=\frac{20}{435}=0.046$$

(iii) What is probability of college level educated person?

Solution: Let C = college level educated person

$$P(C)=\frac{n(C)}{n(S)}=\frac{58}{435}=0.133$$

(iv) What is probability of university level or college level educated male?

Solution: Let D = university level or college level educated male

$$P(D)=\frac{n(D)}{n(S)}=\frac{30+20}{435}=\frac{50}{435}=0.115$$

(v) What is probability of university level educated person or female?

Solution: Let E = university level educated person.

Let F = female

$$P(E)=\frac{n(E)}{n(S)}=\frac{32}{435}=0.074 \quad P(F)=\frac{n(F)}{n(S)}=\frac{250}{435}=0.575$$

$$P(E\cap F)=\frac{n(E\cap F)}{n(S)}=\frac{12}{435}=0.028$$

$$P(E \text{ or } F)=P(E\cup F)=P(E)+P(F)-P(E\cap F)$$

$$=0.074+0.575-0.028=0.621$$

(vi) What is probability of university level educated male and uneducated female?

Solution: 0

5.8 Conditional probability

If A and B are any two events then occurrence of A given that B has already occurred is:

$$P(A|B)=\frac{P(A\cap B)}{P(B)}; \quad P(B)\neq 0$$

Corollary:

$P(A|B)=P(A)$ When A and B are independent events.

$P(A|B)=0$ When A and B are mutually exclusive events.

5.9 Multiplicative law of probability

If A and B are any two events then occurrence of A and B together at a time is given as: $P(A\cap B)=P(A)P(B|A)$ OR $P(A\cap B)=P(B)P(A|B)$

Corollary

$P(A\cap B)=P(A)P(B)$ if A and B are independent events.

$P(A\cap B)=0$ if A and B are mutually exclusive events.

5.10 Additive law of probability

If A and B are any two events then occurrence of A or B or both / at least one of them is given as:

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

Corollaries

$P(A \cup B) = P(A) + P(B) - P(A)P(B)$ When A and B are independent events

$P(A \cup B) = P(A) + P(B) - P(A)P(B|A)$ When A and B are dependent events

$P(A \cup B) = P(A) + P(B)$ When A and B are mutually exclusive events.

Example 5.11: A pair of die is rolled what is probability of getting

(i) Sum is 6
(ii) Sum is 6 and same numbers
(iii) Sum is 6 or same numbers
(iv) Sum is 6 given that same numbers.

Solution:

$$S = \begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{Bmatrix}; \; n(S) = 36$$

(i) Let A = sum is 6

$A = \{(1,5), (2,4), (3,3), (4,2), (5,1)\}; \; n(A) = 5$ $\qquad P(A) = \frac{n(A)}{n(S)} = \frac{5}{36}$

(ii) Sum is 6 and same numbers

Let A = sum is 6; B = same numbers

$A = \{(1,5), (2,4), (3,3), (4,2), (5,1)\}$;

$B = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$

$A \cap B = \{(3,3)\}; \; n(A \cap B) = 1$

$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{1}{36}$

(iii) Sum is 6 or same numbers

Let A = sum is 6; B = same numbers

$A = \{(1,5), (2,4), (3,3), (4,2), (5,1)\}$; $\qquad P(A) = \frac{n(A)}{n(S)} = \frac{5}{36}$

$B = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$; $\qquad P(B) = \frac{n(B)}{n(S)} = \frac{6}{36}$

$A \cap B = \{(3,3)\}$; $\qquad P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{1}{36}$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A \cup B) = \frac{5}{36} + \frac{6}{36} - \frac{1}{36} = \frac{10}{36} = \frac{5}{18}$

(iv) Sum is 6 given that same numbers

Let A = sum is 6; B = same numbers

$A = \{(1,5), (2,4), (3,3), (4,2), (5,1)\}$;

$B = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$; $\qquad P(B) = \frac{n(B)}{n(S)} = \frac{6}{36}$

$A \cap B = \{(3,3)\}$; $\qquad P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{1}{36}$;

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/36}{6/36} = \frac{1}{6}$

Multiple Choice Questions

1. If A is an empty set and S is the sample space then:
(a) $P(A \cup S)=1$ (b) $P(A \cup S)=0$ (c) $P(A \cap S)=0$ (d) $P(A \cap S)=1$

2. If three coins are tossed, the probability of two heads is:
(a) 1/8 (b) 2/8 (c) 3/8 (d) 4/8

3. The term sample space is used for:
(a) All possible outcomes (b) All possible coins (c) Probability (d) Sample

4. A coin and a die can throw together:
(a) 6 ways (b) 2 ways (c) 36 ways (d) 12 ways

5. Elementary event is an event which consist of:
(a) One sample point (b) Two sample points
(c) Three sample point (d) Many sample points

6. The number of ways in which four books can be arranged on a shelf is:
(a) 8 Ways (b) 12 ways (c) 16 ways (d) 24 ways

7. A set having zero as its element is called:
(a) Null set (b) Empty set (c) Singleton set (d) Infinite set

8. ${}^4C_1$ is equal to: (a) 4 (b) 3 (c) 24 (d) 1

9. When two coins are tossed simultaneously, then *P(One head)* is:
(a) 0.25 (b) 0.50 (c) 0.75 (d) 1.00

10. A measure of chance that an uncertain event will occur:
(a) An experiment (b) An event (c) A probability (d) A trial

11. Two events A and B are mutually exclusive if:
(a) $A \cup B = \phi$ (b) $A \cup B = S$ (c) $A \cap B = S$ (d) $A \cap B = \phi$

12. ${}^nC_x$: (a) $\frac{x!}{n!(n-x)!}$ (b) $\frac{n!}{(n-x)!}$ (c) $\frac{n!}{x!(n-x)!}$ (d) $\frac{n!}{x!(x-n)!}$

13. The probability of a red card out of 52 cards is:
(a) 1/2 (b) 1/4 (c) 4/52 (d) Zero

14. The probability of drawing any one spade card is:
(a) 1/52 (b) 1/13 (c) 4/13 (d) 1/4

15. In tossing of two perfect coins, the probability of at least one head occur is:
(a) 1/2 (b) 1/4 (c) 3/4 (d) 1

16. Tossing two dice, possible outcomes are:
(a) 6 (b) 1 (c) 36 (b) 2

17. Number of ways a committee of 3 members can be selected from 5 members of a club is: (a) 10 (b) 60 (c) 15 (d) 120

18. In a set of 'n' elements, the total number of subsets are:
(a) 2^n (b) n^2 (c) n! (d) None of these

19. The experiment means a well-defined:
(a) Action (b) Outcome (c) Sample space (d) None of these

20. The probability of drawing one white ball from a bag containing 6 red, 8 black, 10 yellow and one green ball is:
(a) 1/25 (b) 0 (c) 4/13 (d) 15/20

21. If two coins are tossed, the probability of getting one head and one tail is:
(a) 1/4 (b) ½ (c) 3/4 (d) 2/3

22. The probability of an event cannot be:
(a) = 0 (b) < 0 (c) < 1 (d) > 0

23. A fair coin is tossed four times. The probability of getting four heads is:
(a) 1/4 (b) 1/2 (c) 1/16 (d) 1

24. An experiment of three fair coins tossed has sample points:
(a) 4 (b) 8 (c) 9 (d) 16

25. A person can choose a tie and a suit from 5 suits and 4 tie in:
(a) 9 ways (b) 25 ways (c) 16 ways (d) 20 ways

26. The term "event" is used for:
(a) Time (b) Total number of outcomes
(c) Subset of the sample space (d) Probability

27. How many possible permutations can be formed from the word COMMITTEE:
(a) 45360 (b) 9! (c) 6! (d) 3!

28. From a set of 10 players, 4 players can be selected in:
(a) 40 ways (b) 210 ways (c) 5040 ways (d) 400 ways

29. An orderly arrangement of things is called:
(a) Combination (b) Permutation (c) Probability (d) Sample space

30. The probability of getting an even number, when a fair die is rolled is:
(a) 1/6 (b) 1/4 (c) 1/2 (d) 1/3

31. When two coins are tossed, the possible outcomes are:
(a) 2 (b) 6 (c) 4 (d) 1

32. The probability of king of heart from the pack of 52 playing cards is:
(a) 1/13 (b) 1/4 (c) 1/2 (d) 1/52

33. Any subset of the sample space is called:
(a) Zero (b) Nothing (c) Event (d) factorial

34. The probability of drawing a picture card from a pack of 52 playing cards is:
(a) 4/52 (b) 12/52 (c) 4/13 (d) 13/52

35. The probability of selecting an even number from a set of first ten natural numbers is: (a) 1/10 (b) 4/10 (c) 0 (d) 1/2

36. A fair die is rolled, the sample space consists of ------ outcomes
(a) 2 (b) 6 (c) 8 (d) 36

37. If A and B are mutually exclusive events, then:
(a) $P(A \cap B) = 0$ (b) $P(A \cap B) = 1$ (c) $P(A \cup B) = 1$ (d) $P(A \cup B) = 0$

38. The probability of sure event is:
(a) Zero (b) More than 1 (c) One (d) 0.5

39. A coin is tossed 3 times then, the sample space is:
(a) {HHH} (b) {HHH, TTT} (c) {HHH, HTT, THT, TTH}
(d) {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}

40. The probability of vowel letters in the word STATISTICS is:
(a) 2/10 (b) 3/10 (c) 0 (d) 1

41. If a player well shuffles the pack of 52 playing cards, then the probability of a black card from 52 playing cards is:
(a) 1/52 (b) 13/52 (c) 26/52 (d) 1

42. The probability of a 'Jack card' from 52 playing cards is:
(a) Zero (b) 1/13 (c) 1/52 (d) 1/4

43. The probability of drawing a white ball from a bag containing 4 red, 8 black and 3 white balls is:
(a) 0 (b) 3/15 (c) 1/2 (d) 4/15

44. When each outcome of a sample space is as equally likely to occur as any other, the outcomes are called:
(a) Mutually exclusive (b) Equally likely (c) Exhaustive (d) Independent

45. The probability of an event always lies between:
(a) -1 and +1 (b) -1 and 0 (c) 0 and 1 (d) 0 and ∞

46. $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$, if A and B are:
(a) Mutually exclusive events (b) Independent events
(c) Not mutually exclusive events (d) equally likely

47. If $B \neq 0$, then conditional probability $P(A|B)$ is:
(a) $\frac{P(A \cup B)}{P(B)}$ (b) $\frac{P(A \cup B)}{P(A)}$ (c) $\frac{P(A \cap B)}{P(A)}$ (d) $\frac{P(A \cap B)}{P(B)}$

48. The probability of appearing a tail when a fair coin is tossed is:
(a) 0 (b) 1 (c) 1/2 (d) ¼

49. The probability of an event can never be:
(a) Zero (b) 1 (c) positive (d) negative

50. The probability of drawing a red queen card from well-shuffled pack of 52 playing cards is: (a) 4/52 (b) 2/52 (c) 13/52 (d) 26/52

51. The probability of selecting an even integer from first 20 positive integers is:
(b) 0.2 (c) 0.5 (d) 2.0

52. AUB means:
(a) The elements of A or B or both (b) The elements of A and B both
(c) The elements A but not of B (d) The elements of A or B but not both

53. The product of first 'n' natural numbers is:
(a) 1.2.3.4..... (b) 1.2.3.....$(n-2)$ $(n-1)$ (n)
(c) $n.(n-1).(n-2)$.....3.2.1.0 (d) 1.2.3.....$(n-2)$ $(n-1)$

54. The number of permutations for "r" objects taken out of total "n" objects (a) ${}^{n}C_{r}$ (b) ${}^{r}C_{n}$ (c) ${}^{n}P_{r}$ (d) ${}^{r}P_{n}$

55. If "A" denotes the males of a town and "B" denotes the females of that town, then A and B are:
(a) Equal sets (b) Over lapping sets
(c) Non-overlapping sets (d) Joint sets

56. If $P(A) = 0.3$, $P(B) = 0.8$, $P(A \cap B) = 0.24$, then events A and B are:
(a) Independent (b) Mutually exclusive
(c) Not independent (d) Exhaustive

57. A student solved 25 questions from first 50 questions of a book to be solved. The probability that he will solve the remaining all questions is:
(a) 0.25 (b) 0.5 (c) 1 (d) 0

58. If A and B are independent events, then $P(A \cap B)$=
(a) $P(A) + P(B)$ (b) $P(A) + P(B) - P(A \cap B)$
(c) $P(A).P(B)$ (d) $P(A) + P(B) + P(A \cap B)$

59. If $P(A) = 0.7$, $P(B) = 0.5$, A and B are independent events, then $P(A \cap B)$ is
(a) 0.85 (b) 0.35 (c) 1.2 (d) 0

60. The events A and B are mutually exclusive. If $P(A)=0.2$, $P(B)=0.4$, $P(A \cup B)=$: (a) 0.60 (b) 0.08 (c) 0.80 (d) 0.72

Key

Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans
1	a	2	c	3	a	4	d	5	a	6	d
7	c	8	a	9	b	10	c	11	d	12	c
13	a	14	d	15	c	16	c	17	a	18	a
19	a	20	b	21	b	22	b	23	c	24	b
25	d	26	c	27	a	28	b	29	b	30	c
31	c	32	d	33	c	34	b	35	d	36	b
37	a	38	c	39	d	40	b	41	c	42	b
43	b	44	b	45	c	46	a	47	d	48	c
49	d	50	b	51	c	52	a	53	b	54	c
55	c	56	a	57	b	58	c	59	b	60	a

Exercise

Q No. 5.1: (a) Discuss the factorial, combination and permutation.
(b) What is probability of an event?
(c) Explain the random experiment and sample space.
(d) Differentiate the following events
1) simple and compound
2) impossible and sure
3) independent and dependent
4) mutually exclusive and not mutually exclusive
5) equally likely and mutually exhaustive

Q No. 5.2: In how many ways a lock pattern of an android mobile set may be formed using (i) 4 points (ii) 5 points?

Q No. 5.3: What is the probability that a mobile set with lock pattern formed by 4 points, be unlocked by someone else than mobile owner?

Q No. 5.4: What is the probability that a mobile set with lock pattern formed by 5 points, be unlocked by someone else than mobile owner?

Q No. 5.5: A card is drawn at random from a well shuffled pack of 52 playing cards. What is the probability that the card is of a) Red card, b) Picture card, c) Red picture card, d) Ace of diamond, e) Black ace card, f) Black heart card and g) 4 of heart.

Q No. 5.6: A man is travelling from Peshawar to Lahore by car via GT road. Let A = heavy traffic somewhere on route and B = Road works somewhere on route. It is estimated that $P(A) = 0.65$ and $P(B) = 0.32$; whilst the probability of encountering both is $P(A \cap B) = 0.17$, what is the probability of encountering heavy traffic or road works?

Q No. 5.7: Sequence of 4 DNA bases in groups of three is called codone (triplet). How many codones are possible when a base is not repeated?

Q No. 5.8: List the sample space in case following genotype matting.
(i) AA×Aa (ii) AA×aa and (iii) Aa×aa

Q No. 5.9: Complete the table by inserting 0, $\frac{1}{4}$, $\frac{1}{2}$ or 1 for the probability of each genotype of the progeny from each type of mating.

Mating	Genotype		
	AA	Aa	aa
$AA \times AA$			
$AA \times Aa$			
$AA \times aa$			
$Aa \times Aa$			
$Aa \times aa$			
$aa \times aa$			

Q No. 5.10: Normal individuals have melanin pigments in their skin, hair and eyes. Albinos totally lack pigment in their body. Albinism is a recessive trait in humans. Two normal parents have an albino child. What is probability that their next child will also be an albino?

Q No. 5.11: Two new born babies get mixed up in the nursery of a hospital. Baby I is type B and baby II is of type O. Determine their parentage from the phenotypes of these two couples Mr. Haris is type A and Mrs. Haris is type AB. Mr. and Mrs. Bilal are both type A.

Q No. 5.12: A sex linked recessive allele "c" produces red blindness. Its normal dominant allele is "C". A normal woman whose father was red blind marries a red blind man. What is the probability of their children can have normal colour vision?

Q No. 5.13: A man is 45 years old and bald. His wife also has pattern baldness. What is the probability that their son will lose his hair?

Q No. 5.14: Two dice are rolled. What is probability that sum of upper faces is (i) 7, (ii) 11, (iii) more than 10, (iv) less than 5 and (v) between 6 and 9.

Q No. 5.15: A die is rolled and a coin is tossed, find the probability that the die shows an even number and the coin shows tail.

Q No. 5.16: Four coins are tossed find probability that there appears (i) only one heads (ii) only heads (iii) all tails (iv) two heads and two tails (v) two heads in a row

Q No. 5.17: A class consists 8 boys and 16 girls of which half the boys and half the girls have blood group B. A student is selected at random what is the probability that the student is a boy or has B blood group?

Q No. 5.18: A box contains 5 red, 6 yellow and 7 black balls. Three balls are drawn from this box together, find probability that (i) all of different colours (ii) all are of the same colours.

Q No. 5.19: Two dice are rolled. What is probability that number appeared on first die is greater than on second die.

Q No. 5.20: A die is rolled twice. What is probability that same number appeared both time.

Q No. 5.21: Two dice are rolled. What is probability that sum of upper faces is (i) 7, (ii) 11, (iii) divisible by 2 but not by 5, (iv) divisible by 2 or 5, (v) neither divisible by 2 nor by 5 and (vi) divisible by 2 and by 5

Q No. 5.22: A card is drawn from a pack of 52 playing cards, let A is the event that card is red and B is the event that card is a picture card. Find the probability of (i) A and B and (ii) either A or B.

Q No. 5.23: A statistics paper of BA/BSc has 10 questions of both sections and section contains five questions each. Candidates are required to attempt any five questions, in how many ways a candidate can attempt this paper if he/she has to select at least two questions from each section.

Q No. 5.24: Sum of two numbers is 10, what is the probability that both numbers are even.

Q No. 5.25: A mushroom of a certain type may produce 0, 1 or two mushrooms for next year, with probabilities 0.2, 0.6 and 0.2 respectively. If there are only two mushrooms, then find the probability that there 3 mushrooms next year.

Solution

Q No 5.2: Total points in used for making pattern for lock = $9(n=9)$

(i) Points selected for pattern = $4(r=4)$

All possible ways = ${}^{n}P_{r} = {}^{9}P_{4} = 3024$

(ii) Points selected for pattern = $5(r=5)$

All possible ways = ${}^{n}P_{r} = {}^{9}P_{5} = 15120$

Q No 5.3 Points selected for pattern = $4(r=4)$

All possible ways = $n(S) = {}^{n}P_{r} = {}^{9}P_{4} = 3024$

Let A = event that the such mobile is unlocked

$$P(A) = \frac{n(A)}{n(S)} = \frac{1}{3024} = 0.0003$$

Q No 5.4 same as 5.3 but with r = 5

Q No 5.5 Solve as example 5.9 is solved.

Q No 5.6: in this question we should use additive law of probability,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cup B) = 0.65 + 0.32 - 0.17$$
$$P(A \cup B) = 0.80$$

Q No 5.7: $n = 4;\ r = 4$

All possible triplets = ${}^{n}P_{r} = {}^{4}P_{3} = 24$

Q No 5.8:(i) $AA \times Aa:\ S = \{AA, Aa\}$ (ii) $AA \times aa:\ S = \{Aa\}$

(i) $Aa \times aa:\ S = \{Aa, aa\}$

Q No 5.9:

Mating	Sample space	Genotype probability		
		AA	Aa	aa
$AA \times AA$	$S = \{AA\}$	1	0	0
$AA \times Aa$	$S = \{AA, Aa\}$	1/2	1/2	0
$AA \times aa$	$S = \{Aa\}$	0	1	0
$Aa \times Aa$	$S = \{AA, Aa, Aa, aa\}$	¼	1/2	1/4

$Aa \times aa$	$S=\{Aa, aa\}$	0	1/2	1/2
$aa \times aa$	$S=\{aa\}$	0	0	1

Q No 5.10: normal parents with an albino child will be $Nn \times Nn$

	N	n
N	NN	Nn
n	Nn	nn

$S=\{NN, Nn, Nn, nn\}$

Let E = child is albino, $E=\{nn\}$ $\quad P(E)=\dfrac{n(E)}{n(S)}=\dfrac{1}{2}$

Q No 5.11: Possibility for Mr. and Mrs. Bilal cross: $I^A i \times I^A i$

	I^A	i
I^A	$I^A I^A$	$I^A i$
i	$I^A i$	ii

$S=\{I^A I^A, I^A i, I^A i, ii\}$

Let E_1 = child is of blood B. $E_1=\phi$ and $P(E_1)=0$ (there is no need to compute probability for second child of blood A)

There is 0% chance for Mr. and Mrs. Bilal having child with blood type B, therefore baby I is of Mr. and Mrs. Haris and baby II is of Mr. and Mrs. Bilal.

Q No 5.12: According to statement women is $X^C X^c$ and man is $X^c Y$

After cross, the sample space is

	X^C	X^c
X^c	$X^C X^c$	$X^c X^c$
Y	$X^C Y$	$X^c Y$

$S=\{X^C X^c, X^C Y, X^c Y, X^c X^c\}$

Let E = infected red blind child $E=\{X^c X^c, X^c Y\}$ $\quad P(E)=\dfrac{n(E)}{n(S)}=\dfrac{2}{4}=\dfrac{1}{2}$

Q No 5.13: According to statement man is bb and women is Bb, after cross

	b	b
B	Bb	Bb
b	bb	bb

the sample space is $S=\{Bb, bb\}$

Let E = son is bald (with no hair), $E=\{bb\}$; $P(E)=\dfrac{n(E)}{n(S)}=\dfrac{1}{2}$.

Q No 5.14: solve it like example 5.7-part b

Q No. 5.15: S=

Coin	Die					
	1	2	3	4	5	6
H	H,1	H,2	H,3	H,4	H,5	H,6
T	T,1	T,2	T,3	T,4	T,5	T,6

$n(S)=12$

Let A = die shows even and die shows tail $A=\{(T,2),(T,4),(T,6)\}$ $\quad n(a)=3$

$$P(A)=\frac{n(A)}{n(S)}=\frac{3}{12}=\frac{1}{4}$$

Q No. 5.16: S={HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT}

(i) Let A = only one head A={HTTT, THTT, TTHT, TTTH};

$$P(A)=\frac{n(A)}{n(S)}=\frac{4}{16}=\frac{1}{4}$$

(ii) Let B = only heads B={HHHH}; $P(B)=\dfrac{n(B)}{n(S)}=\dfrac{1}{16}$

(iii) Let C = All tails C={TTTT}; $P(C)=\dfrac{n(C)}{n(S)}=\dfrac{1}{16}$

(iv) Let D = two heads and two tails D={HHTT, HTHT, HTTH, THHT, THTH, TTHH}; $P(C)=\dfrac{n(D)}{n(S)}=\dfrac{6}{16}=\dfrac{3}{8}$

(v) Let E = two heads in a row E={HHTH, HHTT, HTHH, THHT, TTHH,}

$$P(E)=\frac{n(E)}{n(S)}=\frac{5}{16}$$

Q No. 5.17:

	Boys	Girls	Total
Blood group B	4	8	12
Not blood group B	4	8	12
Total	8	16	24

Let A = selected student is a boy

B = selected student has blood group B

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

$$P(A\cup B)=\frac{8}{24}+\frac{12}{24}-\frac{4}{24}$$

$$P(A\cup B)=\frac{16}{24}=\frac{2}{3}$$

Q No. 5.18:

Red	Yellow	Black	Total
5	6	7	N=18

$n=3, n(S)=\binom{N}{n}=\binom{18}{3}=816$

(i) Let A = all are of the different colour $n(A)=\binom{5}{1}\binom{6}{1}\binom{7}{1}=5\times6\times7=210$

$$P(A)=\frac{n(A)}{n(S)}=\frac{210}{816}=0.257$$

(ii) (ii) Let B = all are of the same colour $n(A)=\binom{5}{3}\binom{13}{0}+\binom{6}{3}\binom{12}{0}+\binom{7}{3}\binom{11}{0}$

$$=10+20+35=65$$

$$P(A)=\frac{n(A)}{n(S)}=\frac{65}{816}=0.080$$

Q No. 5.19:

$$S=\begin{Bmatrix}(1,1)&(1,2)&(1,3)&(1,4)&(1,5)&(1,6)\\(2,1)&(2,2)&(2,3)&(2,4)&(2,5)&(2,6)\\(3,1)&(3,2)&(3,3)&(3,4)&(3,5)&(3,6)\\(4,1)&(4,2)&(4,3)&(4,4)&(4,5)&(4,6)\\(5,1)&(5,2)&(5,3)&(5,4)&(5,5)&(5,6)\\(6,1)&(6,2)&(6,3)&(6,4)&(6,5)&(6,6)\end{Bmatrix};\qquad n(S)=36$$

Let A = Number appeared on first die is greater than on second die

$$A=\{(2,1),(3,1),(3,2),(4,1),(4,2),(4,3),(5,1),(5,2),(5,3),(5,4),(6,1),(6,2),(6,3),(6,4),(6,5)\}$$

$$n(A)=15, P(A)=\frac{n(A)}{n(S)}=\frac{15}{36}$$

Q No. 5.20: Let A = Numbers appeared on both die are same.

$$A=\{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$$

$$n(A)=6, P(A)=\frac{n(A)}{n(S)}=\frac{6}{36}=\frac{1}{6}$$

Q No. 5.21: (i) Let A = sum of numbers appeared is 7.

$$A=\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$$

$$n(A)=6, P(A)=\frac{n(A)}{n(S)}=\frac{6}{36}=\frac{1}{6}$$

(ii) Let B = sum of numbers appeared is 11.

$$B=\{(5,6),(6,5)\}\ n(B)=2, P(B)=\frac{n(B)}{n(S)}=\frac{2}{36}=\frac{1}{18}$$

(iii) Let C = sum of numbers appeared is divisible by 2 but not by 5.

$$C=\{(1,1),(1,3),(1,5),(2,2),(2,4),(2,6),(3,1),(3,3),(3,5),(4,2),(4,4),(5,1),(5,3),(6,2),(6,6),\}$$

$$n(C) = 6, P(C) = \frac{n(C)}{n(S)} = \frac{15}{36}$$

(iv) Let D = sum of numbers appeared is divisible by 2 or by 5.

$$D = \begin{Bmatrix} (1,1),(1,3),(1,4),(1,5),(2,2),(2,3),(2,4),(2,6),(3,1),(3,2),(3,3),(3,5), \\ (4,1),(4,2),(4,4),(4,6),(5,1),(5,3),(5,5),(6,2),(6,4),(6,6) \end{Bmatrix}$$

$$n(D) = 22, P(D) = \frac{n(D)}{n(S)} = \frac{22}{36}$$

(v) Let E = sum of numbers appeared is neither divisible by 2 nor by 5.

$$E = \begin{Bmatrix} (1,2),(1,6),(2,1),(2,5),(3,4),(3,6),(4,3),(4,5),(5,2),(5,4), \\ (5,6),(6,1),(6,3),(6,5) \end{Bmatrix}$$

$$n(E) = 14, P(E) = \frac{n(E)}{n(S)} = \frac{14}{36} = \frac{7}{18}$$

(vi) Let F = sum of numbers appeared is divisible by 2 and by 5.

$$F = \{(4,6),(5,5),(6,4)\}$$

$$n(F) = 3, P(F) = \frac{n(F)}{n(S)} = \frac{3}{36} = \frac{1}{12}$$

Q No. 5.22: There are 52 cards in a deck of playing cards.

Let A = red card $n(A) = 26, P(A) = \frac{n(A)}{n(S)} = \frac{26}{52}$

Let B = picture card $n(B) = 12, P(B) = \frac{n(B)}{n(S)} = \frac{12}{52}$

$$n(A \cap B) = 6, P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{6}{52}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{26}{52} + \frac{12}{52} - \frac{6}{52} = \frac{32}{52} = \frac{8}{13}$$

Q No. 5.23: All possible ways $= \binom{5}{3}\binom{5}{2} + \binom{5}{2}\binom{5}{3} = 10 \times 10 + 10 \times 10 = 100 + 100 = 200$

Q No. 5.24: $S = \{(0,10),(1,9),(2,8),(3,7),(4,6),(5,5)\}, n(S) = 6$

Let A = both numbers are even

$$A = \{(0,10),(2,8),(4,6)\}, n(A) = 3, P(A) = \frac{n(A)}{n(S)} = \frac{3}{6} = \frac{1}{2}$$

Q No. 5.25: Let A = there are 3 mushrooms next year

$$P(A) = P(2MRI \text{ and } 1MRII) + P(1MRI \text{ and } 2MRII)$$

$$= P(2MRI)\,P(1MRII) + P(1MRI)\,P(2MRII)$$

$$= 0.2 \times 0.6 + 0.6 \times 0.2 = 0.12 + 0.12 = 0.24$$

Q No. 5.26: Given information can be presented as the following

	Rural	Urban	Total
Male		**1/2×2/3=1/3**	2/3
Female	**1/2×1/3=1/6**		1/3
Total	½	1/2	1

(i) Let A = student is male and urban;
P(A) = P(male ∩ urban)=P(male)×P(urban)=1/2×2/3=1/3
Percentage of male and urban students = 1/3×100=33.33%

(ii) Let B = student is rural female;
P(B) = P(rural ∩ female)=P(rural)×P(female)=1/2×1/3=1/6
Percentage of rural female students = 1/6×100=16.67%

Q No. 5.27: S={HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT}

Let A = head and tail appear alternately A={HTHT, THTH}; $P(A) = \frac{n(A)}{n(S)} = \frac{2}{16} = \frac{1}{8}$

Chapter 6

PROBABILITY DISTRIBUTIONS

Statistician discovered the laws working in physical and non-physical world in the form of probability distributions.

6.1 Random variable

A variable that assumes numerical values from outcomes of a random phenomenon is called random variable, it is usually written as X, Y or Z. random variable is abbreviated as *r.v.* and is also known as stochastic or chance variable. For example (i) if a coin is tossed 5 times and interest is in number of heads, then random variable denoted by X and defined as number of heads in these five tosses of a coin has value: $X = 0, 1, 2, 3, 4$ and 5. (ii) weight of a plant measured by an experimenter.

There are two types of random variables, discrete and continuous.

6.1.1 Discrete random variable

A random variable which may take on only a countable or specific values such as 0,1,2,... is called discrete random variable, for examples number of children in a family, number of patients in a clinic and number of defective items in a box containing "n" total items.

6.1.2 Continuous random variable

A random variable which may take an infinite number of possible values or any value within a range is said to be continuous random variable. Continuous random variables are usually measurements. Examples are height and weight of a person, the amount of sugar in an orange and the life time of an object.

6.2 Probability distribution

Probability distribution is a statistical model that relates the values of a random variable to their probability of occurrence.

A list of values of the discrete random variable associated with its probability of occurrence is known as discrete probability distribution, for example probability distribution for the number of heads when a coin is tossed twice is

X_i	0	1	2
f	1/4	1/2	1/4

Its mathematical model is called probability mass function (pmf). Popular discrete probability distributions are binomial distribution, hyper-geometric distribution and Poisson distribution etc.

A mathematical model that expresses the probability for values of a continuous random variable is called continuous probability distribution. It is also termed as probability

density function (*pdf*). Popular continuous probability distribution are normal, exponential and gamma distribution etc.

6.2.1 Parameter

A parameter is a quantity that characterizes the probability distribution of a random variable.

6.3 Discrete Probability Distributions

6.3.1 Bernoulli trial

A trial which has only two outcomes "**success**" and "**failure**" and with constant probability is called Bernoulli trial named after Swiss mathematician **Jacob Bernoulli (1654 – 1705)**. Sowing the seedling results in survive and not survive of seedling, birth of a child results in boy or girl are examples of Bernoulli trials, because there are only two outcomes.

6.3.2 Bernoulli probability distribution

Bernoulli distribution is a discrete distribution with *pmf*.

$f(x)=p^x q^{1-x},\ x=0,1$

6.3.3 Binomial experiment

A binomial experiment is a statistical experiment that has more than one Bernoulli trials. It has the following properties

(i) The experiment consists of n repeated trials.

(ii) Each trial can result in just two possible outcomes. Call one of these outcomes as a success and the other, a failure.

(iii) The probability of success, denoted by p, is the same on every trial.

(iv) The trials are independent; it means the outcome of one trial does not affect the outcome of other trials.

6.3.4 Binomial probability distribution

A binomial random variable is the number of successes x in n repeated trials of a binomial experiment. The "probability distribution" of a binomial random variable is called a binomial distribution. Its p.d. is $f(X=x)={}^nC_x p^x q^{n-x}$, $X=0,1,2,\ldots,n$

Its notation is $X \sim b(x;n,p)$

6.3.5 Properties of binomial distribution

A binomial distribution has the following properties

01) It is a discrete probability distribution.
02) Its range is $x = 0$ to n.
03) It has two parameters "n" and "p".
04) Its mean is np and variance is npq.
05) Mean > Variance
06) (a) If $p = q$ then distribution is symmetrical.
    (b) If $p < q$ then distribution is positively skewed.
    (c) If $p > q$ then distribution is negatively skewed.

Example 6.1: A coin is tossed 4 times, find the probability of (i) two heads (ii) at least one head (iii) at most one head (iv) between 2 and 4 (both inclusive) heads and (v) find the complete binomial distribution.

Solution: $n=4$

Let X = Number of heads, then $p=\frac{1}{2}$ and $q=1-p=1-\frac{1}{2}=\frac{1}{2}$

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

$P(X=x)={}^4C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{4-x}={}^4C_x\left(\frac{1}{2}\right)^4$

(i) $P(X=2)={}^4C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{4-2}=0.375={}^4C_1\left(\frac{1}{2}\right)^4+{}^4C_2\left(\frac{1}{2}\right)^4+{}^4C_3\left(\frac{1}{2}\right)^4+{}^4C_4\left(\frac{1}{2}\right)^4=0.9375$

(ii) $P(X \geq 1)$

Or

$P(X \geq 1)=1-P(X<1)=1-P(X=0)$

$=1-{}^4C_0\left(\frac{1}{2}\right)^4=1-0.0625=0.9375$

(iii) $P(X \leq 1)={}^4C_0\left(\frac{1}{2}\right)^4+{}^4C_1\left(\frac{1}{2}\right)^4=0.3125$

(iv) $P(2 \leq X \leq 4)={}^4C_2\left(\frac{1}{2}\right)^4+{}^4C_3\left(\frac{1}{2}\right)^4+{}^4C_4\left(\frac{1}{2}\right)^4=0.6875$

(v) Complete binomial probability distribution

X_i	$P(X_i = x) = {}^4C_x \left(\frac{1}{2}\right)^4$
0	$P(X_i = 0) = {}^4C_0 \left(\frac{1}{2}\right)^4 = 0.0625$
1	$P(X_i = 1) = {}^4C_1 \left(\frac{1}{2}\right)^4 = 0.2500$
2	$P(X_i = 2) = {}^4C_2 \left(\frac{1}{2}\right)^4 = 0.3750$
3	$P(X_i = 3) = {}^4C_3 \left(\frac{1}{2}\right)^4 = 0.2500$
4	$P(X_i = 4) = {}^4C_4 \left(\frac{1}{2}\right)^4 = 0.0625$

Example 6.2: In a locality, chance for a person with blood type O^- is 0.15. A person need this type of blood. 10 persons were tested for blood group. What is the probability that four out of these ten persons' blood group match with the patient's blood group?

Solution: Here $n = 10$, X = Number of persons with blood type O^-

$p = 0.15$ and $q = 1 - p = 1 - 0.15 = 0.85$

Probability mass function for binomial distribution is:

$$P(X = x) = {}^nC_x p^x q^{n-x} \quad ; x = 0,1,2,...,10$$

$$P(X = x) = {}^{10}C_x (0.15)^x (0.85)^{10-x}$$

$$P(X = 4) = {}^{10}C_4 (0.15)^4 (0.85)^6$$

$$P(X = 4) = 0.0401$$

Example 6.3: suppose that 70% adults with allergies report symptomatic relief with a specific medication. If the medication is given to 7 new patients with allergies, what is the probability that it is effective in exactly four?

Solution: Here $n = 7$, X = Number of patients for which medication is effective

$p = 0.70$ and $q = 1 - p = 1 - 0.70 = 0.30$

Probability mass function for binomial distribution is:

$$P(X = x) = {}^nC_x p^x q^{n-x} \quad ; x = 0,1,2,...,7$$

$$P(X = x) = {}^7C_x (0.70)^x (0.30)^{7-x}$$

$$P(X = 4) = {}^7C_4 (0.70)^4 (0.30)^4$$

$$P(X = 4) = 0.2269$$

Example 6.4: In a binomial distribution $n = 5$, $p = 1/3$. Find its mean and SD.

Solution: $n = 5, p = \frac{1}{3}$ then $q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}$

So $mean = np = 5\left(\frac{1}{3}\right) = \frac{5}{3}$ and

$$SD = \sqrt{npq} = \sqrt{5\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)} = \sqrt{\frac{10}{9}} = 1.054$$

Example 6.5: In a binomial distribution mean = 4 and variance is 3. Find its parameters.

Solution: As variance $= npq \Rightarrow 3 = npq ----(i)$

$mean = np \quad \Rightarrow 4 = np ----(ii)$ Put in (i)

$3 = 4q \Rightarrow q = \frac{3}{4}, \quad p = 1 - q = 1 - \frac{3}{4} = \frac{1}{4}$ Put value of p in (ii)

$4 = n\frac{1}{4} \Rightarrow n = 16$, Hence parameters are $n = 16$ *and* $p = \frac{1}{4}$

6.3.6 Hyper-geometric experiment

A hyper-geometric experiment is a statistical experiment that has the following properties:

01) Experiment consists of "n" fixed trials.

02) Each trial can a result in just two possible outcomes. Call one of them as success and other as a failure.

03) The probability of success, denoted by p, is not same on every trial.

04) The trials are dependent; that is, the outcome on one trial affect the outcome on other trials.

6.3.7 Hyper-geometric probability distribution

The probability distribution of a hyper-geometric random variable is called a hyper-geometric distribution. It is given as $P(X=x)=\frac{{}^{k}C_{x}\,{}^{N-k}C_{n-x}}{{}^{N}C_{n}}$,

If $n<N-k$ then $X=0,1,2,\ldots,\min(n,k)$

If $n\geq N-k$ then $X=n+k-N$ to $\min(n,k)$

Its notation is $X\sim h(x,N,n,k)$.

We use hyper-geometric probability distribution when

01) A sample of size "n" is randomly selected without replacement from a population of N items.

02) In the population, k items can be classified as successes, and $N-k$ items can be classified as failures.

6.3.8 Properties of hyper-geometric distribution

01) It is discrete distribution.

02) It has three parameters n, k and N.

03) Its mean is $\frac{nk}{N}$

04) Its variance is $\frac{nk}{N}\left(\frac{N-k}{N}\right)\left(\frac{N-n}{N-1}\right)$.

Example 6.6: In a hyper-geometric distribution with $N=10$, $n=4$ and $k=5$ find $P(X=3)$ and $P(X=0)$

Solution:

Hyper-geometric distribution has probability function as

$$P(X=x)=\frac{{}^{k}C_{x}\,{}^{N-k}C_{n-x}}{{}^{N}C_{n}}=\frac{{}^{5}C_{x}\,{}^{5}C_{4-x}}{{}^{10}C_{4}}$$

$$P(X=3)=\frac{{}^{5}C_{3}\,{}^{5}C_{1}}{{}^{10}C_{4}}=\frac{50}{210}=0.238$$

$$P(X=0)=\frac{{}^{5}C_{0}\,{}^{5}C_{4}}{{}^{10}C_{4}}=\frac{5}{210}=0.0238$$

Example 6.7: In a Class there are 50 students, out of which only 4 have blood type O- Negative. If a person want two bags of this blood group and he asks 8 students at random for blood donation. What is the probability that his demand is achieved?

Solution:

Let X = Number of students with blood type O- Negative, $N=50$, $k=4$ and $n=8$

$$P(X=x)=\frac{{}^{k}C_{x}\,{}^{N-k}C_{n-x}}{{}^{N}C_{n}}=\frac{{}^{4}C_{x}\,{}^{46}C_{8-x}}{{}^{50}C_{8}}$$

$$P(X=2)=\frac{{}^{4}C_{2}\,{}^{46}C_{6}}{{}^{50}C_{8}}=\frac{56200914}{536878650}=0.105$$

6.3.9 Poisson experiment

A Poisson experiment is a statistical experiment that has the following properties:

01) The experiment results in outcomes that can be classified as successes or failures.

02) The average number of successes (μ) that occurs in a specified region is known.

03) The probability that a success will occur is proportional to the size of the region.

04) The probability that a success will occur in an extremely small region is virtually zero.

6.3.10 Poisson probability distribution

A Poisson random variable is the number of successes that result from a Poisson experiment. The probability distribution of a Poisson random variable is called a Poisson distribution. Its function is given as $P(X=x)=\frac{e^{-\mu}\mu^{x}}{x!}$, $x=0,1,2,\ldots$

Its notation is $P(x;\mu)$

6.3.11 Properties of Poisson distribution

01) It is a discrete probability distribution.

02) Its range is $x=0$ to ∞.

03) It has only one parameters μ.

04) Its mean is μ and variance is μ.

∴ Mean = Variance

Note: We use Poisson distribution when $n \geq 20$ and $p \leq 0.05$.

Example 6.8: If 150 treated seeds of wheat are sown by a researcher. The probability is 0.02 that an ear-head is attacked by smut disease. Find the probability that exactly 2 ear-heads are attacked by this disease.

Solution: Let X = Number of ear-heads attacked by smut disease.

$n = 150, p = 0.02, \mu = np = 150(0.02) = 3$

$P(X = x) = \frac{e^{-\mu}\mu^{x}}{x!}$

$P(X = 2) = \frac{e^{-3}3^{2}}{2!} = 0.2240$

6.3.12 Geometric probability distribution

A random variable that denotes the number of trials until the first success is distributed as geometric distribution. Its probability distribution is given as

$P(X = x) = (1-p)^{x-1}p,\ x = 1, 2, 3, \ldots, \infty$

has only one parameter i.e. p.

$X \sim g(x; p)$

[illegible] distribution:

and the probability mass function is $P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^{r};\ x = r, r+1, r+2, \ldots$

Its notation is $X \sim b^{*}(x; r, p)$

6.3.14 Properties of negative binomial distribution:

(i) It is a discrete distribution

(ii) It has two parameter p and r

(iii) Its mean is $\frac{r}{p}(1-p)$

(iv) Its variance is $\frac{r(1-p)}{p^{2}}$

Example 6.9: A couple decide to continue to have children until they have three males. What is the probability that their third male baby is their 5th child?

Solution: Let X = Number male of children.

$$n = 5, p = 0.5 \text{ and } q = 1 - p = 1 - 0.5 = 0.5$$

$$P(X = x) = \binom{n-1}{x-1}p^{x}q^{n-x}$$

$$P(X = 5) = \binom{5-1}{3-1}(0.5)^{2}(0.5)^{2} = 6(0.125)(0.25) = 0.1875$$

6.4 Continuous probability distribution

6.4.1 Normal distribution

Abraham de Moivre introduced the normal distribution as a limiting form of binomial distribution, when the parameter n (number of trials) is large and other parameter p is nearly equal to q. German mathematician Carl Freidrich Gauss and a French mathematician Pierre Simon Laplace derived normal distribution in early nineteenth [illegible] therefore another name of normal distribution is Gaussian probability [illegible] not assigned to a particular type, but in fact it is limiting distribution of [illegible]ibution is a standard distribution for majority of natural [illegible] first scientist who used the term "normal" for this

6.4.2 Definition

A random variable is normally distributed if its pdf is

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad -\infty<x<\infty, \qquad \text{Its notation is } X\sim N(x;\mu,\sigma^2)$$

6.4.3 The standard normal distribution

The random variable Z following a normal distribution with mean 0 and standard deviation 1 is said to follow the standard normal distribution, written as $Z\sim N(0,1)$.

6.4.4 Properties of normal distribution

01) It is continuous distribution.
02) It ranges from -∞ to +∞.
03) Its graph is bell shaped.
04) It has two parameters μ and σ^2, where μ is location and σ^2 is shape parameter.
05) Its mean is μ and variance is σ^2.
06) It is a symmetrical distribution around its mean, in which mean = median = mode = μ.
07) Normal distributions are denser in the center and less dense in the tails.
08) Special areas: $P(\mu-\sigma<X<\mu+\sigma)=0.6826$

$P(\mu-2\sigma<X<\mu+2\sigma)=0.9544$

$P(\mu-3\sigma<X<\mu+3\sigma)=0.9974$

09) Total area under normal curve is unity.
10) The curve has points of inflexion at $\mu-\sigma$ and $\mu+\sigma$
11) $Q_1=\mu-0.6745\sigma;\ Q_3=\mu+0.6745\sigma$
12) $\mu=\frac{Q_3+Q_1}{2};\ \sigma=\frac{Q_3-Q_1}{1.349}$
13) $MD=0.7979\sigma\approx\frac{4}{5}\sigma;\ QD=0.6745\sigma\approx\frac{2}{3}\sigma$

6.4.5 Importance

01) Normal curve provides a useful approximation to other probability distributions.
02) Many real data sets follow normal distribution.

6.4.6 Important formulae

01) $P(Z<a)=\Phi(a)$
02) $P(Z>a)=1-\Phi(a)$
03) $P(a<Z<b)=\Phi(b)-\Phi(a)$
04) $\Phi(-\infty)=0$ *and* $\Phi(+\infty)=1$

Example 6.10: Height of a certain plant is distributed as normal distribution with mean 135 cm and SD 20 cm. what percent of that plants have height between 130 and 160 cm.

Solution: Let X = height of the plant, $\mu=135;\ \sigma=20$

$$P(130<X<160)=P\left(\frac{130-135}{20}<\frac{X-\mu}{\sigma}<\frac{160-135}{20}\right)=P(-0.25<z<1.25)$$

$$=\Phi(1.25)-\Phi(-0.25)=0.8944-0.4013=0.4931$$

Percentage of plants with height between 130 and 160 cm is 49.31%

Example 6.11: A college has 2000 students whose mean height is 64 inches with variance 9 inches square. Find how many students have height (i) less than 60 inches (ii) more than 60 inches. **[GCUF BS Computer Science Spring 2018]**

Solution: Let X = height of a student

(i) $P(X<60)=P\left(\frac{X-\mu}{\sigma}<\frac{60-64}{3}\right)=P(Z<-1.33)=\Phi(-1.33)=0.0918$

No. of students with height less than 60 inches $=2000\times0.0918=183.6\approx184$

(ii) $P(X>60)=P\left(\frac{X-\mu}{\sigma}>\frac{60-64}{3}\right)=P(Z>-1.33)=1-\Phi(-1.33)=1-0.0918=0.9082$

No. of students with height more than 60 inches $=2000\times0.9082=1816.4\approx1816$

Multiple Choice Questions

1. A hyper geometric random variable is:
(a) Discrete (b) Continuous (c) Independent (d) none of these

2. np is the mean of :
(a) Binomial probability distribution (b) Hyper-geometric probability distribution
(c) Poisson distribution (d) Normal distribution

3. The binomial distribution is negatively skewed , if :
(a) $p<q$ (b) $p=q$ (c) $p>q$ (d) $np=npq$

4. The formula used for binomial probability distribution is :
(a) $\binom{n}{x}p^x q^{n-x}$ (b) $\dfrac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$ (c) $\dfrac{e^{-\mu}\mu^x}{x!}$ (d) $\dfrac{1}{x}e^{-x}$

5. The mean of a binomial distribution is always :
(a) Equal to variance (b) Less than variance
(c) Greater than variance (d) None of these

6. In binomial experiment the successive trials are :
(a) Dependent (b) Independent (c) Mutually exclusive (d) Exhaustive

7. When X denotes the number of successes in binomial experiment , it is called :
(a) Random variable (b) Binomial random variable
(c) Continuous random variable (d) Stochastic variable

8. A binomial random variable can only assume values :
(a) 1 to n (b) 0 to $+\infty$ (c) 0 to n (d) 0 to k

9. The binomial distribution deals with -------- variable:
(a) Discrete (b) Continuous (c) Qualitative (d) Categorical

10. In binomial experiment with three trials , the variable can take:
(a) 2 values (b) 5 values (c) 3 values (d) 4 values

11. If in binomial distribution , $\mu = 6$, $p = 3/5$, the number of trials are:
(a) 6 (b) 10 (c) 18 (d) 30

12. Which of the following probability distribution has 3 parameters :
(a) Binomial (b) Hyper-geometric (c) Poisson (d) Normal

13. In a binomial probability distribution it is impossible to find :
(a) $P(X>0)$ (b) $P(X=0)$ (c) $P(0\le X\le n)$ (d) $P(X<0)$

14. The parameters of hyper-geometric probability distribution are
(a) n,k,p (b) n,k,q (c) n,p,q (d) N,n,k

15. If we do not replace the drawn cards back into the pack before the next draw, the used probability distribution will be :
(a) Binomial (b) Hyper-geometric (c) Poisson (d) Normal

16. For a binomial probability distribution : n = 10 and the probability of failure (q = 0.6) , then mean of the distribution is :
(a) 0.6 (b) 6.0 (c) 10 (d) 4

17. In hyper-geometric probability distribution , the successive trials are:
(a) Dependent (b) Independent (c) Constant (d) Bernoulli

18. The probability of success changes from trial to trial is the property of :
(a) Binomial (b) Hyper geometric (c) Poisson (d) Normal

19. For a binomial distribution , if n = 6 , which of the following is impossible :
(a) $P(X=6)$ (b) $P(0\le X\le 6)$ (c) $P(X<6)$ (d) $P(X>6)$

20. N , n and k are parameters of :
(a) Binomial (b) Hyper-geometric (c) Poisson (d) Normal

21. For positively skewed binomial distribution
(a) $p > q$ (b) $p < q$ (c) $p = 0.5$ (d) $p = q$

22. The binomial distribution is symmetrical when :
(a) $p = q$ (b) $p = 0.25$ (c) $p > q$ (d) $p < q$

23. If in binomial distribution $p = q$, the distribution is
(a) Positively skewed (b) Negatively skewed (c) Skewed (d) symmetrical

24. In hyper geometric experiment the probability of each outcome------from trial to trial
(a) Remains constant (b) Changes (c) Depends on N (d) Depends on n

25. The sum of p and q is always
(a) 0 (b) 2 (c) 1 (d) n

26. In a binomial distribution, n = 20 and p = 3/5. then variance of this distribution is: (a) 12 (b) 24 (c) 4.8 (d) 2.4

27. In a binomial distribution, $n = 10$ and $p = 0.5$, then mean is
(a) 5 (b) 0.5 (c) 10 (d) 0.05

28. If mean of the binomial probability distribution is 4.8, then variance of this distribution may be
(a) -2.3 (b) 5.3 (c) -4.8 (d) 2.3

29. If in a binomial probability distribution mean = 6 and variance is 4, then $p =$ (a) 2/3 (b) 1 (c) 1/3 (d) 1/4

30. The mean and standard deviation of $(q + p)^5$ is
(a) np and npq (b) $5p$ and $\sqrt{npq}$ (c) $5p$ and $5\sqrt{pq}$ (d) $5p$ and $\sqrt{5pq}$

31. The bipomial distribution has ..... Parameters
(a) 1 (b) 2 (c) 3 (d) 4

32. The shape of binomial distribution depends upon
(a) q (b) n (c) p (d) n and p

33. In hyper-geometric distribution with $N = 5$, $n = 2$, $k = 3$, is it possible, $P(X = 1)$? (a) Often (b) Sometimes (c) Never (d) always

34. For a binomial distribution, $P(X = x) = {}^{5}C_x(0.5)^x(0.5)^{5-x}$, the mean is
(a) 1 (b) 1.5 (c) 3 (d) 6

35. A normal distribution has parameters
a) 1 b) 2 c) 3 d) 4

36. Marks of a class is normally distributed with mean as 71 and SD as 4. Approximately 95% students have marks between
a) 50 and 85 b) 67 and 75 c) 63 and 79 d) 59 and 83

37. Which of the following is characteristic of the normal distribution?
a) mean = SD b) three parameters c) Positively skewed d) symmetrical

38. Area to the right of the mean of the normal distribution.
a) 0 b) 0.5 c) 1 d) ∞

39. A standard normal distribution has.
a) mean = 0 b) Variance =0 c) mean = μ d) mean = SD

40. P(Z < 0.00)
a) 1 b) 0 c) 0.5 d) 1.00

41. P(-2< Z < 2)
a) 67% b) 90% c) 95% d) 99%

42. $X \sim b(30, 0.70)$, a student have to find P(17 < Z < 24). Which is correct expression?
a) $P(17.5 < X < 24.5)$ b) $P(16.5 < X < 23.5)$
c) $P(16.5 < X < 24.5)$ d) $P(17.5 < X < 23.5)$

key

Sr. No	Ans	Sr. No	Ans	Sr. No	Ans
1	a	2	a	3	c
4	a	5	c	6	b
7	b	8	c	9	a
10	d	11	b	12	b
13	d	14	d	15	b
16	d	17	a	18	b
19	d	20	b	21	b
22	a	23	c	24	b
25	c	26	c	27	a
28	d	29	c	30	d
31	b	32	d	33	d
34	c	35	b	36	c
37	d	38	b	39	a
40	c	41	c	42	d

Exercise

Q No. 6.1: (a) Define Binomial experiment, binomial distribution, hyper-geometric experiment, hyper-geometric distribution, normal distribution.
(b) What are properties of binomial probability distribution, hyper-geometric probability distribution, Poisson probability distribution and normal distribution?

Q No. 6.2: Information about random variable or parameters of a distribution is given below identify the probability distribution.

Sr. No.	Random variable/parameter	Probability Distribution
1	Leaf length of a particular plant	
2	Number of errors per page in a book	
3	$\mu = 50; \sigma^2 = 10$	
4	Number of heads in tossing a coin 4 times	
5	Marks obtained by students of a class	
6	Number of girls in a family with "n" children	
7	$\mu = 7$	
8	$n = 6; k = 5$ *and* $N = 12$	
9	$n = 8; p = 0.40$	
10	number of customers entering in a shop during a specific time interval	
11	Number of red flowers in crossing red and white flowers seeds	
12	Height of individuals in a population	
13	weight of individuals in a population	

Q No. 6.3: Against information given for each case calculate the required probabilities.

Sr. No	Information	Required probabilities
1	$n = 7; p = 0.45$	$P(X = 5); P(X \geq 5)$
2	$\mu = 3.5$	$P(X = 1); P(X = 0)$
3	$n = 4; k = 5; N = 10$	$P(X = 2); P(X = 6)$

4	$n=5; p=0.60$	$P(X=5); P(X\geq 3)$
5	$n=10; p=0.15$	$P(X=9); P(X=2)$
6	$n=8; p=0.55$	$P(X=5); P(X=10)$
7	$\mu=15; \sigma=3$	$P(X<10); P(5\leq X\leq 10)$
8	$\mu=100; \sigma=20$	$P(X>90); P(75<X<115)$
9	No. of trials=10 Probability of success=0.45	$P(X=7); P(X=0)$ $P(X=6.5)$
10	Average number of calls received by a person during 10:00 PM to 12:00 PM is 4 calls	Probability that this specific Person is not disturbed by any one's call during this Time interval.
11	Total objects = 20 Defective objectives = 4 Objects selected for Inspection = 3	X= No. of defective objects in selection $P(X=1); P(X=0)$
12	$n=7; p=0.65$	$P(X=3); P(X=-2)$
13	$n=100; p=0.40$	$P(25<X<50)$
14	No. of children = 5	There are 3 boys
15	Six coins are tossed	3 heads appeared
16	3 cards are drawn from a Pack of playing cards	(i) There are 2 diamond cards. (ii) There are all pictures cards
17	A dice is rolled 5 times	(i) There are 3 sixes (ii) there are 3 even numbers
18	$n=7; p=\frac{3}{5}$	$P(X=6); P\left(X=\frac{5}{2}\right)$

Q No. 6.4: If $X \sim \text{bin}(10,0.4)$. Find $mean(2X+5)$ and $Var(2X+5)$.

Q No. 6.5: Is it possible in binomial distribution? If not why?

(a) Mean = 5 and variance = 4
(b) $P(X=-3) = 0.25$
(c) $P(X=0) = 0.111$
(d) $P(X=½) = 0.150$
(e) $P(X=3) = 1.05$

(f) Mean = 10 and SD = 5
(g) $n=8, p=1.3$
(h) $n=7, p=0.65$
(i) There are three parameters

Q No. 6.6: is it possible in hyper-geometric distribution? If not why?

a) Mean = 10 and variance = 15
b) $N=10; k=4; n=3 \Rightarrow P(X=5)=0.1500$
c) There are two parameters N and K
d) $N=7; k=3; n=5 \Rightarrow P(X=0)=0.1800$
e) $N=8; k=3; n=5 \Rightarrow P(X=1)=0.2679$
f) $mean=\frac{nk}{N}; Variance=\left(1-\frac{nk}{N}\right)\frac{N-n}{N-1}$
g) Trials are independent.
h) Probability of success changes from trial to trial.

Q No. 6.7: Haemophilia is an X linked recessive genetic trait. If a man with haemophilia is married with a non haemophilia [hetrozyygous] woman. Their each offspring has 50% chance of having haeomophilia . this couple has 5 children, what is the probability that there are (i) two haeomophilic (ii) all heomophilic (iii) No haeomophilic and (iv) at most one heomophilic children.

Q No. 6.8: If two carriers of the gene for blue colour blind marry, each of their children has probability 0.25 of being blue colour blind. If such a couple has six children. What is the probability that none of these children are blue colour blind?

Q No. 6.9: Suppose that 80% of adults with allergies report symptomatic relief with a specific medication. If the medication is given to 10 new patients with allergies, what is the probability that it is effective in exactly seven?

Q No. 6.10: Suppose individuals with a certain gene have a 0.70 probability of eventually contracting a certain disease. If 15 individuals with the gene participate in a lifetime study, then the distribution of the random variable describing the number of individuals who will contract the disease is distributed $B(15,0.7)$. Find $P(X\leq 3)$.

Q No. 6.11: Cross-fertilizing a red and a white flower produces red flowers 25% of the time. Now we cross-fertilize five pairs of red and white flowers and produce five offspring. Find the probability that there will be no red flowered plants in the five offspring.

Q No. 6.12: An agronomist knows from past experience that 80% of a citrus variety seedling will survive being transplanted. If we take a random sample of 6 seedlings from current stock, what is the probability that exactly 2 seedlings will survive?

Q No. 6.13: If the probability of being a smoker among a group of cases with lung cancer is 0.6, what's the probability that in a group of 8 cases you have less than 2 smokers? More than 5?

Q No. 6.14: A BS Class has IQ scores with mean = 100 and SD = 15. If a student of this class has an IQ of 125, what percentage of students have higher IQs than him/her?

Q No. 6.15: A Class consists of 35 female and 15 male students. 5 students are selected to from a committee, what is the probability that the selected committee contains only male members?

Q No. 6.16: From past experience the management of a well-known fast food restaurant estimates that a number of weekly customers at a particular location follow a normal distribution with a mean of 5000 and standard deviation of 500 customer. What is the probability that on a given week the number of customers will be 4760 to 5800? **[GCUF BS Botany 2019]**

Q No. 6.17: Using the binomial distribution find the probability of 5 successes in 7 trials, when $p = 0.65$. **[GCUF BS computer Science 2019]**

Q No. 6.18: In a certain country 12% people have green eyes. If 100 people of this country are inspected, find the probability that out of these 100 people, 20 or less have green eyes.

Q No. 6.19: A contestant in a game show of a well-known TV channel has 60% chance of winning the game. In a specific episode six contestants take part, find the probability that at most 3 win the game.

Q No. 6.20: A normal random variable has mean 35 and variance 16, find $P(30 < X < 37)$.

Q No. 6.21: Let $X \sim b(40,0.65)$, find $P(20 < X < 35)$, using normal approximation.

Q No. 6.22: 60% of mice inoculated with a serum are considered protected from a certain disease. 7 mice are inoculated find the probability that four out of these mice contract the disease. Find average number of mice that can contract the disease.

Q No. 6.23: Certain cross-fertilizing of guinea pigs results in red, black and white offspring in the ratio 2:1:1. Find the probability that among 5 such offspring 4 are red.

Solution

Q No. 6.2:

Sr. No.	Random variable/parameter	Probability Distribution
1	Leaf length of a particular plant	Normal
2	Number of errors per page in a book	Poisson
3	$\mu = 50, \sigma^2 = 10$	Normal
4	Number of heads in tossing a coin 4 times	Binomial
5	Marks obtained by students of a class	Normal
6	Number of girls in a family with "n" children	Binomial
7	$\mu = 7$	Poisson
8	$n = 6, k = 5$ and $N = 12$	Hypergeometric
9	$n = 8, p = 0.40$	Binomial
10	number of customers entering in a shop during a specific time interval	Poisson
11	Number of red flowers in crossing red and white flowers seeds	Binomial
12	Height of individuals in a population	Normal
13	weight of individuals in a population	Normal

Q No. 6.3:

Sr. No	Information	Required probabilities
1	$n = 7; p = 0.45$	$P(X = 5) = 0.1172 ; P(X \geq 5) = 0.1529$
2	$\mu = 3.5$	$P(X = 1) = 0.1057 ; P(X = 0) = 0.0302$
3	$n = 4, k = 5; N = 10$	$P(X = 2) = 0.4762;\ P(X = 6) = 0$
4	$n = 5; p = 0.60$	$P(X = 5) = 0.0778;\ P(X \geq 3) = 0.6826$
5	$n = 10; p = 0.15$	$P(X = 9) = 0.0000;\ P(X = 2) = 0.2759$
6	$n = 8; p = 0.55$	$P(X = 5) = 0.2568;\ P(X = 10) = 0$
7	$\mu = 15, \sigma = 3$	$P(X < 10) = 0.0475;\ P(5 \leq X \leq 10) = 0.0471$
8	$\mu = 100, \sigma = 20$	$P(X > 90) = 0.6915;\ P(75 < X < 115) = 0.6678$

9	No. of trials=10 Probability of success=0.45	$P(X = 7) = 0.0746;\ P(X = 6) = 0.0425;\ P(X = 4.5) = 0$
10	Average number of calls received by a person during 10:00 PM to 12:00 PM is 4 calls	Probability that this specific Person is not disturbed by any one's call during this Time interval. 0.0183
11	Total objects = 20 Defective objectives = 4 Objects selected for Inspection = 3	X= No. of defective objects in selection $P(X = 1) = 0.42;\ P(X = 0) = 0.49$
12	$n = 7; p = 0.65$	$P(X = 3) = 0.1442,\ P(X = -2) = 0$
13	$n = 100, p = 0.40$	$P(25 < X < 50) = 0.9782$
14	No. of children = 5	There are 3 boys 0.3125
15	Six coins are tossed	3 heads appeared 0.3125
16	3 cards are drawn from a Pack of playing cards	(i) There are 2 diamond cards. 0.1376 (ii) There are all pictures cards 0.0099
17	A dice is rolled 5 times	(i) There are 3 sixes 0.0322 (ii) there are 3 even numbers 0.3125
18	$n = 7; p = \frac{3}{5}$	$P(X = 6) = 0.2419;\ P\left(X = \frac{5}{2}\right) = 0$

Q No. 6.4: mean = 13 and Variance = 9.6.

Q No. 6.5:

(a) Mean = 5 and variance = 4 — yes
(b) P(X = -3) = 0.25 — No (X can take only positive values)
(c) P(X = 0) = 0.111 — yes
(d) P(X = ½) = 0.150 — No (X can take only integer value)
(e) P(X = 3) = 1.05 — No (probability cannot be more than 1)
(f) Mean = 10 and SD = 5 — No (mean should be greater than variance)
(g) $n = 8; p = 1.3$ — No (value of p cannot be more than 1)

(h) $n = 7, p = 0.65$ yes

(i) There are three parameters No (binomial distribution has only two parameters)

Q No. 6.6:

a) Mean = 10 and variance = 15 No (mean < variance)

b) $N = 10; k = 4; n = 3 \Rightarrow P(X = 5) = 0.1500$ No (X cannot take its value as 5)

c) There are two parameters N and K No (hyper-geometric distribution has three parameters)

d) $N = 7; k = 3; n = 5 \Rightarrow P(X = 0) = 0.1800$ No (wrong calculation)

e) $N = 8; k = 3; n = 5 \Rightarrow P(X = 1) = 0.2679$ yes

f) $mean = \frac{nk}{N}; Variance = \left(1 - \frac{nk}{N}\right)\frac{N-n}{N-1}$ No (formula for variance is incorrect)

g) Trials are independent. No (trials are dependent)

h) Probability of success changes from trial to trial. yes

Q No. 6.7: *Let X = No. of haeomophilic children*

$$n = 5; p = 0.5; q = 0.5$$

(i) $P(X = 2) = \binom{5}{2}(0.5)^2(0.5)^3 = 10 \times 0.25 \times 0.125 = 0.3125$

(ii) $P(X = 5) = \binom{5}{5}(0.5)^5(0.5)^0 = 1 \times 0.0313 \times 1 = 0.0313$

(iii) $P(X = 0) = \binom{5}{0}(0.5)^0(0.5)^5 = 1 \times 1 \times 0.0313 = 0.0313$

(iv) $P(X \leq 1) = \binom{5}{1}(0.5)^1(0.5)^4 + \binom{5}{0}(0.5)^0(0.5)^5$

$$= 5 \times 0.5 \times 0.0625 + 1 \times 1 \times 0.03125 = 0.1875$$

Q No. 6.8: *Let X = No. of blue color blind children*

$$n = 6; p = 0.25; q = 0.75$$

$$P(X = 0) = \binom{6}{0}(0.25)^0(0.75)^6 = 1 \times 1 \times 0.1780 = 0.1780$$

Q No. 6.9: *Let X = No. of patients report relief from alergy*

$$n = 10; p = 0.80; q = 0.20$$

$$P(X = 7) = \binom{10}{7}(0.8)^7(0.2)^3 = 0.2013$$

Q No. 6.10: *Let X = No. of individual contracting a certain disease*

$$n = 15; p = 0.70; q = 0.30$$

$$P(X \leq 3) = \binom{15}{3}(0.7)^3(0.3)^{12} + \binom{15}{2}(0.7)^2(0.3)^{13} + \binom{15}{1}(0.7)^1(0.3)^{14} + \binom{15}{0}(0.7)^0(0.3)^{15}$$

$$= 0.0001$$

Q No. 6.11: *Let X = No. of red flowers* $n = 5; p = 0.25; q = 0.75$

$$P(X = 0) = \binom{5}{0}(0.25)^0(0.75)^5 = 0.2373$$

Q No. 6.12: *Let X = No. of seedling survived* $n = 6; p = 0.80; q = 0.20$

$$P(X = 2) = \binom{6}{2}(0.8)^2(0.2)^4 = 0.0154$$

Q No. 6.13: *Let X = No. of lung cancer patients being smoker*

$$n = 8; p = 0.6; q = 0.4$$

(i) $P(X < 2) = \binom{8}{1}(0.6)^1(0.4)^7 + \binom{8}{0}(0.6)^0(0.4)^8 = 0.0084$

(ii) $P(X > 5) = \binom{8}{6}(0.6)^6(0.4)^2 + \binom{8}{7}(0.6)^7(0.4)^1 + \binom{8}{8}(0.6)^8(0.4)^0 = 0.3153$

Q No. 6.14: $\mu = 100;\ \sigma = 15$

$$P(X > 125) = P\left(Z > \frac{125-100}{15}\right) = P(Z > 1.67) = 1 - \Phi(1.67) = 1 - 0.9525 = 0.0475$$

Percentage of students with higher IQ than 125 is 4.75%

Q No. 6.15:

Female	Male	Total	Sample size
35	15 = k	50=N	n = 5

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = \frac{\binom{15}{x}\binom{35}{5-x}}{\binom{50}{5}}$$

$$P(X = 5) = \frac{\binom{15}{5}\binom{35}{0}}{\binom{50}{5}} = \frac{3003 \times 1}{2118760} = 0.0014$$

Q No. 6.16: $\mu = 5000;\ \sigma = 500$

$$P(4760 < X < 5800) = P\left(\frac{4760-5000}{500} Z > \frac{5800-5000}{500}\right) = P(-0.48 < Z < 1.60)$$

$$= \Phi(1.60) - \Phi(-0.48) = 0.9452 - 0.3156 = 0.6296$$

Q No. 6.17: *Let X = No. of success* $n = 7; p = 0.65; q = 0.35$

$$P(X = 5) = \binom{7}{5}(0.65)^5(0.35)^2 = 0.2985$$

Q No. 6.18: *Let X = No. of persons with green eyes* $n = 100; p = 0.12; q = 0.88$

$\mu = np = 100(0.12) = 12$ *and* $\sigma = \sqrt{npq} = \sqrt{100(0.12)(0.88)} = 3.25$

$$P(X < 20) = P(Z < 2.46) = \Phi(2.46) = 0.9931$$

Q No. 6.19: *Let X = No. of contested that win the game*

$$n = 6; p = 0.6; q = 0.4$$

$$P(X \leq 3) = \binom{6}{3}(0.6)^3(0.4)^3 + \binom{6}{2}(0.6)^2(0.4)^4 + \binom{6}{1}(0.6)^1(0.4)^5 + \binom{6}{0}(0.6)^0(0.4)^6 = 0.45568$$

Q No. 6.20: $\mu = 35, \sigma^2 = 16, \sigma = 4$

$$P(30 < X < 37) = P(-1.25 < Z < 0.50) = \Phi(0.50) - \Phi(-1.25)$$
$$= 0.69146 - 0.10565 = 0.58581$$

Q No. 6.21: $n = 40; p = 0.65; q = 0.35$ $\mu = 26$ *and* $\sigma = 3.02$

$$P(20 < X < 35) = P(-1.98 < Z < 2.98) = 0.99856 - 0.02385 = 0.97471$$

Q No. 6.22: *Let X = No. of inoculated mice protected from a certain disease*

$$n = 7; p = 0.60; q = 0.40$$

$P(X = 4) = \binom{7}{4}(0.60)^4(0.40)^3 = 0.2903$, Average number of mice that can contract the disease is np = 7(0.60) = 4.2

Q No. 6.23: *Let X = No. of red offspring* $n = 5; p = 0.5; q = 0.5$

$$P(X = 4) = \binom{5}{4}(0.5)^4(0.5)^1 = 0.15625$$

SAMPLING

Perform your own data analyses

7.1 Population:

A real or hypothetical collection of measurements related to investigation of any problem with which an investigator is concerned is called a population or universe. Population can also be defined as the totality of observations with which we are concerned. For example if we are concerned with the study about height of plants of a specific species in an area then total plants of that type in that area is population. "*N*" is used to denote its size. If we want to research about age of students of UO, then total students in this university is population.

If there are approximately100 lions in Pakistan and we want to conduct a research about their life pattern then value of *N* is 100.

7.1.1 Types of population

(i) **Finite population:** If the size of the population is limited then it is finite population.

(ii) **Infinite population:** If the size of the population is unlimited then it is infinite population.

(iii) **Target population:** A population for which the researcher wishes to draw inference.

(iv) **Sampled population:** A population from the sample is selected.

(v) **Tangible/existent Population:** A Population whose members physically exist. e.g. ages of GCUF undergraduate students.

(vi) **Conceptual/hypothetical Population:** A Population whose members exist only in our imaginations. e.g. all possible outcome for effectiveness of a new drug for lowering blood pressure.

7.2 Sample

A sample is a representative part of the population which is selected to obtain the information concerning the characteristics of the population. Size of sample is denoted by "*n*". If we take 5 lions for our research study then $n = 5$.

7.3 Sampling

The process of selecting the sample from the population is called sampling.

Moreover an experiment that generates data for use is known as sampling. Data generated by the sampling is called sample.

7.4 Survey

The process of collecting information about a specific task is called survey.

7.4.1 Survey sampling

The process of collecting information about specific purpose using sampling about target population is called survey sampling.

7.4.2 Sampling unit

Any basic item which is selected for the purpose of sampling is called sampling unit. According to the examples discussed above lion and a plant of the specific species are sampling unit.

7.4.3 Advantages of sampling

The followings are some important advantages of sampling over complete enumeration.

01) It saves time and money.
02) Sampling provides more detailed information than census.
03) It provides more reliable results than census.
04) Sample data is used to check accuracy of census data.
05) It is essential when some units under study are destroyed.
06) Sampling is only solution if population is infinite.
07) Sampling has much smaller non response.

7.4.4 Disadvantages (Limitations) of sampling:

The followings are important disadvantages of sampling over complete enumeration.

01) Results are less reliable than census.
02) Results may mislead due to faulty sampling frame.
03) Selection of suitable sampling technique is difficult.

7.5 Parameter

A numerical quantity calculated from population is called parameter. Parameters are constants and are usually denoted by Greek letters. For example mean of population is denoted by μ, SD by σ and correlation coefficient by ρ etc.

7.6 Statistic

A numerical quantity calculated from sample is called statistic. Statistics are variables and are usually denoted by Latin letters. For example mean of sample is denoted by $\bar{x}$, SD by S and correlation coefficient by r etc.

7.7 Sampling fraction

If N is the size of population and n is sample size then n/N is called sampling fraction.

7.8 Sampling frame

List or map of all sampling units of a population is called sampling frame.

7.9 Sampling design

A statistical plan with all the steps taken in the selection of a sample is called sampling design.

7.10 Error: Difference between statistic and parameter is called error. It is given as: $E = t - \theta$

Here t is an estimator (statistic) of parameter θ.

7.10.1 Types of error

There are two types of error (i) Sampling error and (ii) non sampling error.

The error that arises as a result of taking a sample from a population rather than using the whole population is called sampling error.

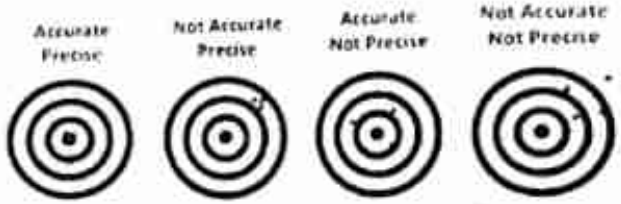
The error that arises in collection and calculation process of the data, is called non sampling error.

7.11 Bias: Bias in sampling is a systematic error which make a sample statistic non representative. It is given as: $B = E(t) - \theta$

Suppose there are 50 students in a class of BS in UO, and we give assignment to each student concerning the estimation of the root length of a plant, by selecting five plants at random. Total plants of that type is population and sample size is $n = 5$. Average root length calculated by i^{th} student is $\bar{X}_i$ (statistic/estimator). Suppose original average root length of that plant is μ (parameter), then error is $\bar{X}_i - \mu$, and bias is $E(\bar{X}_i) - \mu$

7.12 Accuracy and precision

Accuracy refers to the closeness of a measured value to a standard or known value. Precision refers to the closeness of two or more measurements to each other.



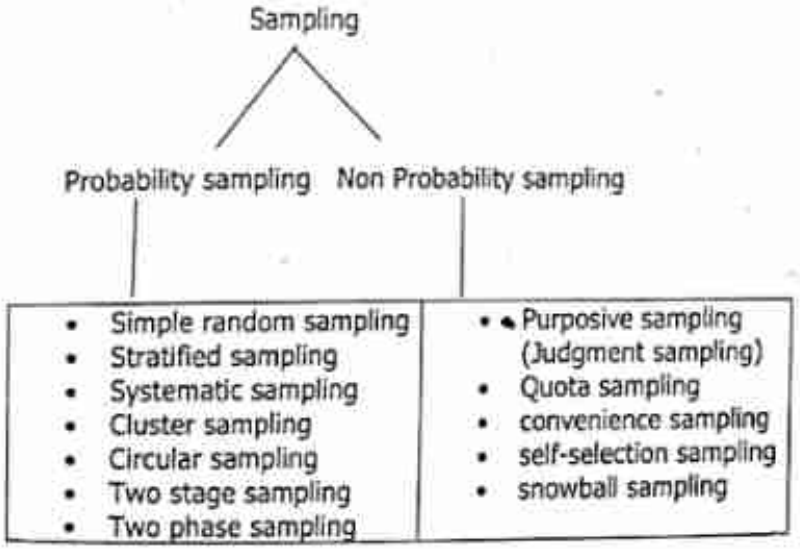
7.13 Sampling with replacement and sampling without replacement

Sampling is called with replacement if the sampling unit selected is returned to the population before drawing the next unit. All possible samples in this case are N^n. On other hand sampling is called without replacement if the sampling unit selected is not returned to the population before drawing the next unit. All possible samples in this case are ${}^{N}C_{n}$.

7.14 Probability and non-probability sampling:

When each sampling unit in a population has a known non zero probability of its being included in the sample, it is called probability sampling, it is also known as random sampling.

When the selection of sampling units is not based on probability theory and the samples are selected using personal judgment then it is called non probability (nonrandom) sampling.



7.15 Probability (random) sampling

7.15.1 Simple random sampling

A sampling technique in which, at any stage of the sampling process, each object or individual (which has not been chosen) in the population has the same probability of being chosen in the sample. SRS may be obtained by (i) Fish bowl method (Lottery method), (ii) Random Number table method and (iii) spinning method.

Simple random sampling is used when population is homogeneous.

For example in botany, selection of simple random sampling is done by (i) sampling by individual and (ii) sampling by area. When sampling units are mature trees and separate trunks define individuals and it is feasible to number all individuals. For this case sampling by individual technique is used. Sampling rhizomatous grasses, mosses, and much of the rest of the plant world by individual is usually not feasible, so sampling by area (throwing the quadrate) is used in this case.

7.15.2 Sampling distribution of a statistic

A sampling distribution of a statistic is the probability distribution of a sample statistic for all possible random samples of the same size from a population.

7.15.3 Standard error

Standard deviation of the sampling distribution of a statistic is called standard error.

7.15.4 Sampling distribution of sample mean

A sampling distribution of sample mean is the probability distribution of sample mean for all possible random samples of the same size from a population.

Measure	Population (parameter)	Sample (statistic)
Size	N	n
Mean	$\mu=\frac{\sum X}{N}$	$\bar{X}=\frac{\sum X}{n}$
Variance	$\sigma^2=\frac{\sum(X-\mu)^2}{N}$	$S^2=\frac{\sum(X-\bar{X})^2}{n}$, biased variance $s^2=\frac{\sum(X-\bar{X})^2}{n-1}$, unbiased variance
Standard deviation	$\sigma=\sqrt{\frac{\sum(X-\mu)^2}{N}}$	$S=\sqrt{\frac{\sum(X-\bar{X})^2}{n}}$ $s=\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$
Proportion	$\pi=\frac{X}{N}$	$p=\frac{X}{n}$

Mean and variance of sampling distribution of sample mean ($\bar{X}$)

$$\mu_{\bar{X}}=\sum \bar{X} f(\bar{X})$$

$$\sigma_{\bar{X}}^2=\sum \bar{X}^2 f(\bar{X})-\left[\sum \bar{X} f(\bar{X})\right]^2$$

7.15.5 Properties of sampling distribution of sample mean ($\bar{X}$)

01) Sampling distribution of sample mean has the following properties.

02) $\mu_{\bar{X}}=\mu$ (With and without replacement)

03) $\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}$ and $\sigma_{\bar{X}}^2=\frac{\sigma^2}{n}$; With replacement case

04) $\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}\sqrt{\left[\frac{N-n}{N-1}\right]}$ and $\sigma_{\bar{X}}^2=\frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right]$; Without replacement case

Example 7.1: select a random sample of size 5 from 50 students of BS botany IV semester of university of Okara, using random number table.

Solution:

There are total 50 students in the class from which we have to draw a sample of size 5.
So $N=50$ and $n=5$
At first we shall assign serial numbers to these 50 students from 01 to 50 or 00 to 49.
Then take a random number table available in books, statistical tables and on internet.
Read this random number table from anywhere combining two columns and note the first five double digit numbers within the range formed above i.e. 01 to 50 / 00 to 49.

Example 7.2: Draw all possible samples of size 2 that can be drawn with replacement from the population 2, 3, 5, 7 and 11. Make sampling distribution of sample mean and verify the following relations.

(i) $\mu_{\bar{X}}=\mu$

(ii) $\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}$

Solution: Population: 2, 3, 5, 7, 11 N = 5

Mean and standard deviation: $X=2,3,5,7,11$

$$X^2=4,9,25,49,121;\quad \sum X=28;\ \sum X^2=208$$

$$\mu=\frac{\sum X}{N}=\frac{28}{5}=5.6$$

$$\sigma=\sqrt{\frac{\sum X^2}{N}-\left(\frac{\sum X}{N}\right)^2}=\sqrt{\frac{208}{5}-\left(\frac{28}{5}\right)^2}=3.2$$

Sample size = n = 2 with replacement

Number of all possible samples = $N^n = 5^2$ = 25 and can be drawn as

2, 2	2, 3	2, 5	2, 7	2, 11	3, 2	3, 3	3, 5	3, 7
3, 11	5, 2	5, 3	5, 5	5, 7	5, 11	7, 2	7, 3	7, 5
7, 7	7, 11	11, 2	11, 3	11, 5	11, 7	11, 11		

Means of these samples $\bar{X}$ are

2	2.5	3.5	4.5	6.5	2.5	3	4	5	7	3.5	4	5	6	8	4.5	5	6	7	9
6.5	7	8	9	11															

Sampling distribution of sample means is

$\bar{X}$	f	$f(\bar{X})=\frac{f}{\sum f}$
2	1	1/25
2.5	2	2/25
3	1	1/25
3.5	2	2/25
4	2	2/25
4.5	2	2/25
5	3	3/25
6	2	2/25
6.5	2	2/25
7	3	3/25
8	2	2/25
9	2	2/25
11	1	1/25

Note:
$f(\bar{X})$ and $P(\bar{X})$ are same

Mean and variance of sampling distribution:

$\bar{X}$	$f(\bar{X})$	$\bar{X}f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
2	1/25	2/25	4/25
2.5	2/25	5/25	12.5/25
3	1/25	3/25	9/25
3.5	2/25	7/25	24.5/25
4	2/25	8/25	32/25

4.5	2/25	9/25	40.5/25
5	3/25	15/25	75/25
6	2/25	12/25	72/25
6.5	2/25	13/25	84.5/25
7	3/25	21/25	147/25
8	2/25	16/25	128/25
9	2/25	18/25	162/25
11	1/25	11/25	121/25
Σ		140/25	912/25

$$\mu_{\bar{X}} = \sum \bar{X} f(\bar{X}) = 140/25 = 5.6$$

$$\sigma_{\bar{X}} = \sqrt{\sum \bar{X}^2 f(\bar{X}) - \left[\sum \bar{X} f(\bar{X})\right]^2} = \sqrt{\frac{912}{25} - \left[\frac{140}{25}\right]^2} = 2.26$$

Verification of relationship:

$\frac{\sigma}{\sqrt{n}} = \frac{3.2}{\sqrt{2}} = 2.26$, So

(i) $\mu_{\bar{X}} = \mu = 5.6$

(ii) $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 2.26$

Example 7.3: Draw all possible sample of size 2 without replacement from the population 6, 8, 10, 12, 14 and 16. Construct sampling distribution of $\bar{X}$ and verify

$\mu_{\bar{X}} = \mu$ *and* $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\left[\frac{N-n}{N-1}\right]}$ [GCUF BS Computer Science 2019]

Solution: Population is 6, 8, 10, 12, 14 and 16. $N = 6$

Mean and standard deviation: $X = 6,8,10,12,14,16$

$$X^2 = 36,64,100,144,196,256$$

$$\sum X = 66; \sum X^2 = 796$$

$$\mu = \frac{\sum X}{N} = \frac{66}{6} = 11$$

$$\sigma=\sqrt{\frac{\sum X^2}{N}-\left(\frac{\sum X}{N}\right)^2}=\sqrt{\frac{796}{6}-\left(\frac{66}{6}\right)^2}=3.42$$

Sample size = $n = 2$ without replacement

All possible samples are ${}^6C_2 = {}^6C_4 = 15$ given with their means as under

Sr. No	Sample	$\bar{X}$	Sr. No	Sample	$\bar{X}$	Sr. No	Sample	$\bar{X}$
1	6, 8	7	6	8, 10	9	11	10, 14	12
2	6, 10	8	7	8, 12	10	12	10, 16	13
3	6, 12	9	8	8, 14	11	13	12, 14	13
4	6, 14	10	9	8, 16	12	14	12, 16	14
5	6, 16	11	10	10, 12	11	15	14, 16	15

Sampling distribution of sample mean

$\bar{X}$	f	$p(\bar{X})$
7	1	1/15
8	1	1/15
9	2	2/15
10	2	2/15
11	3	3/15
12	2	2/15
13	2	2/15
14	1	1/15
15	1	1/15

Mean and standard deviation of sampling distribution of mean.

$\bar{X}$	$p(\bar{X})$	$\bar{X}\,p(\bar{X})$	$\bar{X}^2 p(\bar{X})$
7	1/15	7/15	49/15
8	1/15	8/15	64/15
9	2/15	18/15	162/15
10	2/15	20/15	200/15
11	3/15	33/15	363/15
12	2/15	24/15	288/15
13	2/15	26/15	338/15
14	1/15	14/15	196/15
15	1/15	15/15	225/15
Σ		165/15	1885/15

$$\mu_{\bar{X}}=\sum \bar{X}\,p(\bar{X})=165/15=11$$

$$\sigma_{\bar{X}}=\sqrt{\sum \bar{X}^2\,p(\bar{X})-\left[\sum \bar{X}\,p(\bar{X})\right]^2}=\sqrt{\frac{1885}{15}-\left[\frac{165}{15}\right]^2}=2.16$$

Verification of relationship:

$$\frac{\sigma}{\sqrt{n}}\sqrt{\left[\frac{N-n}{N-1}\right]}=\frac{3.42}{\sqrt{2}}\sqrt{\frac{6-2}{6-1}}=2.16,\ \text{So}$$

(i) $\mu_{\bar{X}}=\mu=11$

(ii) $\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}\sqrt{\left[\frac{N-n}{N-1}\right]}=2.16$

Example 7.4: Draw all possible sample of size 3 without replacement from the population 6, 8, 10, 12, 14 and 16. Construct sampling distribution of $\bar{X}$ and verify that

$$\mu_{\bar{X}}=\mu \text{ and } \sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}\sqrt{\left[\frac{N-n}{N-1}\right]}$$

Solution: Population is 6, 8, 10, 12, 14 and 16. $N = 6$

Mean and standard deviation: $X = 6,8,10,12,14,16$

$$X^2 = 36,64,100,144,196,256$$

$$\sum X = 66;\ \sum X^2 = 796$$

$$\mu=\frac{\sum X}{N}=\frac{66}{6}=11$$

$$\sigma=\sqrt{\frac{\sum X^2}{N}-\left(\frac{\sum X}{N}\right)^2}=\sqrt{\frac{796}{6}-\left(\frac{66}{6}\right)^2}=3.42$$

Sample size = $n = 3$ without replacement

All possible samples are ${}^6C_3 = {}^6C_3 = 20$ given with their means as under

Sr. No	Sample	$\bar{X}$	Sr. No	Sample	$\bar{X}$	Sr. No	Sample	$\bar{X}$
1	6, 8, 10	8	8	6, 12, 14	10.67	15	8, 12, 16	12
2	6, 8, 12	8.67	9	6, 12, 16	11.33	16	8, 14, 16	12.67
3	6, 8, 14	9.33	10	6, 14, 16	12	17	10, 12, 14	12
4	6, 8, 16	10	11	8, 10, 12	10	18	10, 12, 16	12.67
5	6, 10, 12	9.33	12	8, 10, 14	10.67	19	10, 14, 16	13.33
6	6, 10, 14	10	13	8, 10, 16	11.33	20	12, 14, 16	14
7	6, 10, 16	10.67	14	8, 12, 14	11.33			

Sampling distribution of sample mean

$\bar{X}$	f	$p(\bar{X})$
8	1	1/20
8.67	1	1/20
9.33	2	2/20
10	3	3/20
10.67	3	3/20
11.33	3	3/20
12	3	3/20
12.67	2	2/20
13.33	1	1/20
14	1	1/20

Mean and standard deviation of sampling distribution of mean.

$\bar{X}$	$p(\bar{X})$	$\bar{X}\,p(\bar{X})$	$\bar{X}^2 p(\bar{X})$
8	1/20	8/20	64/20
8.67	1/20	8.67/20	75.1689/20
9.33	2/20	18.66/20	174.0978/20
10	3/20	30/20	300/20
10.67	3/20	32.01/20	341.5467/20
11.33	3/20	33.99/20	385.1067/20
12	3/20	36/20	432/20
12.67	2/20	25.34/20	321.0578/20
13.33	1/20	13.33/20	177.6889/20
14	1/20	14/20	196/20
		220/20	2466.6668/20

$$\mu_{\bar{X}} = \sum \bar{X}\, p(\bar{X}) = 220/20 = 11$$

$$\sigma_{\bar{X}} = \sqrt{\sum \bar{X}^2 p(\bar{X}) - \left[\sum \bar{X}\, p(\bar{X})\right]^2} = \sqrt{\frac{2466.6668}{20} - \left[\frac{220}{20}\right]^2} = 1.53$$

Verification of relationship:

$$\frac{\sigma}{\sqrt{n}}\sqrt{\left[\frac{N-n}{N-1}\right]} = \frac{3.42}{\sqrt{3}}\sqrt{\frac{6-3}{6-1}} = 1.53, \text{ So}$$

(i) $\mu_{\bar{X}} = \mu = 11$

(ii) $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\left[\frac{N-n}{N-1}\right]} = 1.53$

Example 7.5: Draw all possible sample of size 3 with replacement from the population 6, 8, 10 and 12. Construct sampling distribution of $\bar{X}$ and verify that

$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

Solution: Population is 6, 8, 10 and 12. $N = 4$

Mean and standard deviation: $X = 6, 8, 10, 12$

$X^2 = 36, 64, 100, 144$ $\qquad \sum X = 36;\ \sum X^2 = 344$

$$\mu = \frac{\sum X}{N} = \frac{36}{4} = 9; \qquad \sigma^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2 = \frac{344}{4} - \left(\frac{36}{4}\right)^2 = 5$$

Sample size = $n = 3$ with replacement

All possible samples are $N^n = 4^3 = 64$ given with their means as under

Sr. No	Sample	$\bar{X}$	Sr. No	Sample	$\bar{X}$	Sr. No	Sample	$\bar{X}$
1	6, 6, 6	6	23	8, 8, 10	8.67	45	10, 12, 6	9.33
2	6, 6, 8	6.67	24	8, 8, 12	9.33	46	10, 12, 8	10
3	6, 6, 10	7.33	25	8, 10, 6	8	47	10, 12, 10	10.67
4	6, 6, 12	8	26	8, 10, 8	8.67	48	10, 12, 12	11.33
5	6, 8, 6	6.67	27	8, 10, 10	9.33	49	12, 6, 6	8
6	6, 8, 8	7.33	28	8, 10, 12	10	50	12, 6, 8	8.67
7	6, 8, 10	8	29	8, 12, 6	8.67	51	12, 6, 10	9.33

Sr. No	Sample	$\bar{X}$	Sr. No	Sample	$\bar{X}$	Sr. No	Sample	$\bar{X}$
8	6, 8, 12	8.67	30	8, 12, 8	9.33	52	12, 6, 12	10
9	6, 10, 6	7.33	31	8, 12, 10	10	53	12, 8, 6	8.67
10	6, 10, 8	8	32	8, 12, 12	10.67	54	12, 8, 8	9.33
11	6, 10, 10	8.67	33	10, 6, 6	7.33	55	12, 8, 10	10
12	6, 10, 12	9.33	34	10, 6, 8	8	56	12, 8, 12	10.67
13	6, 12, 6	8	35	10, 6, 10	8.67	57	12, 10, 6	9.33
14	6, 12, 8	8.67	36	10, 6, 12	9.33	58	12, 10, 8	10
15	6, 12, 10	9.33	37	10, 8, 6	8	59	12, 10, 10	10.67
16	6, 12, 12	10	38	10, 8, 8	8.67	60	12, 10, 12	11.33
17	8, 6, 6	6.67	39	10, 8, 10	9.33	61	12, 12, 6	10
18	8, 6, 8	7.33	40	10, 8, 12	10	62	12, 12, 8	10.67
19	8, 6, 10	8	41	10, 10, 6	8.67	63	12, 12, 10	11.33
20	8, 6, 12	8.67	42	10, 10, 8	9.33	64	12, 12, 12	12
21	8, 8, 6	7.33	43	10, 10, 10	10			
22	8, 8, 8	8	44	10, 10, 12	10.67			

Sampling distribution of sample mean

$\bar{X}$	f	$f(\bar{X})$
6	1	1/64
6.67	3	3/64
7.33	6	6/64
8	10	10/64
8.67	12	12/64
9.33	12	12/64
10	10	10/64
10.67	6	6/64
11.33	3	3/64
12	1	1/64

Mean and standard deviation of sampling distribution of mean.

$\bar{X}$	$p(\bar{X})$	$\bar{X}\,p(\bar{X})$	$\bar{X}^2 p(\bar{X})$
6	1/64	6/64	36/64
6.67	3/64	20/64	133.5/64
7.33	6/64	44/64	322.4/64
8	10/64	80/64	640/64
8.67	12/64	104/64	902/64
9.33	12/64	112/64	1044.6/64
10	10/64	100/64	1000/64
10.67	6/64	64/64	683.1/64
11.33	3/64	34/64	385.1/64
12	1/20	12/64	144/64
		576/64	5290.7/64

$$\mu_{\bar{X}} = \sum \bar{X}\, p(\bar{X}) = 576/64 = 9$$

$$\sigma^2_{\bar{X}} = \sum \bar{X}^2\, p(\bar{X}) - \left[\sum \bar{X}\, p(\bar{X})\right]^2 = \frac{5290.7}{64} - \left[\frac{576}{64}\right]^2 = 1.67$$

Verification of relationship:

$\frac{\sigma^2}{n} = \frac{5}{3} = 1.67$, So

(i) $\mu_{\bar{X}} = \mu = 9$

(ii) $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = 1.67$

7.15.6 Stratified sampling

In stratified sampling, the population is partitioned into non-overlapping groups, called strata and a sample is selected from each stratum on the basis of its size. This sampling technique is appropriate when population of interest is heterogeneous.

Advantages of stratified random sampling:

01) It is more precise than simple random sampling technique.
02) It is more convenient.
03) It is more representative than simple random sample.

04) It is efficient than simple random sample.
05) Estimation for each stratum can be established.

Example 7.6: you are interested in estimating the biomass of sword fern in a 20-hectare study area. Within the area, prairie covers 15 hectares and contains just an occasional sword fern, and forest with lots of sword fern covers the remaining 5 hectares. In this example, prairie and forest would be your strata.

Example 7.7: A public sector college has 1000 total students, including 400 in FSc, 300 in FA, 150 in BA and 150 in Bsc. If 20 students are to select through stratified sampling to form a committee. Find the sample size for each stratum if proportional allocation is used.

Solution:

Population

Stratum	No. of students	
FSc	400	$N_1 = 400$
FA	300	$N_2 = 300$
BA	150	$N_3 = 150$
BSc	150	$N_4 = 150$
Total		$N = 1000$

Sample: $n_i = \frac{N_i}{N} \times n$

Stratum	Sample size
FSc	$n_1 = \frac{N_1}{N} \times n = \frac{400}{1000} \times 20 = 8$
FA	$n_2 = \frac{N_2}{N} \times n = \frac{300}{1000} \times 20 = 6$
BA	$n_3 = \frac{N_3}{N} \times n = \frac{150}{1000} \times 20 = 3$
BSc	$n_4 = \frac{N_4}{N} \times n = \frac{150}{1000} \times 20 = 3$
Total	$n = 20$

Example 7.8: A public sector college has 1000 total students, including 400 in FSc, 300 in FA, 150 in BA and 150 in Bsc. If 20 students are to select through stratified sampling to form a committee. Find the sample size for each stratum if equal allocation is used.

Solution:

population

Stratum	No. of students	
FSc	400	$N_1 = 400$
FA	300	$N_2 = 300$
BA	150	$N_3 = 150$
BSc	150	$N_4 = 150$
Total		$N = 1000$

Sample: $n_i = \frac{n}{k} = \frac{20}{4} = 5$

Stratum	Sample size
FSc	5
FA	5
BA	5
BSc	5
Total	$n = 20$

7.15.6 Systematic sampling

A method of sampling from a list of the ordered population in which the sample is made up of every k^{th} member of each group on the list, after randomly selecting a starting point from 1 to k. where $k = \frac{N}{n}$.

Example 7.9: In a class there are 50 students, we want to select 5 students from this class to form committee. Select the required 5 students from this class by using systematic sampling.

Solution: There are total 50 students in the class, so N = 50.

We have to select 5 students, i.e. $n = 5$, So $k = \frac{N}{n} = \frac{50}{5} = 10$

Step I. Form 5 groups consisting of 10 students sequentially as

Group #	Serial number
1	01 – 10
2	11 – 20
3	21 – 30
4	31 – 40
5	41 – 50

Step II. Select one student at random from the first group using random number table. Let the first selected student is of serial number 6, then the all 5 selected students for committee by systematic sampling are 6, 16, 26, 36 and 46.

7.15.7 Cluster sampling

Cluster sampling is used in statistics when groups / natural groups are present in a population. The whole population is subdivided into clusters(groups), and random samples are then collected from each cluster population. It is also known as area sampling when clusters are related to geographical regions.

7.16 Non-probability sampling

7.16.1 Judgment / Purposive Sampling

The Judgment Sampling is the non-random sampling technique whereas the choice of sample items depends exclusively on the investigator's knowledge and professional judgment.

In other words, the investigator chooses only those sample items which he feels to be the best representative of the population with regard to the attributes or characteristics under investigation. It is also known as purposive sampling because it serves the purpose of the investigator.

7.16.2 Convenience Sampling

Convenience sampling (also known as availability sampling) is a specific type of non-probability sampling method that relies on data collection from population members who are conveniently available to participate in study.

Facebook polls or questions can be mentioned as a popular example for convenience sampling.

7.16.3 Quota Sampling

Quota sampling is a non-probability sampling, in which information is collected from a specified group. It is a non-probabilistic version of stratified sampling.

Stratified	Quota
It is random sampling	It is non-random sampling
Allocation is proportional, equal or optimum	Allocation totally based on researcher's personal judgement.

7.16.4 Self-selection Sampling

when sampling units (individuals / organizations) take part in the process of sampling voluntarily, it is called self-selection sampling.

7.16.5 Snowball Sampling

Snowball sampling also known as chain referral sampling is a non-probability sampling in which selected sampling unit provide reference to other sampling units.

Multiple Choice Questions

1. If $\bar{X} = 10$ and $\mu = 12$ then sampling error is equal to.
(a) -2 (b) 2 (c) 10 (d) 12

2. In sampling with replacement, the following is always true:
(a) $n = N$ (b) $n < N$ (c) $n > N$ (d) All of these

3. In sampling without replacement, all possible number of samples:
(a) q (b) N^n (c) N (d) $\binom{N}{n}$

4. The complete list of sampling units is called:
(a) Sampling frame (b) Target population
(c) Sampling design (d) Sampling fraction

5. Population parameters are donated by ---- letters
(a) Roman (b) Greek (c) Latin (d) English

6. In sampling with replacement, a sampling unit can be selected:
(a) Only once (b) More than once (c) Less than once (d) none of these

7. If $\Sigma X = 18$, $N = 3$ then $\mu =$ ------
(a) 6 (b) 9 (c) 3 (d) 10

8. In 2,3,4,5,6 and 8, the proportion of odd number is:
(a) 1/3 (b) 1/2 (c) 1/5 (d) 1/4

9. In sampling without replacement, the relation hold:
(a) $n \neq N$ (b) $n > N$ (c) $n < N$ (d) $n \geq N$

10. In sampling without replacement a sample unit may be:
(a) Fixed (b) Not repeated (c) Repeated (d) variable

11. In random sampling, the probability of selecting an item from the population
(a) Unknown (b) Known (c) Undecided (d) One

12. Selection of questions by students to solve a paper is
(a) Random sampling (b) Non random sampling
(c) Probability (d) Sampling W.R.

13. A —— consists of the totality of the observations with which we are concerned:
(a) Population (b) Sample (c) Parameter (d) None of these

14. A population which consists of limited number of units is called
(a) Finite population (b) Infinite population (c) Uncountable (d) None of these

15. A population which consists of unlimited number of units is called:
(a) Finite population (b) Infinite population
(c) Target population (d) Sampled population

16. A representative part of the population is called
(a) Parameter (b) Statistic (c) Sampling (d) Sample

17. The process of selecting sample from the population is called:
(a) Census (b) Complete Enumeration (c) Both (A) and (B) (d) Sampling

18. A value calculated from sample is called
(a) Proportion (b) Mean (c) Statistic (d) Parameter

19. Each and every unit in the population is enumerated in
(a) Census (b) Sampling (c) Both (a) and (b) (d) None of these

20. Selection of cricket team by the selectors is :
(a) Random Sampling (b) Stratified Sampling
(c) Non random sampling (d) Probability sampling

Multiple Choice Questions

1. If $\bar{X} = 10$ and $\mu = 12$ then sampling error is equal to.
(a) -2 (b) 2 (c) 10 (d) 12

2. In sampling with replacement, the following is always true:
(a) $n = N$ (b) $n < N$ (c) $n > N$ (d) All of these

3. In sampling without replacement, all possible number of samples:
(a) q (b) N^n (c) N (d) $\binom{N}{n}$

4. The complete list of sampling units is called:
(a) Sampling frame (b) Target population
(c) Sampling design (d) Sampling fraction

5. Population parameters are donated by ---- letters
(a) Roman (b) Greek (c) Latin (d) English

6. In sampling with replacement, a sampling unit can be selected:
(a) Only once (b) More than once (c) Less than once (d) none of these

7. If $\Sigma X = 18$, $N = 3$ then μ = ------
(a) 6 (b) 9 (c) 3 (d) 10

8. In 2,3,4,5,6 and 8, the proportion of odd number is:
(a) 1/3 (b) 1/2 (c) 1/5 (d) 1/4

9. In sampling without replacement, the relation hold:
(a) $n \neq N$ (b) $n > N$ (c) $n < N$ (d) $n \geq N$

10. In sampling without replacement a sample unit may be:
(a) Fixed (b) Not repeated (c) Repeated (d) variable

11. In random sampling, the probability of selecting an item from the population
(a) Unknown (b) Known (c) Undecided (d) One

12. Selection of questions by students to solve a paper is
(a) Random sampling (b) Non random sampling
(c) Probability (d) Sampling W.R.

13. A —— consists of the totality of the observations with which we are concerned:
(a) Population (b) Sample (c) Parameter (d) None of these

14. A population which consists of limited number of units is called
(a) Finite population (b) Infinite population (c) Uncountable (d) None of these

15. A population which consists of unlimited number of units is called:
(a) Finite population (b) Infinite population
(c) Target population (d) Sampled population

16. A representative part of the population is called
(a) Parameter (b) Statistic (c) Sampling (d) Sample

17. The process of selecting sample from the population is called:
(a) Census (b) Complete Enumeration (c) Both (A) and (B) (d) Sampling

18. A value calculated from sample is called
(a) Proportion (b) Mean (c) Statistic (d) Parameter

19. Each and every unit in the population is enumerated in
(a) Census (b) Sampling (c) Both (a) and (b) (d) None of these

20. Selection of cricket team by the selectors is :
(a) Random Sampling (b) Stratified Sampling
(c) Non random sampling (d) Probability sampling

21. Sampling errors are reduced by
(a) Increasing n (b) Decreasing n (c) Increasing population (d) None of these

22. The difference between estimated and actual value of parameter is:
(a) Type I error (b) Standard error (c) Sampling error (d) Type II error

23. For making voters lists in Pakistan we need :
(a) Simple random sampling (b) Systematic sampling
(c) Quota sampling (d) Census

24. When $N = 5$, $n = 2$, then all the possible samples, drawn with replacement are:
(a) 5 (b) 7 (c) 10 (d) 25

25. For a population consisting of 2, 5, 9, 11 and 17 the sample of size "2"is taken without replacement, then the number of all possible samples:
(a) 25 (b) 10 (c) 15 (d) 32

26. The standard deviation of any sampling distribution is called:
(a) Standard error (b) Non -sampling error (c) Type I error (d) Biased error

27. Sample is subset of:
(a) Population (b) Data (c) Set (d) Distribution

Key

Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans
1	a	2	d	3	d	4	a	5	b	6	b
7	a	8	a	9	c	10	b	11	b	12	b
13	a	14	a	15	b	16	d	17	d	18	c
19	a	20	c	21	a	22	c	23	d	24	d
25	b	26	a	27	a						

Exercise

Q No. 7.1: (a) What is sampling and sampling distribution?
(b) Define the terms, population, sample, standard error, sampling frame and sampling design
(c) Differentiate between the followings
1) finite and infinite population
2) sampled and target population
3) parameter and statistic
4) probability and non-probability sampling
5) sampling and non-sampling error

(d) State one major advantage as well as disadvantage of non-probability sampling. Write name of any four non probability sampling technique. [GCUF 2019]

Q No. 7.2: Suppose you want to conduct a survey of the attitude of psychology graduate students studying clinical psychology toward psychoanalytic methods of psychotherapy. One approach would be to contact every psychology graduate student you know and ask them to fill out a questionnaire about it. What kind of sampling method is this?

Q No. 7.3: Take all possible samples of size 2 without replacement from the population comprising 2, 4, 6, 8 and 10. Make a sampling distribution of mean and verify the following relations.

i) $\mu_{\bar{x}} = \mu$ ii) $\sigma_{\bar{x}}\sqrt{n(N-1)} = \sigma\sqrt{(N-n)}$ [GCUF 2019]

Q No. 7.4: Take all possible samples of size 2 with replacement from the population comprising 1, 3, 5, 7 and 9. Make a sampling distribution of mean and verify the following relations.

i) $\mu_{\bar{x}} = \mu$ ii) $\sigma_{\bar{x}}\sqrt{n} = \sigma$

Q No. 7.5: Take all possible samples of size 3 without replacement from the population comprising 10, 12, 14, 16, 18 and 20. Make a sampling distribution of mean and verify the following relations.

i) $\mu_{\bar{x}} = \mu$ ii) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

Q N0. 7.6: Take all possible samples of size 3 with replacement from the population comprising 10, 12, 14, 16 and 18. Make a sampling distribution of mean and verify the following relations.

i) $\mu_{\bar{x}} = \mu$ ii) $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

Q No. 7.7: If $N = 7; n = 3, \mu = 25$ and $\sigma = 4$ then calculate standard error ($\sigma_{\bar{x}}$). If sampling is done with-out replacement.

Q No. 7.8: If $N = 10; n = 4, \mu = 2.5$ and $\sigma = 0.75$ then calculate mean and variance of sample mean ($\mu_{\bar{x}}, \sigma_{\bar{x}}^2$) if sampling is done without replacement.

Q No. 7.9: If $n = 2$ and $\sigma_{\bar{x}} = 1.75$ then calculate variance of the population (σ^2). If sampling is done with replacement.

Q No. 7.10: If $N = 8; n = 3$ and $\sigma_{\bar{x}} = 1.4$ then calculate standard deviation of population (σ). If sampling is done without replacement.

Q No. 7.11: If $\sigma = 2.5$ and $\sigma_{\bar{x}} = 1.25$ then find the sample size, if sampling is done with replacement.

Q No. 7.2: Convenience sampling

Q No. 7.3:

Population: 2, 4, 6, 8, 10 N = 5 all possible samples $= \binom{N}{n} = \binom{5}{2} = 10$

mean and variance of the population:

X	2	4	6	8	10	**30**
$(X-\mu)^2$	16	4	0	4	16	**40**

$$\mu = \frac{\sum X}{N} = \frac{30}{5} = 6; \quad \sigma^2 = \frac{\sum (X-\mu)^2}{N} = \frac{40}{5} = 8; \quad \sigma = 2\sqrt{2}$$

Samples:

Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$
2, 4	**3**	2, 6	**4**	2, 8	**5**	2, 10	**6**	4, 6	**5**	4, 8	**6**
4, 10	**7**	6, 8	**7**	6, 10	**8**	8, 10	**9**				

Sampling distribution of sample mean $\bar{X}$:

$\bar{X}$	f	$f(\bar{X})$	$\bar{X}f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
3	1	1/10	3/10	9/10
4	1	1/10	4/10	16/10
5	2	2/10	10/10	50/10
6	2	2/10	12/10	72/10
7	2	2/10	14/10	98/10
8	1	1/10	8/10	64/10
9	1	1/10	9/10	81/10
Total	**10**	**1**	**60/10=6**	**390/10=39**

Mean and standard deviation of sampling distribution of the sample mean $\bar{X}$:

$$\mu_{\bar{X}} = \sum \bar{X} f(\bar{X}) = 6$$

$$\sigma_{\bar{X}} = \sqrt{\sum \bar{X}^2 f(\bar{X}) - \left(\sum \bar{X} f(\bar{X})\right)^2} = \sqrt{39-(6)^2} = \sqrt{3}$$

Verification:

(i) $\mu = 6; \quad \mu_{\bar{X}} = 6 \Rightarrow \mu_{\bar{X}} = \mu$

(ii) $\sigma_{\bar{X}}\sqrt{n(N-1)} = \sqrt{3}\sqrt{2(5-1)} = \sqrt{24} = 2\sqrt{6}$

$\sigma\sqrt{(N-n)} = 2\sqrt{2}\sqrt{5-2} = 2\sqrt{6}$

$\sigma_{\bar{X}}\sqrt{n(N-1)} = \sigma\sqrt{(N-n)}$

Q No. 7.4:

Population: 1, 3, 5, 7, 9 N = 5 all possible samples $= N^n = 5^2 =$ 25

mean and variance of the population:

X	1	3	5	7	9	**25**
$(X-\mu)^2$	16	4	0	4	16	**40**

$$\mu = \frac{\sum X}{N} = \frac{25}{5} = 5;\ \sigma^2 = \frac{\sum (X-\mu)^2}{N} = \frac{40}{5} = 8;\ \sigma = 2\sqrt{2}$$

Samples:

Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$
1, 1	**1**	1, 3	**2**	1, 5	**3**	1, 7	**4**	1, 9	**5**	3, 1	**2**
3, 3	**3**	3, 5	**4**	3, 7	**5**	3, 9	**6**	5, 1	**3**	5, 3	**4**
5, 5	**5**	5, 7	**6**	5, 9	**7**	7, 1	**4**	7, 3	**5**	7, 5	**6**
7, 7	**7**	7, 9	**8**	9, 1	**5**	9, 3	**6**	9, 5	**7**	9, 7	**8**
9, 9	**9**										

Sampling distribution of sample mean $\bar{X}$:

$\bar{X}$	f	$f(\bar{X})$	$\bar{X}f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
1	1	1/25	1/25	1/25
2	2	2/25	4/10	8/25
3	3	3/25	9/10	27/25
4	4	4/25	16/25	64/25
5	5	5/25	25/25	125/25
6	4	4/25	24/25	144/25
7	3	3/25	21/25	147/25
8	2	2/25	16/25	128/25
9	1	1/25	9/25	81/25
Total	**25**	**1**	**125/25=5**	**725/25=29**

Mean and standard deviation of sampling distribution of the sample mean $\bar{X}$:

$$\mu_{\bar{X}} = \sum \bar{X} f(\bar{X}) = 5$$

$$\sigma_{\bar{X}} = \sqrt{\sum \bar{X}^2 f(\bar{X}) - \left(\sum \bar{X} f(\bar{X})\right)^2} = \sqrt{29 - (5)^2} = \sqrt{4} = 2$$

Verification:

$$\mu = 5;\ \mu_{\bar{X}} = 5 \Rightarrow \mu_{\bar{X}} = \mu$$

$$\frac{\sigma}{\sqrt{n}} = \frac{2\sqrt{2}}{\sqrt{2}} = \frac{2\sqrt{2}}{\sqrt{2}} = 2$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \Rightarrow \sigma = \sqrt{n}\sigma_{\bar{X}}$$

Q No. 7.5:

Population: 10, 12, 14, 16, 18, 20, N = 6, all possible samples $\binom{N}{n} = \binom{6}{3} = 20$

mean and variance of the population:

X	10	12	14	16	18	20	**90**
$(X-\mu)^2$	25	9	1	1	9	25	**70**

$$\mu = \frac{\sum X}{N} = \frac{90}{6} = 15;\ \sigma^2 = \frac{\sum (X-\mu)^2}{N} = \frac{70}{6} = 11.67;\ \sigma = 3.42$$

Samples:

Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$
10,12,14	12	10,12,16	38/3	10,12,18	40/3	10,12,20	14	10,14,16	40/3
10,14,18	14	10,14,20	44/3	10,16,18	44/3	10,16,20	46/3	10,18,20	16
12,14,16	14	12,14,18	44/3	12,14,20	46/3	12,16,18	46/3	12,16,20	16
12,18,20	50/3	14,16,18	16	14,16,20	50/3	14,18,20	52/3	16,18,20	18

Sampling distribution of sample mean $\bar{X}$:

$\bar{X}$	f	$f(\bar{X})$	$\bar{X}f(\bar{X})$	$\bar{X}^2 f(X)$
12	1	1/20	12/20	144/20
38/3	1	1/20	38/60	1444/180
40/3	2	2/20	80/60	3200/180
14	3	3/20	42/20	588/20
44/3	3	3/20	132/60	5808/180
46/3	3	3/20	46/20	2116/60
16	3	3/20	48/20	768/20
50/3	2	2/20	100/60	5000/180
52/3	1	1/20	52/60	2704/180
18	1	1/20	18/20	324/20
Total	**20**	**1**	**900/60=15**	**40920/180=227.33**

Mean and standard deviation of sampling distribution of the sample mean $\bar{X}$:

$$\mu_{\bar{X}} = \sum \bar{X} f(\bar{X}) = 15$$

$$\sigma_{\bar{X}} = \sqrt{\sum \bar{X}^2 f(\bar{X}) - \left(\sum \bar{X} f(\bar{X})\right)^2} = \sqrt{227.33 - (15)^2} = \sqrt{2.33} = 1.53$$

Verification:

$$\because \mu = 15;\ \mu_{\bar{X}} = 15 \Rightarrow \mu_{\bar{X}} = \mu$$

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{3.42}{\sqrt{3}}\sqrt{\frac{6-3}{6-1}} = \frac{3.42}{\sqrt{3}}\sqrt{\frac{3}{5}} = 1.53$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Q NO. 7.6: Population: 10, 12, 14, 16, 18, 20, N = 5, all possible samples =125
mean and variance of the population:

X	10	12	14	16	18	**70**
$(X-\mu)^2$	16	4	0	4	16	**40**

$$\mu = \frac{\sum X}{N} = \frac{70}{5} = 14;\ \sigma^2 = \frac{\sum (X-\mu)^2}{N} = \frac{40}{5} = 8$$

Samples:

Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$
10,10,10	10	10,10,12	32/3	10,10,14	34/3	10,10,16	12	10,10,18	38/3
10,12,10	32/3	10,12,12	34/3	10,12,14	12	10,12,16	38/3	10,12,18	40/3
10,14,10	34/3	10,14,12	12	10,14,14	38/3	10,14,16	40/3	10,14,18	14
10,16,10	12	10,16,12	38/3	10,16,14	40/3	10,16,16	14	10,16,18	44/3
10,18,10	38/3	10,18,12	40/3	10,18,14	14	10,18,16	44/3	10,18,18	46/3
12,10,10	32/3	12,10,12	34/3	12,10,14	12	12,10,16	38/3	12,10,18	40/3
12,12,10	34/3	12,12,12	12	12,12,14	38/3	12,12,16	40/3	12,12,18	14
12,14,10	12	12,14,12	38/3	12,14,14	40/3	12,14,16	14	12,14,18	44/3
12,16,10	38/3	12,16,12	40/3	12,16,14	14	12,16,16	44/3	12,16,18	46/3
12,18,10	40/3	12,18,12	14	12,18,14	44/3	12,18,16	46/3	12,18,18	16
14,10,10	34/3	14,10,12	12	14,10,14	38/3	14,10,16	40/3	14,10,18	14
14,12,10	12	14,12,12	38/3	14,12,14	40/3	14,12,16	14	14,12,18	44/3
14,14,10	38/3	14,14,12	40/3	14,14,14	14	14,14,16	44/3	14,14,18	46/3
14,16,10	40/3	14,16,12	14	14,16,14	44/3	14,1616	46/3	14,16,18	16
14,18,10	14	14,18,12	44/3	14,18,14	46/3	14,28,16	16	14,18,18	50/3
16,10,10	12	16,10,12	38/3	16,10,14	40/3	16,10,16	14	16,10,18	44/3
16,12,10	38/3	16,12,12	40/3	16,12,14	14	16,12,16	44/3	16,12,18	46/3
16,14,10	40/3	16,14,12	14	16,14,14	44/3	16,14,16	46/3	16,14,18	16

16,16,10	14	16,16,12	44/3	16,16,14	46/3	16,16,16	16	16,16,18	50/3
16,18,10	44/3	16,18,12	46/3	16,18,14	16	16,18,16	50/3	16,18,18	52/3
18,10,10	38/3	18,10,12	40/3	18,10,14	14	18,10,16	44/3	18,10,18	46/3
18,12,10	40/3	18,12,12	14	18,12,14	44/3	18,12,16	46/3	18,12,18	16
18,14,10	14	18,14,12	44/3	18,14,14	46/3	18,14,16	16	18,14,18	50/3
18,16,10	44/3	18,16,12	46/3	18,16,14	16	18,16,16	50/3	18,16,18	52/3
18,18,10	46/3	18,18,12	16	18,18,14	50/3	18,18,16	52/3	18,18,18	18

i) Sampling distribution of sample mean $\bar{X}$:

$\bar{X}$	f	$f(\bar{X})$	$\bar{X} f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
10	1	1/125	10/125	100/125
32/3	3	3/125	32/125	1024/375
34/3	6	6/125	68/125	2312/375
12	10	10/125	120/125	1440/125
38/3	15	15/125	190/125	7220/375
40/3	18	18/125	240/125	9600/375
14	19	19/125	266/125	3724/125
44/3	18	18/125	264/125	11616/375
46/3	15	15/125	230/125	10580/375
16	10	10/125	160/125	2560/125
50/3	6	6/125	100/125	5000/375
52/3	3	3/125	52/125	2704/375
18	1	1/125	18/125	324/125
Total	**125**	**1**	**1750/125=14**	**74500/375=198.67**

Mean and standard deviation of sampling distribution of the sample mean $\bar{X}$:

$$\mu_{\bar{X}} = \sum \bar{X} f(\bar{X}) = 14$$

$$\sigma_{\bar{X}}^2 = \sum \bar{X}^2 f(\bar{X}) - \left(\sum \bar{X} f(\bar{X})\right)^2 = 198.67 - (14)^2 = 2.67$$

Verification:

$$\mu = 14;\ \mu_{\bar{X}} = 14 \Rightarrow \mu_{\bar{X}} = \mu$$

$$\frac{\sigma^2}{n} = \frac{8}{3} = 2.67$$

$$\frac{\sigma^2}{n} = \sigma_{\bar{X}}^2$$

Q No. 7.7:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{4}{\sqrt{3}}\sqrt{\frac{7-3}{7-1}} = \frac{4}{\sqrt{3}}\sqrt{\frac{4}{6}} = 1.89$$

Q No. 7.8:

$$\mu_{\bar{x}} = \mu = 2.5$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right] = \frac{(0.75)^2}{4}\left[\frac{10-4}{10-1}\right] = \frac{(0.75)^2}{4} \times \frac{6}{9} = 0.094$$

Q No. 7.9:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \Rightarrow (1.75)^2 = \frac{\sigma^2}{2} \Rightarrow \sigma^2 = 2 \times (1.75)^2 = 6.125$$

Q No. 7.10:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \Rightarrow 1.4 = \frac{\sigma}{\sqrt{3}}\sqrt{\frac{8-3}{8-1}} \Rightarrow 1.4 = \frac{\sigma}{\sqrt{3}}\sqrt{\frac{5}{7}} \Rightarrow \sigma = 1.4\sqrt{3}\sqrt{\frac{7}{5}} = 2.86$$

Q No. 7.11:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \Rightarrow (1.25)^2 = \frac{(2.5)^2}{n} \Rightarrow n = \frac{(2.5)^2}{(1.25)^2} = 4$$

Chapter 8

TESTING OF HYPOTHESIS

(t and Z test)

Statistics is the back bone of research

8.1 Testing of hypothesis

Testing of hypothesis is defined as the formal procedures used by a researcher to accept or reject a statistical hypothesis.

8.2 Statistical hypothesis

A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true.

8.2.1 Null hypothesis

The null hypothesis, denoted by H_0, is the hypothesis which is to be tested for possible rejection assuming that it is true.

8.2.2 Alternative hypothesis

The alternative hypothesis, denoted by H_1 or H_a, is the hypothesis that is formulated by sample observations. It is also known as research hypothesis.
Null and alternative hypotheses are opposite to each other.

8.2.3 Simple hypothesis

A simple hypothesis is one in which all parameters of the distribution are specified, e.g. $H_0: \mu = 3, H_0: \sigma^2 = 12, H_0: \mu_1 = \mu_2$ etc.

8.2.4 Composite hypothesis

A hypothesis which is not simple (i.e. in which parameters are not specified) is called a composite hypothesis, $H_0: \mu \geq 3$, $H_0: \sigma^2 \leq 12$, $H_0: \mu_1 \geq \mu_2$ etc.

8.3 Type 1 error (False positive)

A Type 1 error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type 1 error is called the significance level. This probability is also called alpha, and is often denoted by α.

8.4 Type II error (False negative)
A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called Beta, and is often denoted by β. The probability of *not* committing a Type II error is called the Power of the test.

8.5 Test statistic
A test statistic is a quantity calculated from our sample data. Its value is used to decide whether or not the null hypothesis should be rejected in our hypothesis test.

8.6 Critical region
The critical region (CR), or rejection region (RR), is a set of values of the test statistic for which the null hypothesis is rejected in a hypothesis test.

8.7 One sided (tailed) and two sided (tailed) test
A statistical hypothesis test in which the critical region is located entirely in one tail of the sampling distribution of the test statistic is called one sided(tailed) test.
A statistical hypothesis test in which the critical region is located in both tails of the sampling distribution of the test statistic is called two sided(tailed) test.

8.9 Level of significance
The probability of rejecting the null hypothesis in a statistical test when it is true is called level of significance. It is denoted by α and is given as $P(\text{rejecting } H_0 | H_0 \text{ is true}) = \alpha$

8.10 Level of confidence:
$1-\alpha$ that is probability of accepting H_0 when it is true is called level of confidence. i.e $P(\text{accepting } H_0 | H_0 \text{ is true}) = 1-\alpha$

8.11 Power of the test:
$1-\beta$ that is probability of rejecting H_0 when it is false is called power of the test. i.e $P(\text{Rejecting } H_0 | H_0 \text{ is } false) = 1-\beta$, where $P(\text{Accepting } H_0 | H_0 \text{ is } false) = \beta$ is probability of type-II error.

8.12 Acceptance region
Values of the sampling distribution of the test statistic which leads to the acceptance of null hypothesis H_0 is known as acceptance region.

8.13 Rejection region
Values of the sampling distribution of the test statistic which leads to the rejection of null hypothesis H_0 is known as rejection region. It is also called the critical region.

8.14 Critical value:
The value(s) that separates the rejection region from the acceptance region is called the critical value(s)

8.15 Important Tests of hypothesis
8.15.1 Single population mean:
Hypothesis: $H_0: \mu = \mu_0 / \mu \le \mu_0 / \mu \ge \mu_0$
$H_1: \mu \ne \mu_0 / \mu > \mu_0 / \mu < \mu_0$

Level of significance: α

Test statistic: $z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$; when σ is known

$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$; when σ is unknown

Note: we can also use z test when σ is unknown and sample size is large ($n > 30$)

Critical value/Critical Region:

Hypothesis	Critical Region		
$H_0: \mu = \mu_0$; $H_1: \mu \ne \mu_0$	$Z > Z_{\alpha/2}$ and $Z < -Z_{\alpha/2}$	$t > t_{\frac{\alpha}{2},\nu}$ and $t < -t_{\frac{\alpha}{2},\nu}$	where
$H_0: \mu \ge \mu_0$; $H_1: \mu < \mu_0$	$Z < -Z_\alpha$	$t < -t_{\alpha,\nu}$	$\nu = n-1$
$H_0: \mu \le \mu_0$; $H_1: \mu > \mu_0$	$Z > Z_\alpha$	$t > t_{\alpha,\nu}$	

Conclusion: We reject H_0 when calculated value falls in CR

Example 8.1(a): A sample of 25 fruits fly [Drosophila melanogaster] larva was incubated at 37 degree centigrade for 30 minutes. It is theorized that such exposure to heat causes polygene chromosome located in the salivary glands of the fly to unwind, creating puffs on

the chromosomes arm, that visible under a microscope. The average number of puffs for the 25 observations was 4.3 with a SD of 2.0. Test the hypothesis that average number of puffs in population is 4.5.

Solution: $H_0: \mu = 4.5 \quad H_1: \mu \neq 4.5$

$\alpha = 0.05$

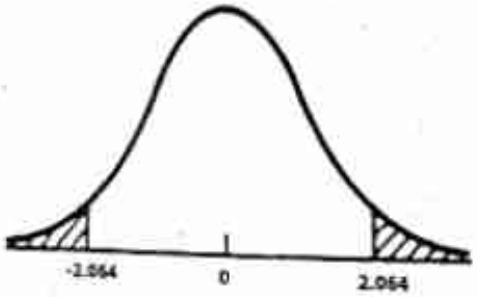
Test statistic: $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$

Critical value: $t_{(\alpha, 2n)} = t_{(0.05, 2n-1)} = t_{(0.025, 24)} = 2.064$

Critical region: $t > 2.064$ *and* $t < -2.064$

Calculation: $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{4.3 - 4.5}{2.0/\sqrt{25}} = -0.500$

Conclusion:



As calculated value t(=-0.500) does not lie in CR so we may accept $H_0: \mu = 4.50$

Example 8.1(b): Average body temperature in a sample of 36 intertidal crabs placed in air at 24.3C° is 25.03C°. variance of the population is known to be 1.80C°², Test the hypothesis $H_0: \mu = 24.3$ against $H_1: \mu \neq 24.3$

Solution: $H_0: \mu = 24.3 \quad H_1: \mu \neq 24.3$

$\alpha = 0.05$

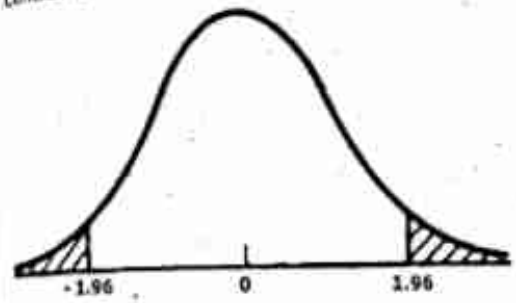
Test statistic: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

Critical value: $Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$

Critical region: $Z > 1.96$ *and* $Z < -1.96$

Calculation: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{25.03 - 24.3}{1.34/\sqrt{36}} = \frac{0.73}{0.224} = 3.259$

Conclusion:



As calculated value (Z=3.259) lies in CR so we may not accept $H_0: \mu = 24.3$

8.15.2 Difference between two population means:

Hypothesis: $H_0: \mu_1 = \mu_2 \,/\, \mu_1 \leq \mu_2 \,/\, \mu_1 \geq \mu_2$

$H_1: \mu_1 \neq \mu_2 \,/\, \mu_1 > \mu_2 \,/\, \mu_1 < \mu_2$

Test statistic: $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$;when σ_1 and σ_2 are known

$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$;when σ_1 and σ_2 are unknown

Where $s_p = \sqrt{\frac{\sum(X_1 - \bar{X}_1)^2 + \sum(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$; $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$; $s_p = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}$

Note: we can also use z test when σ_1 and σ_2 are unknown and sample sizes are large ($n_1, n_2 > 30$)

Critical value/Critical Region:

Hypothesis	Critical Region		
$H_0: \mu_1 = \mu_2;\ H_1: \mu_1 \neq \mu_2$	$Z > Z_{\alpha/2}$ *and* $Z < -Z_{\alpha/2}$	$t > t_{\frac{\alpha}{2},\nu}$ and $t < -t_{\frac{\alpha}{2},\nu}$	Where $\nu = n_1 + n_2 - 1$
$H_0: \mu_1 \leq \mu_2;\ H_1: \mu_1 > \mu_2$	$Z > Z_\alpha$	$t > t_{\alpha,\nu}$	
$H_0: \mu_1 \geq \mu_2;\ H_1: \mu_1 < \mu_2$	$Z < -Z_\alpha$	$t < -t_{\alpha,\nu}$	

Example 8.2(a): The following data are of heights (cm) of plants, each grown with one of two different fertilizers. Test the hypothesis that fertilizer II is better than fertilizer I

Fertilizer I	48.2	54.6	58.3	47.8	51.4	49.1	49.9
Fertilizer II	52.3	57.4	55.6	53.2	61.3	54.8	

Solution:

X_I	X_{II}	X_I^2	X_{II}^2
48.2	52.3	2323.4	2735.29
54.6	57.4	2981.16	3294.76
58.3	55.6	3398.89	3091.36
47.8	53.2	2284.84	2830.24
51.4	61.3	2641.96	3757.69
49.1	54.8	2410.81	3003.04
49.9		2490.01	
359.3	**334.6**	**18530.91**	**18712.38**

$$\bar{X}_I = \frac{\sum X_I}{n_I} = \frac{359.3}{7} = 51.33;\ \bar{X}_{II} = \frac{\sum X_{II}}{n_{II}} = \frac{334.6}{6} = 55.77$$

$$S_I^2 = \frac{\sum X_I^2}{n_I} - \left(\frac{\sum X_I}{n_I}\right)^2 = \frac{18530.91}{7} - \left(\frac{359.3}{7}\right)^2 = 12.65$$

$$S_{II}^2 = \frac{\sum X_{II}^2}{n_{II}} - \left(\frac{\sum X_{II}}{n_{II}}\right)^2 = \frac{18712.38}{6} - \left(\frac{334.6}{6}\right)^2 = 8.81$$

$$s_p = \sqrt{\frac{n_I S_I^2 + n_{II} S_{II}^2}{n_I + n_{II} - 2}} = \sqrt{\frac{7\times 12.65 + 6\times 8.81}{7+6-2}} = \sqrt{\frac{141.41}{11}} = 3.58$$

$H_0: \mu_{II} = \mu_I \quad H_1: \mu_{II} > \mu_I$

$\alpha = 0.05$

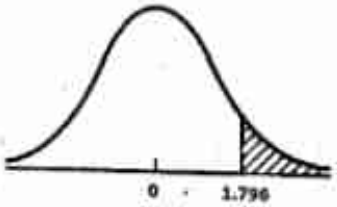
Test statistic: $t = \dfrac{(\bar{X}_{II} - \bar{X}_I) - (\mu_{II} - \mu_I)}{s_p\sqrt{\dfrac{1}{n_{II}} + \dfrac{1}{n_I}}}$

Critical value: $t_{0.05} = t_{(0.05,\ 7+6-2)} = t_{(0.05,\ 11)} = 1.796$

Critical region: $t > 1.796$

Calculation: $t = \dfrac{(\bar{X}_{II} - \bar{X}_I) - (\mu_{II} - \mu_I)}{s_p\sqrt{\dfrac{1}{n_{II}} + \dfrac{1}{n_I}}} = \dfrac{55.77 - 51.33 - 0}{3.58\sqrt{\dfrac{1}{6} + \dfrac{1}{7}}} = \dfrac{4.44}{2.0} = 2.22$

Conclusion:



As calculated value of t(=2.22) lies in CR so we may reject $H_0: \mu_{II} = \mu_I$ in favour of $H_1: \mu_{II} > \mu_I$. It means new fertilizer is better than present fertilizer.

Example 8.2(b): A sample of 30 plants from a population of a certain species has a mean height 10 cm whereas a second sample of 25 plants from another population of plants of same species has mean height 13 cm. Variance of this species of plant is known to be 14. Test the hypothesis that second population taller than first population

Solution: $n_1 = 30, n_2 = 25; \bar{X}_1 = 10, \bar{X}_2 = 13; \sigma^2 = \sigma_1^2 = \sigma_2^2 = 14; \sigma = \sqrt{14} = 3.74$

$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 < \mu_2$

$\alpha = 0.05$

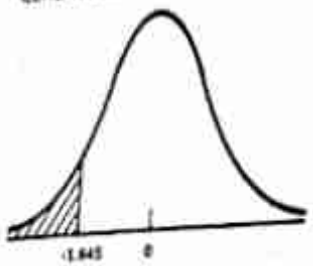
Test statistic: $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

Critical value: $Z_\alpha = Z_{0.05} = 1.645$

Critical region: $Z < -1.645$

Calculation: $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{10 - 13 - 0}{3.74\sqrt{\dfrac{1}{30} + \dfrac{1}{25}}} = \dfrac{-3}{1.01} = -2.970$

Conclusion:



As calculated value of Z(= -2.970) lies in CR so we may not accept $H_0: \mu_1 = \mu_2$ in favour of $H_1: \mu_1 < \mu_2$ It means second population is taller than first population.

8.15.3 Paired observations case:

Hypothesis: $H_0: \mu_d = 0 / \mu_d \leq 0 / \mu_d \geq 0$

$H_1: \mu_d \neq 0 / \mu_d > 0 / \mu_d < 0$

Test statistic: $t = \dfrac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$

Where $s_d = \sqrt{\dfrac{1}{n-1}\left[\sum d^2 - \dfrac{(\sum d)^2}{n}\right]}$

Critical value/Critical Region:

Hypothesis	Critical value	Critical Region
$H_0: \mu_d = 0;\ H_1: \mu_d \neq 0$	$t_{(\alpha/2,\upsilon)};\ \upsilon = n-1$	$t > t_{(\alpha/2,\upsilon)}$ and $t < -t_{(\alpha/2,\upsilon)}$
$H_0: \mu_d \leq 0;\ H_1: \mu_d > 0$	$t_{(\alpha,\upsilon)};\ \upsilon = n-1$	$t > t_{(\alpha,\upsilon)}$
$H_0: \mu_d \geq 0;\ H_1: \mu_d < 0$	$t_{(\alpha,\upsilon)};\ \upsilon = n-1$	$t < -t_{(\alpha,\upsilon)}$

Example 8.3 (a): In a study on the comparison of sorbic acid in red meat before and after storage the following data on sorbic acid residuals in parts per million of 7 slices of red meat immediately after dipping in a sorbate solution and after 60 days of storage were recorded.

Sorbic acid residuals								
	Before storage	124	100	250	344	440	660	700
	After storage	116	96	239	329	437	597	689

Assuming the populations to be normally distributed, is there sufficient evidence, at the 0.05 level of significance, to say that the length of storage influences sorbic acid residual concentrations?

Solution: $\mathbf{d = X_B - X_A}$

X_B	X_A	d	d^2
124	116	8	64
100	96	4	16
250	239	11	121
344	329	15	225
440	437	3	9
660	597	63	3969
700	689	11	121
Σ		**115**	**4525**

$$\bar{d} = \frac{\sum d}{n} = \frac{115}{7} = 16.43$$

$$s_d = \sqrt{\frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]} = \sqrt{\frac{1}{6}\left[4525 - \frac{(115)^2}{7}\right]} = 20.96$$

$H_0: \mu_d = 0 \quad H_1: \mu_d > 0$

$\alpha = 0.05$

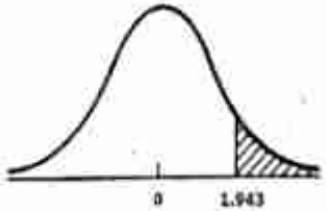
Test statistic: $t = \dfrac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$

Critical value: $t_{(\alpha,\upsilon)} = t_{(0.05,7-1)} = t_{(0.05,6)} = 1.943$

Critical region: $t > 1.943$

Calculation: $t = \dfrac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \dfrac{16.43 - 0}{20.96 / \sqrt{7}} = \dfrac{16.43}{7.922} = 2.074$

Conclusion:



As calculated value t(=2.074) lies in CR so we may reject $H_0: \mu_d = 0$. It means length of storage has influence on sorbic acid residual

Example 8.3 (b): Difference between hindleg and foreleg length of 10 deer (cm) are given below 3, 2, -3, 4, -1, 4, -5, 4, 5, -1. Test the hypothesis that there is no difference between length of hindleg and foreleg of deer.

Solution:

d	d^2
3	9
2	4
-3	9
4	16
-1	1
4	16
-5	25
4	16
5	25
-1	1
22	122

$$\bar{d} = \frac{\sum d}{n} = \frac{22}{10} = 2.2$$

$$s_d = \sqrt{\frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]} = \sqrt{\frac{1}{9}\left[122 - \frac{(22)^2}{10}\right]} = 2.86$$

$H_0: \mu_d = 0 \quad H_1: \mu_d \neq 0$

$\alpha = 0.05$

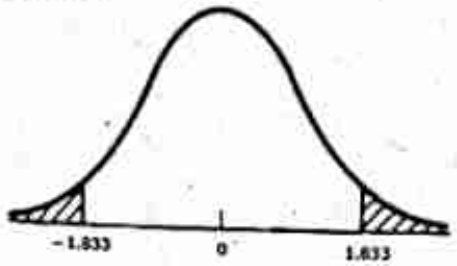
Test statistic: $t = \dfrac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$

Critical value: $t_{(\alpha/2,n)} = t_{(0.05/2,10-1)} = t_{(0.025,9)} = 2.262$

Critical region: $t > 2.262$ *and* $t < -2.262$

Calculation: $t = \dfrac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \dfrac{2.2 - 0}{2.86/\sqrt{10}} = \dfrac{2.2}{0.904} = 2.433$

Conclusion:



As calculated value t(=2.433) lies in CF may reject $H_0: \mu_d = 0$. It means length hindleg and foreleg of deer is different

8.15.4 Single population proportion:

Hypothesis: $H_0: \pi = \pi_0 / \pi \leq \pi_0 / \pi \geq \pi_0$

$H_1: \pi \neq \pi_0 / \pi > \pi_0 / \pi < \pi_0$

Test statistic: $Z = \dfrac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$

Critical value/Critical Region:

Hypothesis	Critical value	Critical Region
$H_0: \pi = \pi_0$ $H_1: \pi \neq \pi_0$	$Z_{\alpha/2}$	$Z > Z_{\alpha/2}$ $Z < -Z_{\alpha/2}$
$H_0: \pi \leq \pi_0$ $H_1: \pi > \pi_0$	Z_α	$Z > Z_\alpha$
$H_0: \pi \geq \pi_0$ $H_1: \pi < \pi_0$	Z_α	$Z < -Z_\alpha$

Example 8.4(a): A botanist has produced a new variety of hybrid wheat that is better able to withstand drought than other varieties. He knows that 80% of the seeds from the parent plants germinate. He claims the hybrid has the same germination rate. To test this claim, 400 seeds from the hybrid plant are tested and 312 germinated. Test the botanist claim at a 5% level of significance.

Solution: $p=\frac{X}{n}=\frac{312}{400}=0.78;\ \pi=0.80$

$H_0:\pi=0.80\ \ H_1:\pi<0.80$

$\alpha=0.05$

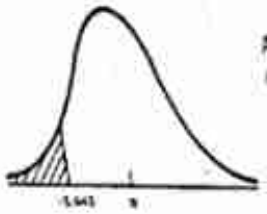
Test statistic: $Z=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

Critical value: $Z_\alpha=Z_{0.05}=1.645$

Critical region: $Z<-1.645$

Calculation: $Z=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}=\frac{0.78-0.80}{\sqrt{\frac{0.80(1-0.80)}{400}}}=\frac{-0.02}{0.02}=-1.00$

Conclusion:



As Calculated value of Z (= - 1.00) does not lie in CR so we may accept $H_0:\pi=0.80$. It means new hybrid variety is better.

Example 8.4(b): Out of 20 seeds taken at random from a population of that seed 6 are male. Is male seed percentage is different than 35% .

Solution: $p=\frac{X}{n}=\frac{6}{20}=0.30;\ \pi=0.35$

$H_0:\pi=0.35\ \ H_1:\pi\neq0.35$

$\alpha=0.05$

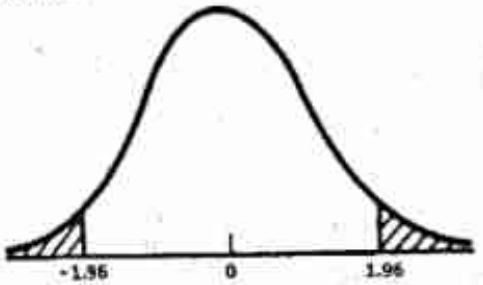
Test statistic: $Z=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

Critical value: $Z_{\alpha/2}=Z_{0.05/2}=Z_{0.025}=1.96$

Critical region: $Z<-1.93$ *and* $Z>1.96$

Calculation: $Z=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}=\frac{0.30-0.35}{\sqrt{\frac{0.35(1-0.35)}{20}}}=\frac{-0.05}{0.107}=-0.462$

Conclusion:



As Calculated value of Z (= -0.462) does not li in CR so we may accept $H_0:\pi=0.35$. It means percentage of male see is 35%.

8.15.5 Difference between two population proportions:

Hypothesis: $H_0:\pi_1=\pi_2\ /\ \pi_1\leq\pi_2\ /\ \pi_1\geq\pi_2$

$H_1:\pi_1\neq\pi_2\ /\ \pi_1>\pi_2\ /\ \pi_1<\pi_2$

Test statistic: $Z=\frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}$

Critical value/Critical Region:

Hypothesis	Z-test	CR
$H_0:\pi_1=\pi_2;\ H_1:\pi_1\neq\pi_2$	$Z_{\alpha/2}$	$Z>Z_{\alpha/2}$ *and* $Z<-Z_{\alpha/2}$
$H_0:\pi_1\leq\pi_2;\ H_1:\pi_1>\pi_2$	Z_α	$Z>Z_\alpha$
$H_0:\pi_1\geq\pi_2;\ H_1:\pi_1<\pi_2$	Z_α	$Z<-Z_\alpha$

Example 8.5(a): A sample of 200 smokers and 95 non-smokers were asked if they've had a night sleep last night. 141 of the smokers and 76 non-smokers responded that they did. At the .05 level of significance, does the data suggest that non-smokers have more restful nights?

Solution: $p_1 = \frac{X_1}{n_1} = \frac{141}{200} = 0.705,\ p_2 = \frac{X_2}{n_2} = \frac{76}{95} = 0.80$

$H_0: \pi_1 = \pi_2$ $H_1: \pi_1 < \pi_2$

$\alpha = 0.05$

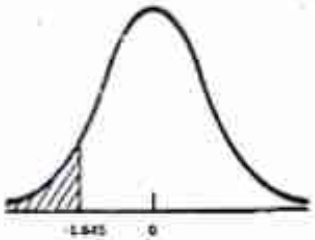
Test statistic: $Z = \dfrac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$

Critical value: $Z_\alpha = Z_{0.05} = 1.645$

Critical region: $Z < -1.645$

Calculation: $Z = \dfrac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} = \dfrac{0.705 - 0.80 - 0}{\sqrt{\dfrac{0.705(1-0.705)}{200} + \dfrac{0.80(1-0.80)}{95}}} = \dfrac{-0.095}{0.052} = -1.827$

Conclusion:



As Calculated value of Z (= -1.827) lies in CR so we may reject $H_0: \pi_1 = \pi_2$. It means nonsmokers sleep well.

Example 8.5(b): Out of 120 boys and 110 girls 22 and 36 are lefthanded respectively. Is there any difference between percentages of left-handedness in boys and girls.

Solution: $p_1 = \frac{X_1}{n_1} = \frac{22}{120} = 0.18,\ p_2 = \frac{X_2}{n_2} = \frac{36}{110} = 0.33$

$H_0: \pi_1 = \pi_2$ $H_1: \pi_1 \neq \pi_2$

$\alpha = 0.05$

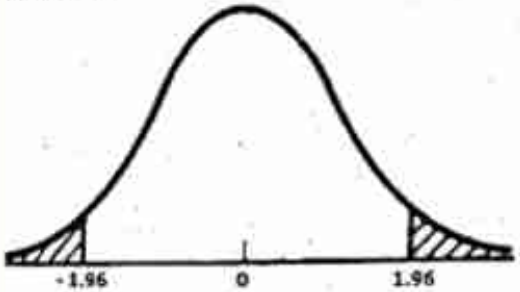
Test statistic: $Z = \dfrac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$

Critical value: $Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$

Critical region: $Z < -1.96$ *and* $Z > 1.96$

Calculation: $Z = \dfrac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} = \dfrac{0.18 - 0.33 - 0}{\sqrt{\dfrac{0.18(1-0.18)}{120} + \dfrac{0.33(1-0.33)}{110}}} = \dfrac{-0.15}{0.057} = -2.63$

Conclusion:



As Calculated value of Z(= -2.63) lies in CR so we may not accept $H_0: \pi_1 = \pi_2$. It means percentage of lefthanded boys and girls are different.

Multiple Choice Questions

1) The Hypothesis $\mu \leq 200$ is:
a) Simple b) Composite c) Alternative d) Researcher

2) The region which leads to the rejection of H_0 is:
a) Critical region b) Rejection region c) Both (a) and (b) d) None of these

3) For one tailed test when $\alpha = 0.05$, the table value of Z is:
a) 1.645 b)-1.96 c) 2.33 d)2.58

4) For two tailed test when $\alpha = 0.05$, the table value of Z is:
a) 2.33 (b) 1.96 c) 2.58 d) 2.45

5) An alternative hypothesis is also called:
a) True hypothesis b) Null hypothesis c) False hypothesis d)Research hypothesis

6) Statistical inference is divided into two major types given below:
a) Point estimation and hypothesis testing
b) Interval estimation and hypothesis testing
c) hypothesis testing and estimation d) point estimation and interval estimation

7) The probability of accepting the true null hypothesis is called:
a) Level of significance b) Level of confidence c) β d) Power

8) If $H_a: \mu \neq \mu_0$ then test is called:
a) One tailed test b) Two tailed c) Left tailed test d) Right tailed test

9) Acceptance of a true null hypothesis is called:
a) Type-I error b) Type-II error c) wrong decision d) correct decision

10) A teacher passed a bad students is an example of:
a) Right decision b) Type-I error c) Type-II error d) Sample error

11) Hypothesis testing and estimation are types of
a) Estimation b) testing of hypothesis c) statistical inference d) sampling

12) The hypothesis which is tested for possible purpose of rejection under the assumption that it is true is:
a)Simple b) Composite c) Null d) alternative

13) Two tailed test is used if:
a) $H_a: \mu < \mu_0$ b) $H_a: \mu > \mu_0$ c) $H_a: \mu \neq \mu_0$ d) $H_a: \mu = \mu_0$

14) Alternative hypothesis against the null hypothesis $H_0: \mu \leq \mu_0$
a) $H_a: \mu < \mu_0$ b) $H_a: \mu > \mu_0$ c) $H_a: \mu \neq \mu_0$ d) $H_a: \mu = \mu_0$

15) It is claimed that an automobile is driven on the average more than 20,000 km/years. Which is the following is true:
a) $H_0: \mu \leq 20{,}000$ $H_1: \mu > 20{,}000$ b) $H_0: \mu < 20{,}000$ $H_1: \mu > 20{,}000$
c) $H_0: \mu > 20{,}000$ $H_1: \mu < 20{,}000$ d) $H_0: \mu \leq 20{,}000$ $H_1: \mu < 20{,}000$

16) Critical region is a region of:
a) Rejection b) Acceptance c) Indecision d) critical value

17) The null hypothesis (H_0) is:
a) Always true b) Always false c) May be true or false d) Always wrong

18) One tailed test is used if:
a) $H_a: \pi < 0.9$ b) $H_a: \pi > 0.9$ c) $H_a: \pi \neq 0.9$ d) both (a) and (b)

19) One sided and two-sided critical regions are based on:
a) Alternative hypothesis b) Sample size c) Null hypothesis d) All of these

20) $H_1: \pi \neq 0.9$, we will use:
a) One sided b) Two sided c) Right sided d) Left sided

21) If $H_1: \pi \neq 0.9$, than test will be:
a) Two sided test b) Left-side test c) Right side test d) two tailed test

22) If the population standard deviation is unknown, than the test to be used for testing of hypothesis about single population mean:
a) t-test b) Z-test c) Chi-square test d) F test

23) In t-test we use:
a) Population variance b) sample variance
c) population SD d) Sample proportion

24) For testing of hypothesis of equality of two population means and known population standard deviation, we used:
a) Z-test b) t-test c) χ^2-test d) F test

25) The df for unpaired samples in t-test with $n_1 = 11$, $n_2 = 10$ are:
a) 19 b) 21 c) 22 d) 20

26) When population S.D is known, than test is:
a) Z-test b) t-test c) χ^2 test d) F-test

27) The test used for paired observation is:
a) Z b) t c) χ^2 d) F

28) The degree of freedom to test the mean for paired observation is:
a) n_1+n_2-1 b) n_1+n_2-2 c) $(n_1-1)(n_2-1)$ d) $n-1$

29) If null hypothesis is H_0; $\pi \geq 0.60$, than we use ---------- test:
a) Z b) t c) χ^2 d) F

30) Pooled standard deviation is used to the test the population means, when populations are:
a) Dependent b) Independent c) paired d) Correlated

31) A test statistic is a ratio between the sampling error and:
a) Bias b) Sample error c) Standard error d) mean

32) Probability of rejecting true null hypothesis is
a) α b) $1-\alpha$ c) β d) $1-\beta$

33) Probability of not rejecting false null hypothesis is
a) α b) $1-\alpha$ c) β d) $1-\beta$

34) For hypotheses $H_0: \mu = 100$ against $H_1: \mu < 100$ critical region is on
a) Left side b) right side c) on both sides d) none

35) Difference between Z test and t test:
a) Z test is two tailed and t test is one tailed
b) Z test is one tailed and t test is two tailed
c) Z test is used for testing of hypothesis about single mean
d) Z requires population variance be known

36) Level of significance
a) α b) $1-\alpha$ c) β d) $1-\beta$

37) Level of confidence
a) α b) $1-\alpha$ c) β d) $1-\beta$

38) Probability of type II error
a) α b) $1-\alpha$ c) β d) $1-\beta$

39) Power of the test
a) α b) $1-\alpha$ c) β d) $1-\beta$

40) Probability of type I error
a) α b) $1-\alpha$ c) β d) $1-\beta$

Key

Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans
1	b	2	c	3	a	4	b	5	d	6	c	7	b
8	b	9	d	10	c	11	c	12	c	13	c	14	b
15	a	16	a	17	c	18	d	19	a	20	b	21	c
22	a	23	b	24	a	25	a	26	a	27	b	28	d
29	a	30	b	31	c	32	a	33	c	34	a	35	d
36	a	37	b	38	c	39	d	40	a				

Exercise

Q No. 8.1: (a) Define the terms, statistical hypothesis, testing of hypothesis, test statistic, level of significance, critical value and analysis of variance.

(b) Differentiate between the followings

(i) Null and alternative hypotheses.

(ii) Simple and composite hypotheses.

(iii) Type I and type II error.

(iv) One sided and two sided test.

(v) Acceptance and rejection region.

Q No. 8.2: Solve the following questions

Sr. No	Sample statistic			hypotheses	α
	Size	Mean	SD		
1	$n = 36$	42.6	5	$H_0: \mu = 45,\ H_1: \mu \neq 45$	2%
2	$n = 10$	1.49	0.02	$H_0: \mu = 1.5,\ H_1: \mu < 1.5$	5%
3	$n = 8$	1200	125	$H_0: \mu = 1150,\ H_1: \mu > 1150$	4%
		Proportion/No. of Success			
4	$n = 400$	$p = 0.89$		$H_0: \pi = 0.95,\ H_1: \pi < 0.95$	2%
5	$n = 100$	$p = 0.38$		$H_0: \pi = 0.50,\ H_1: \pi \neq 0.50$	10%
6	$n = 500$	$p = 0.056$		$H_0: \pi \leq 0.04,\ H_1: \pi > 0.04$	5%
7	$n_1 = 25$ $n_2 = 36$	$\bar{X}_1 = 16$ $\bar{X}_2 = 15$	$s_1 = 0.65$ $s_2 = 0.35$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	5%
8	$n_1 = 8$ $n_2 = 10$	$\bar{X}_1 = 27.5$ $\bar{X}_2 = 25.8$	$s_1 = 3.50$ $s_2 = 2.95$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	2%
9	$n_1 = 50$ $n_2 = 40$	$\bar{X}_1 = 125.5$ $\bar{X}_2 = 133.0$	$s_1 = 21.5$ $s_2 = 23.8$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	1%
10	$n_1 = 12$ $n_2 = 15$	$\bar{X}_1 = 100$ $\bar{X}_2 = 250$	$s_1 = 400$ $s_2 = 300$	$H_0: \mu_2 - \mu_1 = 100$ $H_1: \mu_2 - \mu_1 > 100$	2%

11	$n_1 = 20$ $n_2 = 15$	$p_1 = 0.28$ $p_2 = 0.23$		$H_0: \pi_1 = \pi_2$ $H_1: \pi_1 > \pi_2$	5%
12	$n_1 = 15$ $n_2 = 20$	$\bar{X}_1 = 75$ $\bar{X}_2 = 80$		$H_0: \pi_1 = \pi_2$ $H_1: \pi_1 \neq \pi_2$	1%
13	$n_1 = 80$ $n_2 = 60$	$\bar{X}_1 = 38$ $\bar{X}_2 = 35$		$H_0: \pi_1 = \pi_2$ $H_1: \pi_1 < \pi_2$	5%
14	$n = 10$	$\bar{d} = -2.5$ $s_d = 1.15$		$H_0: \mu_d = 0$ $H_1: \mu_d < 0$	1%
15	$n = 7$	$\bar{d} = 3.0$ $s_d = 1.50$		$H_0: \mu_d = 0$ $H_1: \mu_d > 0$	1%
16	$n = 6$	$\bar{d} = 1.5$ $s_d = 0.75$		$H_0: \mu_d = 0$ $H_1: \mu_d \neq 0$	5%
17	$n = 15$	$\bar{X} = 11.5$ $\sigma = 3.75$		$H_0: \mu = 10$ $H_1: \mu \neq 10$	5%
18	$n = 100$	$\bar{X} = 50$ $\sigma^2 = 49$		$H_0: \mu = 55$ $H_1: \mu < 55$	2%
19	$n_1 = 20$ $n_2 = 25$	$\bar{X}_1 = 70$ $\bar{X}_2 = 65$	$\sigma_1 = 4$ $\sigma_2 = 3$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	2%
20	$n_1 = 55$ $n_2 = 50$	$\bar{X}_1 = 7$ $\bar{X}_2 = 6$	$\sigma_1^2 = 4$ $\sigma_2^2 = 9$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \quad \mu_2$	1%

Q No. 8.3: Ambulance service 1122 authority in Punjab claims that average speed of emergency response is less than 20 minutes. There are no past records; so the actual standard deviation of such response times cannot be determined. The response times for a selected 10 days are 16, 20, 18, 19, 15, 18, 22, 25, 14, and 13 minutes. Test claim of the authority.

Q No. 8.4: Consider the analysis of a soil sample for arsenic content. We might want to know whether the experimental value exceeds the maximum allowable concentration

(MAC). Suppose a set of 7 replicate measurements on a soil sample returned a mean concentration of 2.5 ppm with standard deviation $s = 0.5$ ppm, and that the MAC was 2.0 ppm. Test the hypothesis.

Q No. 8.5: Data about fluric acid in corn seedling grown in two different conditions are as follows. Test for equality of means for both conditions.

Measurement	Condition 1	Condition 2
Sample	10	15
Average	98	116
SD	12	9

Q No. 8.6: A pain reliever currently being used in a hospital is known to bring relief to patient in a mean time of 3.5 minutes. To compare a new pain reliever with the one currently being used, the new drug is administered to a random sample of 50 patients. The mean time to relief for the sample of patients is 2.8 minutes and standard deviation is 1.14 minutes. Do the data provide sufficient evidence to conclude that the new dru was effective in reducing the mean time until a patient receives relief from pain? Test using $\alpha = 0.10$.

Q No. 8.7: Organic chemists often purify organic compounds by fractional crystallization. A laboratory technician desired to prepare and purify several samples, each of 4.85 grams aniline. He claims that his procedure theoretically will produce 3.43 grams of acetanilide. In a sample of 16, the mean was found to be 3.50 with a standard deviation of 0.55. Test the claim made by the chemist.

Q No. 8.8: Suppose the normal level of hemoglobin (Hb) in children is 13.2 g/dl. A study on a random sample of 10 children with chronic diarrhea revealed that the mean is 12.6 g/dl and standard deviation is 1.77 g/dl. The objective is to find out whether children with chronic diarrhea, on average, have less Hb level or not. What are the null and the alternative hypotheses for this study? And test the hypothesis.

Q No. 8.9: The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a sample mean of 134 calories with a standard deviation of 17 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips?

Q No. 8.10: According to advertisement, a strain of soya beans planted on soil prepared with a specified fertilizer treatment has a mean yield of 500 bushels per acre fifty farmers plant the soya bean. The mean and variance for the sample of 50 farms are 485 and 10045. Does the data provide sufficient of evidence to indicate that the mean yield for soya beans is different from that advertised?

Q No. 8.11: The following data are times (days) for seven cockroach eggs to hatch at 30° laboratory temperature and for eight eggs to hatch at 10° temperature. Is there any difference between two temperature levels?

At 30° C	40	38	32	37	39	41	35	
At 10° C	36	45	32	52	59	41	48	55

Q No. 8.12: The following data are of germination time for pine seeds planted in a greenhouse and outside. Test the hypothesis that there is no significant difference between both methods

Greenhouse	69.3	75.5	81.0	74.7	72.3	78.7	76.4		
Outside	69.5	64.6	74.0	84.8	76.0	93.9	81.2	73.4	66.0

Q No. 8.13: It is proposed that animals with a northerly distribution have shorter appendages than animals from a southern distribution. Test an appropriate hypothesis. Using the following wing length data for birds in mm

Northern	120	113	125	118	116	114	119	
Southern	116	117	121	114	116	118	123	120

Q No. 8.14: If $\bar{X}_1 = 334.6g$, $\bar{X}_2 = 349.8g$, $\sum(X_1 - \bar{X}_1)^2 = 336.64g^2$, $\sum(X_2 - \bar{X}_2)^2 = 286.78g^2$, $n_1 = 19$ *and* $n_2 = 24$, test the hypothesis that the mean weight of second population is more than 10 g greater than the mean weight of first population.

Q No. 8.15: A claim is made that boys start walking at a younger age than girls. A random sample of 18 girls started walking at a mean age of 12.5 months with a standard deviation of 0.8 months. A sample of 12 boys had a mean of 12.1 and a standard deviation of 0.7. Test the hypothesis at a 1% significance level.

Q No. 8.16: It is known that the white shark grows to a mean length of 21 feet. A marine biologist believes that the white shark of the Bermuda Coast grow much longer

due to unusual feeding habits. To test this claim, three full grown great white sharks are captured of the Bermuda Coast, their length are 24, 20 and 22 feet. Do the data provide sufficient evidence to support the marine biologists claim? Use $\alpha = 0.05$

Q No. 8.17: Two types of fertilizers are used in the cultivation of Cabbage. They grow Cabbage in two different fields. The sample mean and variance of weights, of 25 cabbage grown with fertilizer 1 are 44.1g and 36g, and from sample of 12 cabbages grown with fertilizer 2 is 31.07g and the variance 44g. Is there any significant difference between both types of fertilizers?

Q No. 8.18: Using following data, the null hypothesis that male and female have same mean serum cholesterol concentrations.

Serum Cholesterol mg/100ml

Male	220.1	218.6	229.6	228.8	222.0	224.1	265.5
Female	223.4	221.5	230.2	224.3	223.8	230.8	227.9

Q No. 8.19: 150 wheat ear heads of variety A gave an average 65 grains/ear head with a SD of 3 and 100 ear heads of variety B gave an average of 75 grains/ear head with a SD of 5. Do you conclude that variety B has more grains/ear head, at 0.05 level of significance.

Q No. 8.20: The following data is about weight change of humans, tabulated after administration of a drug proposed to result in weight loss. Each weight change (in Kg) is the weight after minus the weight before drug administration: 0.2, - 0.5, - 1.3, - 1.6, - 0.7, 0.4, - 0.1, 0.0, - 0.6, - 1.1, - 1.2, - 0.8. Test the hypothesis the there is no effect of using the drugs.

Q No. 8.21: Is the proportion of babies born male different from 0.50? In a sample of 200 babies, 96 were male.

Q No. 8.22: A botanist has produced a new variety of hybrid wheat that is better able to withstand drought than other varieties. He knows that 80% of the seeds from the parent plants germinate. He claims the hybrid has the same germination rate. To test this claim, 400 seeds from the hybrid plant are tested and 312 germinated. Test the botanist claim at a 5% level of significance.

Q 8.23: Level of a particular hormone in the blood in six patients before and after they begin taking a hormone treatment program. Results are as follows:

Patient	Before	After
1	0.21	0.19
2	0.16	0.17
3	0.22	0.22
4	0.25	0.20
5	0.17	0.16
6	0.23	0.21

Using the .05 significance level, was there a significant change in the level of this hormone?

Q 8.24: Study was done of personality characteristics of 100 students who were tested in their fall and spring semester of their first year of BS. The researchers reported the results:

	Fall semester	Spring semester	Difference
Mean	16.82	15.32	1.50
SD	4.21	3.84	1.85

Test for significance of difference, when
(a) In both semesters students are different.
(b) In both semesters students are same.

Q 8.25: Do students at two different universities differ in how sociable they are? Twenty-five students were randomly selected from each of two universities in a region and were asked to report on the amount of time they spent socializing each day with other students. The result for University 1 was a mean of 5 hours, with variance of 4 hours; for University 2, mean = 4, Standard deviation = 1.5; what should you conclude? Use the 0.05 level.

Q 8.26: Typing speed of 10 participants in two different conditions (social and alone) are given below test the hypothesis that time for typing alone is less than typing in company.

Social	9	14	13	15	10	17	13	18	12	7
Alone	8	12	14	11	10	18	12	15	13	5

Q 8.27: Marks of ten students in "Statistics" in first year class and second year class are given below, test students have improved their study in second year.

Student	A	B	C	D	E	F	G	H	I	J
First year	44	54	81	76	50	47	65	60	42	79
Second year	46	55	78	79	45	47	66	66	48	78

Q 8.28: A sample of 15 students of 10th class of school "A" have an average IQ of 107.3; whilst a sample of 12 students from school "B" have an average IQ of 104.1, the true variances of the IQ's for the children at the two schools are 39 and 58 respectively; do you consider that there is strong evidence that students of "A" school have higher mean IQ than those "B" school?

Solution

Q. No	Hypothesis	CR	Calculations	Test statistic	Conclusion
8.3	$H_0: \mu \geq 20$ $H_1: \mu < 20$	$t < -1.833$	$\bar{X} = 18$ $s = 3.71$	$t = -1.709$	Accept H_0
8.4	$H_0: \mu \leq 2.0$ $H_1: \mu > 2.0$	$t > 1.943$	$\bar{X} = 2.5$ $s = 0.5$	$t = 2.645$	Reject H_0
8.5	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$t > 2.069$ $t < -2.069$	$S_p = 10.74$	$t = -4.29$	Reject H_0
8.6	$H_0: \mu = 3.5$ $H_1: \mu < 3.5$	$t < -1.299$	$\bar{X} = 2.8$ $s = 1.14$	$t = -4.348$	Reject H_0
8.7	$H_0: \mu = 3.43$ $H_1: \mu > 3.43$	$t > 1.753$	$\bar{X} = 3.50$ $s = 0.55$	$t = 0.509$	Accept H_0
8.8	$H_0: \mu = 13.2$ $H_1: \mu < 13.2$	$t < -1.833$	$\bar{X} = 12.6$ $s = 1.77$	$t = -1.071$	Accept H_0
8.9	$H_0: \mu = 130$ $H_1: \mu \neq 130$	$t > 2.032$ $t < -2.032$	$\bar{X} = 134$ $s = 17$	$t = 1.392$	Accept H_0
8.10	$H_0: \mu = 500$ $H_1: \mu \neq 500$	$t > 2.010$ $t < -2.010$	$\bar{X} = 485$ $s = 100.2$	$t = -1.058$	Accept H_0
8.11	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$t > 2.160$ $t < -2.160$	$\bar{X}_1 = 37.43$ $\bar{X}_2 = 46$ $s_p = 7.18$	$t = -2.31$	Reject H_0
8.12	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$t > 2.145$ $t < -2.145$	$\bar{X}_1 = 75.41$ $\bar{X}_2 = 75.93$ $s_p = 7.55$	$t = -0.136$	Accept H_0
8.13	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	$t < -1.771$	$\bar{X}_1 = 117.86$ $\bar{X}_2 = 118.13$ $s_p = 3.53$	$t = -0.14$	Accept H_0
8.14	$H_0: \mu_2 - \mu_1 \leq 10$ $H_1: \mu_2 - \mu_1 > 10$	$t > 1.683$	$s_p = 3.9$	$t = 4.344$	Reject H_0
8.15	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	$t > 2.467$	$s_p = 0.79$	$t = 1.36$	Accept H_0
8.16	$H_0: \mu = 21$ $H_1: \mu > 21$	$t > 2.920$	$s = 2$	$t = 0.866$	Accept H_0
8.17	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$t > 2.030$ $t < -2.030$	$s_p = 6.39$	$t = 5.81$	Reject H_0

8.18	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$t > 2.179$ $t < -2.179$	$s_p = 11.79$ $\bar{X}_1 = 229.81$ $\bar{X}_2 = 225.99$	$t = 0.60$	Accept H_0
8.19	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	$Z < -1.645$		$Z = -17.953$	Reject H_0
8.20	$H_0: \mu_d = 0$ $H_1: \mu_d < 0$	$t < -2.200$	$\bar{d} = -0.608$ $s_d = 0.633$	$t = -3.33$	Reject H_0
8.21	$H_0: \pi = 0.50$ $H_1: \pi \neq 0.50$	$Z > 1.96$ $Z < -1.96$		$Z = -0.57$	Accept H_0
8.22	$H_0: \pi = 0.80$ $H_1: \pi < 0.80$	$Z < -1.645$		$Z = -1.00$	Accept H_0
8.23	$H_0: \mu_d = 0$ $H_1: \mu_d \neq 0$	$t > 2.571$ $t < -2.571$	$\bar{d} = 0.015$ $s_d = 0.021$	$t = 1.749$	Accept H_0
8.24 (a)	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$t > 1.972$ $t < -1.972$		$t = 2.618$	Reject H_0
(b)	$H_0: \mu_d = 0$ $H_1: \mu_d \neq 0$	$t > 1.984$ $t < -1.984$	$\bar{d} = 1.05$ $s_d = 1.85$	$t = 8.108$	Reject H_0
8.25	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$t > 2.011$ $t < -2.011$	$s_p = 1.804$	$t = 1.96$	Accept H_0
8.26	$H_0: \mu_d = 0$ $H_1: \mu_d > 0$	$t > 1.833$	$\bar{d} = 1$ $s_d = 1.76$	$t = 1.797$	Accept H_0
8.27	$H_0: \mu_d = 0$ $H_1: \mu_d > 0$	$t > 1.833$	$\bar{d} = 1$ $s_d = 3.53$	$t = 0.896$	Accept H_0
8.28	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	$Z > 1.645$		$Z = 1.174$	Accept H_0

Chapter 9
TESTING OF HYPOTHESIS

(Chi square)

Statistics is the back bone of research

9.1 Attribute
A qualitative characteristic of an individual is known as an attribute.

9.2 Contingency table
A contingency table is a table that consists of n paired observations related to different categories. Each cell gives the number of observations that fall into the category defined by its row and column heading.

9.3 Association
Relationship between attributes is called association.

9.4 Testing of hypothesis
Testing of hypothesis is defined as the formal procedures used by a researcher to accept or reject a statistical hypothesis.

9.4.1 Testing of hypothesis about independence of attributes:

Hypothesis: H_0: *Attributes are independent*

H_1: *Attributes are dependent*

Test statistic: $\chi^2 = \sum_i \left(\frac{(O_i - E_i)^2}{E_i} \right)$

Critical value/Critical Region:

CV	CR
$\chi^2_{(\alpha,\upsilon)}$; $\upsilon = (r-1)(c-1)$ r = *No. of rows* and c = *No. of columns*	$\chi^2 > \chi^2_{(\alpha,\upsilon)}$

Example 9.1: The following data relate to the number of children classified according to the type of feed and nature of teeth.

Type of teeth	Normal teeth	Defective teeth
Breast	14	19
Bottle	13	35

Use chi square test and draw inference from this data.

Solution: H_0: *Attributes are independent*

H_1: *Attributes are dependent*

Test statistic: $\chi^2 = \sum_i \left(\frac{(O_i - E_i)^2}{E_i} \right)$

$\alpha = 0.05$

Critical value: $\chi^2_{(\alpha,\upsilon)} = \chi^2_{(0.05,1)} = 3.841$

Critical region: $\chi^2 > 3.841$

Calculation:

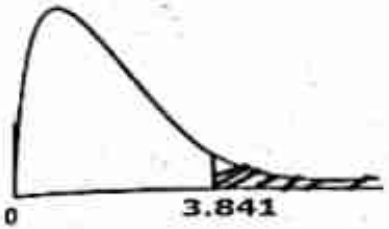
Type of teeth	Normal teeth	Defective teeth	Total
Breast	14	19	33
Bottle	13	35	48
Total	27	54	81

Expected frequency table:

Type of teeth	Normal teeth	Defective teeth	Total
Breast	11	22	33
Bottle	16	32	48
Total	27	54	81

O_i	E_i	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
14	11	3	0.8181
19	22	-3	0.4091
13	16	-3	0.5625
35	32	3	0.2813
Σ			2.071

Conclusion:



As calculated value of χ^2 (= 2.071) is less than critical value (3.841) so we may accept H_0: Attributes are independent

9.4.2 Testing of hypothesis about Several proportions:

Hypothesis: $H_0: \pi_1 = \pi_{10}, \pi_2 = \pi_{20}, \ldots, \pi_k = \pi_{k0}$

$H_1: \pi_i \neq \pi_{i0}$

Test statistic: $\chi^2 = \sum_i \left(\frac{(O_i - E_i)^2}{E_i} \right)$

Critical value/Critical Region:

CV	CR
$\chi^2_{(\alpha,\upsilon)}$; $\upsilon = k-1$	$\chi^2 > \chi^2_{(\alpha,\upsilon)}$

Example 9.2: According to genetic theory the numbers of color strains red, yellow, blue and white in a certain flower appear in the ratio 4:12:5:4. For 800 plants, the results were as follows:

Color	Red	Yellow	Blue	White
Number of plants	110	410	150	130

Test the theory.

Solution: $H_0: \pi_1 = \frac{4}{25}, \pi_2 = \frac{12}{25}, \pi_3 = \frac{5}{25}, \pi_4 = \frac{4}{25}$

$H_1: \pi_i \neq \pi_{i0}$

$\alpha = 0.05$

Test statistic: $\chi^2 = \sum_i \left(\frac{(O_i - E_i)^2}{E_i} \right)$

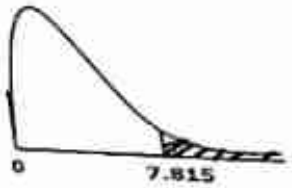
Critical value: $\chi^2_{(\alpha, v)} = \chi^2_{(0.05,3)} = 7.815$

Critical region: $\chi^2 > 7.815$

Calculation:

O_i	$E_i = np_i$	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
110	$800 \times \frac{4}{25} = 128$	-18	2.5313
410	384	26	1.7601
150	160	-10	0.6250
130	128	2	0.0313
800			**4.9477**

Conclusion:



As calculated value of χ^2 (=4.9477) does not lie in CR so we may accept $H_0: \pi_1 = \frac{4}{25}, \pi_2 = \frac{12}{25}, \pi_3 = \frac{5}{25}, \pi_4 = \frac{4}{25}$

Multiple Choice Questions

1) For testing of hypothesis of equality of more than two population means and known population standard deviation, we used:

a) Z-test b) t-test c) χ^2-test d) *F* test

2) The *df* for chi squared test with $n = 11$ is:

a) 11 b) 10 c) 8 d) degree of freedom does not apply

3) To test several population proportions:

a) Z b) t c) χ^2 d) *F*

4) If null hypothesis is H0; $\pi = \pi_0$; than we use ---------- test:

a) *Z* b) *t* c) χ^2 d) *F*

5) If $r = 3$ and $c = 4$ in a contingency table, then df is:

a) 12 b) 9 c) 8 d) 6

6) Null hypothesis for independence of attributes:

a) Attributes are dependent b) Attributes are not associated
c) Attributes are associated d) Attributes are correlated

7) If calculated value of chi square test is more than critical value :

a) Do not reject null hypothesis b) Accept null hypothesis
c) not accept null hypothesis d) not accept alternative hypothesis

8) If calculated value of chi square test is less than critical value :

a) Do not accept null hypothesis b) Accept null hypothesis
c) not accept null hypothesis d) accept alternative hypothesis

9) Which of the following values cannot be of chi square:

a) 1 b) -1 c) 0.5 d) 100

10) Chi square test cannot be used:

a) To compare two population variances
b) To compare several population proportions
c) For independence of attribute d) To make inference about a population variance

11) Must be true for expected value in a chi square test:

a) ≥ 2 b) ≥ 5 c) ≥ 10 d) ≥ 15

12) Is gender a factor in a person's favorite color? Which test should be used: a) F b) t c) Z d) χ^2

13) The level of significance:

a) Probability of Type II error b) Probability of Type I error
c) Confidence coefficient d) same as the p-value

14) Which one is incorrect for chi square test:

a) It is used for goodness of fit b) It has a degree of freedom
c) It is negatively skewed d) Its calculated value is always positive

Key

Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans
1	d	2	b	3	c	4	c	5	d	6	b	7	c
8	b	9	b	10	a	11	b	12	d	13	b	14	c

Exercise

Q No. 9.1: Define the terms, contingency table, Attribute and association.

Q No. 9.2: Is the proportion of babies born male different from 0.50? In a sample of 200 babies, 96 were male.

Q No. 9.3: The theory predicts the proposition of beans in the four groups A, B, C and D should be 9:3:3:1. In an experiment among 1600 beans, the numbers in four groups were 882, 313, 287 and 118. Does the experimental result support the theory?

Q No. 9.4: According to a genetics theory, a crossing of red and white snapdragons should produce offspring that are 25% red, 50% pink, and 25% white. An experiment conducted to test theory produces 30 red, 78 pink and 36 white offspring in 144 crossing. Do the data provide sufficient evidence to contradict the genetics theory?

Q No. 9.5: Five different foods were placed before 120 animals of a certain species. Animals choose the foods as follows.

Food 1	1	2	3	4	5
No. of animals	18	20	26	30	26

Test that animals like all five foods equally.

Q No. 9.6: Out of a group of 320 people exposed to infection, 255 had not been immunized, and of these 95 contracted the disease. Of these who had been immunized, 15 were infected. Does it seem that treatment gave any protection against infection using 1% level of significance?

Q No. 9.7: A recent experiment investigated the relationship between smoking and urinary incontinence. Of the 322 subjects in the study who were incontinent, 113 were smokers, 51 were former smokers, and 158 had never smoked. Of the 284 control subjects who were not incontinent, 68 were smokers, 23 were former smokers, and 193 had never smoked. (a) Create a table displaying this data. (b) What is the expected frequency in each cell? (c) Conduct a significance test to see if there is a relationship between smoking and incontinence. What Chi Square value do you get?

Q 9.8: The following table gives the census data of orchards. Test the hypothesis that the two variables of classification are independent

Classes	Shaded	Unshaded	Total
High yielders	350	205	555
Low yielders	250	195	445
Total	600	400	1000

Q 9.9: Given the following contingency table for hair colour and eye colour, test independence of attributes.

Eye colour	Hair colour		
	Fair	Grey	Brown
Blue	69	49	28
Black	91	56	27
Dark blue	57	34	33

Q No. 9.10: A survey was conducted about behavior of children and using the social websites. 4000 children of age 10 – 15 years old and using social websites for three hours daily were observed. The result is as follows. Test for association between these attributes.

	Used social websites	Not Used social websites	Total
Argued with parents	880	400	1280
Not Argued with parents	1120	1600	2720
Total	2000	2000	4000

Q No. 9.11: Psychologist tends to believe that there is a relationship between aggressiveness and order of birth. To this belief, a Psychologist chooses 500 elementary school students at random and admistered each a test designed to measure the student's aggressiveness. Each student was classified according to one of four, categories. The percentage of students falling in the four categories is shown here.

	Second birth	Other
Aggressive	70	110
Non aggressive	130	190

Q 9.12: from the following data test for independence of attributes.

Relationship with siblings	Father	Mother
Change	43	47
Not change	14	7

Q 9.13: A random sample of 200 retired married men was classified according to education and number of children.

Education	Number of children		
	0 – 1	2 – 3	Over 3
Elementary	14	37	32
Secondary	19	42	17
College	12	17	10

Test the hypothesis, at the 0.05 level of significance, that the size of a family is independent of the level of education attained by the father.

Q 9.14: Number of out of school children for all levels is given below test is there any difference between gender and level

Level	Gender	
	Male	Female
Primary	2.03	3.03
Middle	3.14	3.37
High	2.42	2.55
Higher secondary	3.09	3.20

Figures of data are in million

Pakistan education statistics 2016-17

Solution

Q 9.2: $H_0: p_1 = \frac{1}{2}, p_2 = \frac{1}{2}$ $H_1: p_i \neq p_{i0}$ $CR: \chi^2 > 3.84$

O	96	104
$E = np_i$	100	100
$(O-E)^2/E$	0.16	0.16

$\chi^2 = 0.32$

Conclusion: Accept H_0

Q 9.3: $H_0: p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$ $H_1: p_i \neq p_{i0}$ $CR: \chi^2 > 7.815$

O	882	313	287	118	
$E = np_i$	900	300	300	100	Conclusion: Accept H_0
$(O-E)^2/E$	0.360	0.563	0.563	3.240	$\chi^2 = 4.727$

Q 9.4: $H_0: p_1 = 0.25, p_2 = 0.50, p_3 = 0.25, H_1: p_i \neq p_{i0}$ $CR: \chi^2 > 5.992$

O	30	78	36	144	
$E = np_i$	36	72	36	144	Conclusion: Accept H_0
$(O-E)^2/E$	1.0	0.5	0.0	1.5	$\chi^2 = 1.5$

Q 9.5: $H_0: p_1 = \frac{1}{5}, p_2 = \frac{1}{5}, p_3 = \frac{1}{5}, p_4 = \frac{1}{5}, p_5 = \frac{1}{5}$ $H_1: p_i \neq p_{i0}$ $CR: \chi^2 > 9.488$

O	18	20	26	30	26	120		
$E = np_i$	24	24	24	24	24	120		Conclusion: Accept H_0
$(O-E)^2/E$	1.500	0.667	1.667	1.500	0.167	4.000	$\chi^2 = 4.000$	

Q 9.6: H_0: Attributes are independent H_1: Attributes are dependent $CR: \chi^2 > 3.842$

O	15	50	95	160	320		
$E = np_i$	22	43	88	167	320		Conclusion: Reject H_0
$(O-E)^2/E$	2.227	1.140	0.557	0.293	4.217	$\chi^2 = 4.217$	

Q 9.7: H_0: Attributes are independent H_1: Attributes are dependent $CR: \chi^2 > 5.992$

O	113	51	158	68	23	193	606		
$E = np_i$	96	39	187	85	35	164	606		
$(O-E)^2/E$	3.010	3.692	4.497	3.400	4.114	5.128	23.842	$\chi^2 = 23.842$	Conclusion: Reject H_0

Q 9.8: H_0: Attributes are independent H_1: Attributes are dependent $CR: \chi^2 > 3.842$

O	350	205	250	195	1000		
$E = np_i$	333	222	267	178	1000		
$(O-E)^2/E$	0.868	1.302	1.082	1.624	4.876	$\chi^2 = 4.876$	Conclusion: Reject H_0

Q 9.9: H_0: Attributes are independent H_1: Attributes are dependent $CR: \chi^2 > 9.488$

O	69	49	28	91	56	27	57	34	33	444
$E = np_i$	71	46	29	85	54	34	61	39	25	444
$(O-E)^2/E$	0.056	0.196	0.034	0.424	0.074	1.441	0.262	0.641	2.560	5.689
$\chi^2 = 5.689$	Conclusion: Acceptt H_0									

Q 9.10: H_0: Attributes are independent H_1: Attributes are dependent $CR: \chi^2 > 3.842$

O	880	400	1120	1600	4000		
$E = np_i$	640	640	1360	1360	4000		
$(O-E)^2/E$	90	90	42.353	42.353	264.706	$\chi^2 = 264.706$	Conclusion: Reject H_0

Q 9.11: H_0: Attributes are independent H_1: Attributes are dependen $CR: \chi^2 > 3.842$

O	70	110	130	190	500		
$E = np_i$	72	108	128	192	500		
$(O-E)^2/E$	0.056	0.037	0.031	0.021	0.145	$\chi^2 = 0.145$	Conclusion: Accept H_0

Q 9.12: H_0: Attributes are independent H_1: Attributes are dependent $CR: \chi^2 > 3.842$

O	43	47	14	7	111		
$E = np_i$	46	44	11	10	111		
$(O-E)^2/E$	0.196	0.205	0.818	0.900	2.118	$\chi^2 = 2.118$	Conclusion: Accept H_0

Q 9.13: H_0: Attributes are independent H_1: Attributes are dependent $CR: \chi^2 > 9.488$

O	14	37	32	19	42	17	12	17	10	200
$E = np_i$	19	40	24	18	37	23	9	19	11	200
$(O-E)^2/E$	1.316	0.225	2.667	0.056	0.676	1.565	1.000	0.211	0.091	7.805
$\chi^2 = 7.805$	Conclusion: Accept H_0									

Q 9.14: H_0: Attributes are independent H_1: Attributes are dependent $CR: \chi^2 > 9.488$

O	2.03	3.03	3.14	3.37	2.42	2.55	3.09	3.20	22.83
$E = np_i$	2.37	2.69	3.05	3.46	2.32	2.65	2.94	3.35	22.83
$(O-E)^2/E$	0.049	0.043	0.003	0.002	0.004	0.004	0.008	0.007	0.119
$\chi^2 = 0.119$	Conclusion: Accept H_0								

Chapter 10

REGRESSION AND CORRELATION

Statistics is the grammar of science

Learning Goals:

01) To model the relationship between two variables.
02) To understand the method of least squares estimation.
03) To learn the idea of prediction/forecasting.
04) To calculate reliability of the model.
05) To calculate and interpret the strength of linear relationship between quantitative variables.
06) To understand the strength of linear relationship between qualitative variables measured on ordinal scale.
07) To model the relationship between more than two variables.
08) To compute correlation coefficient between more than two variables.
09) To fit the model other than straight line
10) To find suitable model for a data set.

10. 1 Regression

The dependence of one variable on one or more other variables is called regression. When we study the dependence of a variable on a single independent variable, it is called ***simple regression*** or two variables regression.

When the dependence of a variable on two or more than two variables is studied, it is called ***multiple regression***. Examples of regression are:

- The height of a child depends on the height of the parents.
- The temperature depends on the shining of the sun.
- The production of a crop depends on the quality of seed and fertilizers used, is an example of multiple regression.

10.2 Regression line

A line that summarizes the linear relationship (or linear trend) between the two variables in a linear regression analysis, from the bivariate data collected.

10.3 Linear regression

When the dependence is represented by a straight line equation .Then the regression is called linear, otherwise, it is said to be a curvilinear.

10.4 Simple linear regression model

The regression is called simple if there is dependence of one variable on only one variable and the regression is called, simple linear regression if ordered pairs (X, Y) are clustered around a line. A simple regression model is written as: $Y_i = \alpha + \beta X_i + \varepsilon_i$

Where Y_i = dependent variable X_i = independent variable

α = Y-intercept, β = Slope of regression line it is also called regression coefficient.

ε_i = Error / disturbance term.

10.5 Dependent variable

A variable whose average value is to be estimated is called dependent variable. Examples of dependent variables are:

- If we want to estimate the heights of children on the basis of ages, then height is dependent variable.
- If we want to estimate the yield of crop on the basis of fertilizers used then yield is a dependent variable.
- If we want to estimate the production of fans on the basis of workers, then the production of fans is a dependent variable.

Dependent variable is also called regressand, predictand, response or explained variable.

10.6 Independent variable

A factor or phenomenon that causes or influences another associated factor or phenomenon called an independent variable. Examples are

- If we want to estimate the height of children on the basis of ages, then age is an independent variable.
- If we want to estimate the yield of a crop on the basis of fertilizer used then the fertilizer is independent variable.
- If we want to estimate the production of fans on the basis of number of workers, then number of workers is an independent variable.

Independent variable or nonrandom variable is also called regressor, predictor, regression variable or explanatory variable or controlled variable.

Note: Independent and dependent variables are usually denoted by X and Y respectively. They are named differently in various field of studies, some of them are as under.

Y	Dependent	Predictand	Explained	Effect	Response
X	Independent	Predictor	Explanatory	Cause	Stimulus
Y	Endogenous	Stochastic	Outcome	Regressand	
X	Exogenous	Non-stochastic	Covariate	Regressor	

10.7 Intercept: Regression equation is: $Y_i = \alpha + \beta X_i + \varepsilon_i$

In this equation α is the value of Y when X = 0 is called the Y intercept. It means intercept "α" is that value of Y when there is no association of independent variable.

10.8 Regression coefficient

It is defined as the coefficient of independent variable in the regression equation. It is the slope of regression line. For example in the regression equation $Y_i = \alpha + \beta X_i + \varepsilon_i$ "β" is the slope of regression line. It indicates the change in Y for a unit change in X.

10.9 Justification for Inclusion of Disturbance Term

01) Error term in the model covers the effect of omitted variables in the model.
02) It captures any specification error related to assumed linear functional form.
03) It covers vagueness of theory(theory is incomplete)
04) It covers unavailability of data.
05) It covers the variation in the line due to human behavior which is an important factor in the relationship of economic variables.
06) It covers the error of measurement.

10.10 Assumptions for Regression Model

01) Sample is representative of the population.
02) Response is at ratio or interval scale.
03) ε_i is a random real variable. (the value which ε_i may assume in any one period depends on chance, it may be positive , negative or zero)
04) The mean value of ε_i in any particular period is zero i.e. $E(\varepsilon_i) = 0$
05) The variance of ε_i is constant in each period (in other words for all values of X, the ε_i show the same dispersion round their mean). It is called as homoscedasticity. $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$.
06) The variable ε_i has a normal distribution. i.e. $\varepsilon_i \sim NIID(0, \sigma^2)$
07) The random terms of different observations ε_i and ε_j are independent (the value which the random term assumed in one period does not depend on the value which it assumed in any other period). It is called no autocorrelation. i.e. $E(\varepsilon_i, \varepsilon_j) = 0$

(08) ε_i is independent of explanatory variable(s) i.e. $E(\varepsilon_i X_i) = 0$.
The X's are a set of fixed values in the hypothetical process of repeated sampling which underlies the linear regression model.
10) The explanatory variable(s) are measured without error.
11) The explanatory variables are not perfectly linearly correlated. That is there is no multi-collinearity.
12) The macro variables should be correctly aggregated.
13) The relationship being estimated is identified.
14) The relationship is correctly specified.

10.11 Sample regression line

A sample regression line is an estimate of the line that describes the true, but unknown, linear relationship between the two variables. The equation of the regression line is used to predict (or estimate) the value of the response variable from a given value of the explanatory variable.

Sample regression line is given as: $\hat{Y} = a + bX_i$.

10.11.1 Interpretation

Regression coefficient "β" or "b" is a marginal change. It give average change in response "Y" with one unit increase in independent variable "X".

$b\left(\frac{x}{y}\right)$ is known as point elasticity and $b\left(\frac{\bar{x}}{\bar{y}}\right)$ is known as average elasticity, where elasticity is defined as percentage change in Y (Dependent variable) with 1% increase in X (independent variable)

b: The dependent variable Y changes b unit for 1 unit increase in the independent va..able X.

a: value of Y for $X = 0$.

10.12 Method of obtaining regression lines

Methods to obtain regression lines are:

01) The scatter diagram method
02) The method of least squares

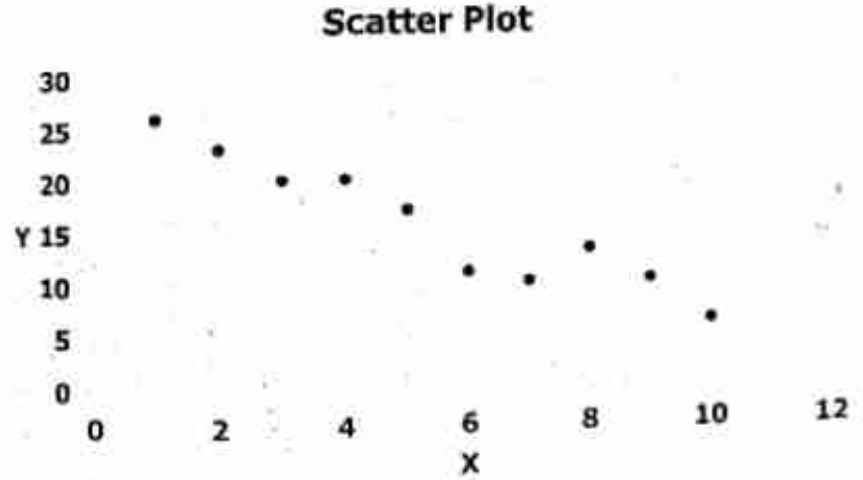
10.12.1 Scatter Diagram

The graphical presentation of a set of "n" pairs of observation by taking values of the independent variable(X) on x-axis and the values of the dependent variable (Y) on y-axis in a graph is called a scatter diagram. Scatter diagram helps in indicating the existing relationship between the two variables. If a relationship between the variables exists, then the points are clustered around a straight line or some curve. Such a line or curve around which the points cluster is called the regression line or regression curve which can be used to estimate the expected value of the random variable y from the values of the non-random variable x.

Example 10.1: Scatter plot for the following data and estimate the model between variables.

X_i	1	2	3	4	5	6	7	8	9	10
Y_i	26	23	20	20	17	11	10	13	10	6

Solution:



As this scatter plot shows that points lies approximately on a line. So linear regression model is appropriate for modelling relationship between these variables.

10.12.2 Method of Ordinary Least Squares

Ordinary least squares (OLS) or linear least square is a method for estimating the unknown parameters in a linear regression model. The goal of OLS is to closely "fit" a function with the data. It does so by minimizing the sum of squared errors from the data.

It is a method which consists of minimizing the sum of the squares of the differences between observed value Y_i and the corresponding estimated values $\hat{Y}_i$ (usually called residuals, ε_i) i.e. It is a method which minimizes $S = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.

10.12.3 Least squares estimates in simple linear regression:

For regression line Y on X: $b_{YX} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$; $a_{YX} = \bar{Y} - b\bar{X}$

10.12.4 Other formulae of b

Line for Y on X ($\hat{Y} = a + bX_i$)

$$b_{yx} = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - (\sum X)^2} \qquad b_{yx} = \frac{\sum XY - \frac{\sum X\sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} \qquad b_{yx} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n(\bar{X})^2}$$

$$b_{yx} = \frac{n\sum D_xD_y - \sum D_x\sum D_y}{n\sum D_x^2 - (\sum D_x)^2} \qquad b_{yx} = \frac{n\sum UV - \sum U\sum V}{n\sum U^2 - (\sum U)^2} \times \frac{k}{h} \qquad b_{yx} = r\frac{S_y}{S_x}$$

Line for X on Y ($\hat{X} = a + bY_i$) $\qquad a_{XY} = \bar{X} - b\bar{Y} \qquad b_{xy} = \frac{n\sum XY - \sum X\sum Y}{n\sum Y^2 - (\sum Y)^2}$

$$b_{xy} = \frac{n\sum XY - \sum X\sum Y}{n\sum Y^2 - (\sum Y)^2} \qquad b_{xy} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n(\bar{Y})^2} \qquad b_{xy} = \frac{n\sum D_xD_y - \sum D_x\sum D_y}{n\sum D_y^2 - (\sum D_y)^2}$$

$$b_{xy} = \frac{n\sum UV - \sum U\sum V}{n\sum V^2 - (\sum V)^2} \times \frac{h}{k} \qquad b_{xy} = r\frac{S_x}{S_y}$$

10.12.5 Regression co-efficient method

Regression equation Y on X is $Y_i - \bar{Y} = b_{yx}(X_i - \bar{X})$ where $b_{yx} = r\frac{S_y}{S_x}$, Similarly regression equation X on Y is $X_i - \bar{X} = b_{xy}(Y_i - \bar{Y})$, where $b_{xy} = r\frac{S_x}{S_y}$

10.13 Properties of regression line by OLS

01) $\sum Y = \sum \hat{Y}$

02) $\sum(Y - \hat{Y}) = 0$

03) $\sum(Y - \hat{Y})^2$ *is least*

04) It always passes through $(\bar{X}, \bar{Y})$.

05) It is the best linear regression line.

06) Regression coefficients are not symmetrical for X and Y. I.e. $b_{yx} \neq b_{xy}$.

07) Regression coefficients are independent of change of origin but not scale.

10.14 Standard error of estimate

The standard deviation which measures the dispersion of ordered pairs (X,Y) points, which lie above and below the estimated regression line is called standard deviation of regression. It is also called standard error of estimate.

It is a measure of the accuracy of predictions made with a regression line. The standard error of estimate is used to determine how well a least squares line equation fits a data set. It is given as $S_{y.x} = \sqrt{\frac{\sum \varepsilon_i^2}{n-2}} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{\sum Y_i^2 - a\sum Y_i - b\sum X_iY_i}{n-2}}$

It is an estimate of badness of the fitting model. It relates to regression line as standard deviation to mean. If lines are drawn parallel to fitted regression line with $\pm\sqrt{S_{y.x}}$ distance, nearly 68% values ordered pairs lies between these two lines.

Example 10.2: Fit a regression equation to predict the yield of crop from the following data

Fertilizer	0	1	2	3	4	5
Yield	9	16	25	38	50	60

[GCUF STA 324 2018]

Solution:

X_i	Y_i	X_iY_i	X_i^2
0	9	0	0
1	16	16	1
2	25	50	4
3	38	114	9
4	50	200	16
5	60	300	25
15	**198**	**680**	**55**

Estimated regression equation for predicting the yield crop Y_i given the fertilizer X_i is

$\hat{Y}_i = a + bX_i$

First method: using normal equation.

Normal equations are:

$$\sum_i Y_i = na + b\sum_i X_i$$

$$\sum_i X_iY_i = a\sum_i X_i + b\sum_i X_i^2$$

Putting values:

$198 = 6a + 15b$ ----(i)

$680 = 15a + 55b$ ----(ii)

Multiplying equation (i) by 15 and (ii) by 6 and subtracting (i) from (ii) then

$4080 = 90a + 330b$ -----(ii)

$2970 = 90a + 225b$ -----(i)

$1110 = 105b \Rightarrow b = 10.57$

Put value of b in (i) we get

$198 = 6a + 15(10.57)$ $a = 6.57$

Hence the required estimated regression line is

$\hat{Y}_i = 6.57 + 10.57X_i$

Second method: using formula

$\bar{X} = \frac{\sum X_i}{n} = \frac{15}{6} = 2.5$ and $\bar{Y} = \frac{\sum Y_i}{n} = \frac{198}{6} = 33$

$$b = \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} = \frac{6(680) - 15(198)}{6(55) - (15)^2} = \frac{1110}{105} = 10.57$$

$a = \bar{Y} - b\bar{X} = 33 - (10.57)(2.5) = 6.57$

Hence the required estimated regression line is $\hat{Y}_i = 6.57 + 10.57X_i$

Example 10.3: (a) Fit a regression equation to predict the yield of crop from the following data

Fertilizer	0	1	2	3	4	5
Yield	9	16	25	38	50	60

(GCUF STA 324 2018)

(b) Prove that $\sum e_i = \sum\left(Y_i - \hat{Y}_i\right) = 0$

(c) Prove that $\sum e_i^2 = \sum Y_i^2 - a\sum Y_i - b\sum X_iY_i$

(d) Compute the standard error of estimate.

(e) Predict value of yield when 3.5 fertilizer is used and when average value of fertilizer is used.

Solution:

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X - \bar{X})^2$
0	9	-2.5	-24	60	6.25
1	16	-1.5	-17	25.5	2.25
2	25	-0.5	-8	4	0.25
3	38	0.5	5	2.5	0.25
4	50	1.5	17	25.5	2.25
5	60	2.5	27	67.5	6.25
15	**198**	**0**	**0**	**185**	**17.5**

Estimated regression equation for predicting the yield crop (Y) given the fertilizer (X) is

$\hat{Y}_i = a + bX_i$

$$\bar{X}=\frac{\sum X_i}{n}=\frac{15}{6}=2.5 \text{ and } \bar{Y}=\frac{\sum Y_i}{n}=\frac{198}{6}=33$$

$$b=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}=\frac{185}{17.5}=10.57$$

$$a=\bar{Y}-b\bar{X}=33-(10.57)(2.5)=6.57$$

Hence the required estimated regression line is $\hat{Y}_i=6.57+10.57X_i$

X_i	Y_i	$\hat{Y}_i$	$Y_i-\hat{Y}$	$(Y_i-\hat{Y})^2$	Y_i^2
0	9	6.57	2.43	5.905	81
1	16	17.14	-1.14	1.300	256
2	25	27.71	-2.71	7.344	625
3	38	38.29	-0.29	0.084	1444
4	50	48.86	1.14	1.300	2500
5	60	59.43	0.57	0.325	3600
15	**198**	**198**	**0**	**16.26**	**8506**

(b) $\sum e_i=\sum(Y_i-\hat{Y})=0$ (From above table)

(c) $\sum Y_i^2-a\sum Y_i-b\sum X_iY_i$

$=8506-(6.57143)(198)-(10.57143)(680)=16.3$

$\sum e_i^2=\sum(Y_i-\hat{Y})^2=16.26\approx16.3$

Hence $\sum e_i^2=\sum Y_i^2-a\sum Y_i-b\sum X_iY_i$

(d) Standard error of estimate $S_{y.x}=\sqrt{\frac{\sum(Y_i-\hat{Y})^2}{n-2}}=\sqrt{\frac{16.26}{6-2}}=2.02$

(e) Predicted value of yield when 3.5 fertilizer is used

$\hat{Y}|X=3.5=6.57+10.57(3.5)=43.56$

Predicted value of yield when average 2.5 fertilizer is used

$E(\hat{Y}|X=2.5)=6.57+10.57(2.5)=33.0$

Example 10.4: Marks in the midterm exam predict marks in the final term exam. The regression constant (intercept) in the linear regression model for predicting final exam marks from midterm exam marks is 40 and the regression coefficient is 1.5. Write the linear regression model for this example. Predict final term exam marks for students whose marks in the midterm were 13, 24, and 15

Solution: X = Marks in midterm exam; Y = Marks in final term exam

Regression Constant (intercept): $a = 40$

Regression coefficient (slope): $b = 1.5$

Linear regression model: $\hat{Y}_i=40+1.5X_i$

Prediction of final exam marks:

$\hat{Y}_i|X=13=40+1.5(13)=59.5$

$\hat{Y}_i|X=24=40+1.5(24)=76$

$\hat{Y}_i|X=15=40+1.5(15)=62.5$

10.15 Correlation

Interdependence between two or more than two variables is called correlation. For example there is correlation between (i) marks obtained by a student and his/her study hour, (ii) level of uric acid and consumption of red meat, (iii) sugar level and exercise time of an individual and (iv) yield of a crop and fertility of the land etc.

10.15.1 Correlation coefficient (r)

The quantity r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of its developer Karl Pearson. It is given as

$$r=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}}$$

It gives

(i) Change in dependent variable in standard deviation unit with one standard deviation increase in independent variable.

(iv) Direction and strength of linear relationship between two variables.

Other formulae

$$r=\frac{\Sigma XY-\frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left[\Sigma X^2-\frac{(\Sigma X)^2}{n}\right]\left[\Sigma Y^2-\frac{(\Sigma Y)^2}{n}\right]}}$$

$$r=\frac{n\Sigma XY-(\Sigma X)(\Sigma Y)}{\sqrt{\left[n\Sigma X^2-(\Sigma X)^2\right]\left[n\Sigma Y^2-(\Sigma Y)^2\right]}}$$

$$r=\frac{n\Sigma D_x D_y-(\Sigma D_x)(\Sigma D_y)}{\sqrt{\left[n\Sigma D_x^2-(\Sigma D_x)^2\right]\left[n\Sigma D_y^2-(\Sigma D_y)^2\right]}}$$

Where $D_x=X-A$ and $D_y=Y-B$

$$r=\frac{\Sigma XY-n\bar{X}\bar{Y}}{\sqrt{\left[\Sigma X^2-n(\bar{X})^2\right]\left[\Sigma Y^2-n(\bar{Y})^2\right]}}$$

$$r=\frac{n\Sigma UV-(\Sigma U)(\Sigma V)}{\sqrt{\left[n\Sigma U^2-(\Sigma U)^2\right]\left[n\Sigma V^2-(\Sigma V)^2\right]}}$$

Where $U=\frac{X-A}{h}$ and $V=\frac{Y-B}{k}$

$$r=\frac{Cov(X,Y)}{SD(X)SD(Y)}=\frac{S_{xy}}{S_xS_y}$$

10.15.2 Properties of correlation coefficient

01) It is a pure number.
02) It lies between -1 and +1.
03) It is symmetrical i.e. $r_{xy}=r_{yx}$.
04) It is independent of change of origin and scale i.e. $r_{uv}=r_{xy}$.
05) It is GM of both regression coefficients b_{xy} and b_{yx} i.e. $r=\pm\sqrt{b_{xy}\times b_{yx}}$.

It may be noted that the sign of the correlation coefficient "r" is the same as that of the regression coefficients.

10.15.3 Interpretation of correlation coefficient r

01) If $r>0$ it means that variables move in the same direction.
02) If $r<0$ it means variables move in the opposite direction.

03) If $r=0$ it means there is no relationship between the variables.
04) If $r=+1$ it means there is perfect positive correlation.
05) If $r=-1$ it means there is perfect negative correlation.

10.16 Coefficient of Determination

The coefficient of determination is.

- The percent of the variation that can be explained by the regression equation.
- The explained variation divided by the total variation
- The square of r

A statistic which indicates the strength of fit between two variables implied by a particular value of the sample correlation coefficient r, designated by r^2 is known as coefficient of determination. It is given as

$$r^2=\frac{Explained\ Variation}{Total\ Variation}=\frac{\sum_{i=1}^{n}(\hat{Y}-\bar{Y})^2}{\sum_{i=1}^{n}(Y-\bar{Y})^2}=1-\frac{\sum_{i=1}^{n}(Y-\hat{Y})^2}{\sum_{i=1}^{n}(Y-\bar{Y})^2}$$

Example 10.5: Calculate coefficient of correlation between the yield of crop and fertilizer used.

Fertilizer	0	1	2	3	4	5
Yield	9	16	25	38	50	60

Solution:

X_i	0	1	2	3	4	5	**15**
Y_i	9	16	25	38	50	60	**198**
X_iY_i	0	16	50	114	200	300	**680**
X_i^2	0	1	4	9	16	25	**55**
Y_i^2	81	256	625	1444	2500	3600	**8506**

$$r=\frac{n\sum X_iY_i-\sum X_i\sum Y_i}{\sqrt{\left\{n\sum X_i^2-\left(\sum X_i\right)^2\right\}\left\{n\sum Y_i^2-\left(\sum Y_i\right)^2\right\}}}=\frac{6(680)-(15)(198)}{\sqrt{\left\{6(55)-(15)^2\right\}\left\{6(8506)-(198)^2\right\}}}$$

$$= \frac{1110}{\sqrt{(105)(11832)}} = \frac{1110}{1114.61} = 0.99$$

Coefficient of determination: $r^2 = 0.98$

Example 10.6: Estimate regression equations (Y on X and X on Y) and prove that coefficient of correlation is the GM of both regression coefficients obtained in the equations.

X	1	2	3	4	5	6	7
Y	5	3	2	6	4	6	7

Solution:

X_i	1	2	3	4	5	6	7	28
Y_i	5	3	2	6	4	6	7	33
X_iY_i	5	6	6	24	20	36	49	146
X_i^2	1	4	9	16	25	36	49	140
Y_i^2	25	9	4	36	16	36	49	175

The regression line Y on X: $\hat{Y}_i = a_{yx} + b_{yx}X_i$

$$b_{yx} = \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{\left\{n\sum X_i^2 - \left(\sum X_i\right)^2\right\}} = \frac{7(146) - (28)(33)}{7(140) - (28)^2} = \frac{98}{196} = 0.50$$

$$a_{yx} = \bar{Y} - b_{yx}\bar{X} = 4.71 - 0.50(4) = 2.71$$

Hence the required estimated line is

$$\hat{Y}_i = 2.71 + 0.50X_i$$

The regression line X on Y: $\hat{X}_i = a_{xy} + b_{xy}Y_i$

$$b_{xy} = \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{\left\{n\sum Y_i^2 - \left(\sum Y_i\right)^2\right\}} = \frac{7(146) - (28)(33)}{7(175) - (33)^2} = \frac{98}{136} = 0.72$$

$$a_{xy} = \bar{X} - b_{xy}\bar{Y} = 4 - 0.72(4.71) = 0.61$$

Hence the required estimated line is $\hat{X}_i = 0.61 + 0.72Y_i$

$$r = \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{\sqrt{\left\{n\sum X_i^2 - \left(\sum X_i\right)^2\right\}\left\{n\sum Y_i^2 - \left(\sum Y_i\right)^2\right\}}}$$

$$= \frac{7(146) - (28)(33)}{\sqrt{\left\{7(140) - (28)^2\right\}\left\{7(175) - (33)^2\right\}}} = \frac{98}{\sqrt{(196)(136)}} = 0.60$$

Now $\sqrt{b_{yx}b_{xy}} = \sqrt{0.50 \times 0.72} = 0.60$; So it is proved that $r = \sqrt{b_{yx}b_{xy}}$.

10.17 Rank correlation

An alternative method introduced by Spearman to find the strength of relationship between two qualitative variables measuring on ordinal scale, by ordering them (in ascending or descending) is known as rank correlation. Ordering of values is called ranks, and correlation between ranks of the variables is called rank correlation.

It may also be used in the situation when quantitative variables having any outlier (extreme value).

Spearman rank correlation coefficient is given by $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$; $-1 \le r_s \le +1$

It is used when

1. The data is qualitative
2. Accurate assessment is not possible
3. The actual measurement is not available.
4. When there exists outlier.

Example 10.7: Marks of students in the subjects of genetics and statistics of the UOO is given below. Find rank correlation between them.

Genetics	50	61	65	85	60	40	80	90	45
Stat	60	65	73	80	55	50	67	81	49

Solution:

Genetics	Stat	Ranks		d	d^2
		Genetics	Stat		
50	60	7	6	1	1
61	65	5	5	0	0

65	73	4	3	1	1
85	80	2	2	0	0
60	55	6	7	-1	1
40	50	9	8	1	1
80	67	3	4	-1	1
90	81	1	1	0	0
45	49	8	9	-1	1
					6

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 6}{9(9^2-1)} = 1 - \frac{36}{720} = 0.95$$

Example 10.8: the members of selection committee ranked nine persons according to their suitability for promotion as follows

Person	A	B	C	D	E	F	G	H	I
Member-I	1	3	5	6	3	7	8	3	9
Member-II	3	4.5	8	4.5	2	7	6	1	9

Calculate coefficient of rank correlation.

Solution:

Person	member-I	Member-II	d	d^2
A	1	3	-2	4
B	3	4.5	-1.5	2.25
C	5	8	-3	9
D	6	4.5	1.5	2.25
E	3	2	1	1
F	7	7	0	0
G	8	6	2	4
H	3	1	2	4
I	9	9	0	0
				26.5

There tied ranks in this ranking of both members.

t	3(B,E and H)	2 (B and D)	
t^3-t	$3^3-3=24$	$2^3-2=6$	30

$$\frac{1}{12}\sum(t^3-t) = \frac{30}{12} = 2.5$$

Adjusted $\sum d^2 = \sum d^2 + \frac{1}{12}\sum(t^3-t) = 26.5 + 2.5 = 29$

$$r_s = 1 - \frac{6\sum d^2}{n(n^3-1)} = 1 - \frac{6\times 29}{9(9^2-1)} = 1 - \frac{174}{720} = 1 - 0.24 = 0.76$$

10.17 MULTIPLE REGRESSION ANALYSIS

10.17.1 Multiple regression

The dependence of one variable on two or more independent variables is called multiple regression. Its general model is $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i,\ i = 1,2,...,n$

Xs have fixed values.

β's are partial regression coefficients, ε_i is the error term normally distributed with mean 0 and variance σ^2.

Multiple regression model with two regressor (independent variables)

$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

Its model for sample data is $\hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i},\ i = 1,2,...,n$

10.17.2 Standard error of estimate in multiple regression

The standard error of the estimate in multiple regression analysis with two independent variables, is given as

$$S_{y.12} = \sqrt{\frac{\sum_{i=1}^{n}\varepsilon_i^2}{n-3}} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-3}} \text{ or } S_{y.12} = \sqrt{\frac{\sum_{i=1}^{n}Y_i^2 - a\sum_{i=1}^{n}Y_i - b_1\sum_{i=1}^{n}X_{1i}Y_i - b_2\sum_{i=1}^{n}X_{2i}Y_i}{n-3}}$$

It is an estimate of badness of the fitting model. It relates to regression line as standard deviation to mean.

10.17.3 Multiple correlation coefficient

Multiple correlation coefficient is defined as the positive square root of the coefficient of multiple determination. It measures the degree of association between Y and both the regressor combined. It is always taken to be positive.

Formulae for different combinations:

$$R_{1.23}=\sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}}{1-r_{23}^2}};\quad R_{2.13}=\sqrt{\frac{r_{12}^2+r_{23}^2-2r_{12}r_{13}r_{23}}{1-r_{13}^2}};\quad R_{3.12}=\sqrt{\frac{r_{13}^2+r_{23}^2-2r_{12}r_{13}r_{23}}{1-r_{12}^2}}$$

10.17.4 Partial correlation

Let there be three or more associated variables, then the ordinary correlation between any two of them for the fixed values of the other is called partial correlation.

10.17.5 Partial correlation coefficient

The partial correlation coefficient measures the strength of linear relationship between any two variables considering the effect of other variables as constant. In case of three variables X_1, X_2 and X_3 problem the various partial correlation coefficients are given below.

$$r_{12.3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{\left(1-r_{13}^2\right)\left(1-r_{23}^2\right)}},$$

Partial correlation coefficient between X_1 and X_2 when X_3 is kept constant.

Similarly $r_{13.2}=\dfrac{r_{13}-r_{12}r_{23}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{23}^2\right)}}$ and $r_{23.1}=\dfrac{r_{23}-r_{12}r_{13}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{13}^2\right)}}$

It lies between -1 and +1.

Note: $r_{12}=r_{21}, r_{13}=r_{31}, r_{23}=r_{32}$

Example 10.9: (a) fit the multiple regression line X_3 on X_1 and X_2 using normal equations and formulae.

(b) Find Standard error of estimate.

(c) Calculate Coefficient of multiple determination.

(d) Compute coefficient of multiple correlation.

X_1	2	3	0	2	4
X_2	1	2	3	4	5
X_3	12	10	9	13	16

Solution:

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_1^2	X_2^2	X_3^2
2	1	12	2	24	12	4	1	144
3	2	10	6	30	20	9	4	100
0	3	9	0	0	27	0	9	81
2	4	13	8	26	52	4	16	169
4	5	16	20	64	80	16	25	256
11	**15**	**60**	**36**	**144**	**191**	**33**	**55**	**750**

Model is: $X_3=a+b_1X_1+b_2X_2$

Normal equations:

$$\sum X_3=na+b_1\sum X_1+b_2\sum X_2$$
$$\sum X_1X_3=a\sum X_1+b_1\sum X_1^2+b_2\sum X_1X_2$$
$$\sum X_2X_3=a\sum X_2+b_1\sum X_1X_2+b_2\sum X_2^2$$

Putting the values:

$60=5a+11b_1+15b_2$ -----(i)

$144=11a+33b_1+36b_2$ -----(ii)

$191=15a+36b_1+55b_2$ -----(iii)

Solution by calculator yield the following results.

$a=7.27, b_1=1.10$ and $b_2=0.77$

Hence required multiple regression line X_3 on X_1 and X_2 is

$$\hat{X}_3=7.27+1.10X_1+0.77X_2$$

Using formulae

$$\sum x_1^2=\sum X_1^2-\frac{\left(\sum X_1\right)^2}{n}=33-\frac{(11)^2}{5}=8.8$$

$$\sum x_2^2=\sum X_2^2-\frac{\left(\sum X_2\right)^2}{n}=55-\frac{(15)^2}{5}=10$$

$$\sum x_3^2=\sum X_3^2-\frac{\left(\sum X_3\right)^2}{n}=750-\frac{(60)^2}{5}=30$$

$$\sum x_1x_2=\sum X_1X_2-\frac{\left(\sum X_1\right)\left(\sum X_2\right)}{n}=36-\frac{(11)(15)}{5}=3$$

$$\sum x_1x_3 = \sum X_1X_3 - \frac{(\sum X_1)(\sum X_3)}{n} = 144 - \frac{(11)(60)}{5} = 12$$

$$\sum x_2x_3 = \sum X_2X_3 - \frac{(\sum X_2)(\sum X_3)}{n} = 191 - \frac{(15)(60)}{5} = 11$$

$$b_1 = \frac{\sum x_1x_3 \sum x_2^2 - \sum x_1x_2 \sum x_2x_3}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} = \frac{(12)(10)-(11)(3)}{(8.8)(10)-(3)^2} = \frac{120-33}{88-9} = \frac{87}{79} = 1.10$$

$$b_2 = \frac{\sum x_2x_3 \sum x_1^2 - \sum x_1x_2 \sum x_1x_3}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} = \frac{(11)(8.8)-(3)(12)}{(8.8)(10)-(3)^2} = \frac{96.8-36}{88-9} = \frac{60.8}{79} = 0.77$$

$$a = \bar{X}_3 - b_1\bar{X}_1 - b_2\bar{X}_2 = 12-(1.10)(2.2)-(0.77)(3) = 7.27$$

Hence required multiple regression line X_3 on X_1 and X_2 is $\hat{X}_3 = 7.27 + 1.10X_1 + 0.77X_2$

Standard error of estimate

$$S_{3.12} = \sqrt{\frac{\sum X_3^2 - a\sum X_3 - b_1\sum X_1X_3 - b_2\sum X_2X_3}{n-3}} = \sqrt{\frac{750-7.27(60)-1.10(144)-0.77(191)}{5-3}}$$

$$S_{3.12} = \sqrt{\frac{8.33}{2}} = \sqrt{4.165} = 2.04$$

$Total$ variation $= \sum x_3^2 = 30$; Unexplained variation $= \sum\left(Y-\hat{Y}\right) = 8.33$

Total variation = Explained variation + Unexplained variation

Explained variation = 30 − 8.33 = 21.67

Coefficient of multiple determination:

$$R^2_{3.12} = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{21.67}{30} = 0.72$$

Coefficient of multiple correlation: $R_{3.12} = \sqrt{0.72} = 0.85$

10.18 CURVE FITTING

Modelling the relationship between variables by nonlinear function is called curve fitting. There are many curves that can be used for this purpose, some are as follows

01) Second degree curve(parabola)
02) Third degree curve
03) Exponential curve
04) Power curve
05) Reciprocal curve

10.18.1 Second Degree Curve:

It is suitable for negatively skewed data with positive values.

10.18.2 Exponential /semi logarithmic curve

When change in dependent variable is proportional to the magnitude of the corresponding value of the independent variable, such relationship is represented by the exponential curve. Mathematically exponential curve is one in which independent variable occur in exponent. It is represented as $Y = ac^{bX}$. It is also known as semi logarithmic curve.

It is used to describe a relation in which one variable form approximately a geometric progression while the other form an arithmetic progression.

01) Data from the field of biology, banking and economics shows the exponential trends.
02) The growth of bacteria is exponentials.
03) Money accumulating at compound interest rate is exponential.
04) Sales or earnings in business may grow exponentially over a short period.

10.18.3 Geometric / Power / logarithmic Curve

Equation of geometric/power curve is $Y = aX^b$ the curve is suitable when both the variable X and Y Change by constant percentage i.e when both X and Y change by geometric progression. Mathematically power curve is one in which one of the parameter quantity occur in exponent.

The curve is suitable for the right skewed data with non-zero positive values, moreover it is applicable for data which form a straight line when the values of $\log X$ and $\log Y$ are plotted on the graph paper.

10.18.4 Criteria for fitting curve

We are given two methods for selecting suitable curve to be fitted by given data.

01) By differencing the value of dependent variables.
02) By plotting the data on graph paper.

By differencing the values

01) A straight line should be used, when the first differences of dependent variable "Y" are constant.
02) A second degree parabola should be used, if second differences of dependent variable "Y" are constant.
03) A third degree parabola should be used, if third differences of dependent variable "Y" are constant.
04) An exponential curve is used if first differences of log Y are constant.
05) Use geometric curve if first differences of $\log X$ and $\log Y$ are constant.

By plotting the data on graph paper:

01) A straight line is used if graph of X and Y is a straight line.
02) A second degree parabola is used if graph of X and Y gives one bend" $\cup, \cap, \subset$ *and* $\supset$".
03) A third degree parabola is used if graph of X and Y gives "S" shape (in reverse s).
04) If graph of the value of "X" and "log Y" gives a straight line, use an exponential or semi logarithmic curve.
05) If graph of the values of "log X" and "log Y" gives a straight line, use a geometric or logarithmic curve.

Example 10.10: From the data given below fit

(i) Straight line $Y = a + bX$
(ii) Second degree parabola $Y = a + bX + cX^2$
(iii) Exponential curve $Y = ab^X$
(iv) Power curve $Y = aX^b$
(v) The curve $Y = a + b\frac{1}{X}$

Select the best one model from these different models fitted.

X	10	12	14	16	18	20
Y	14	20	25	28	22	17

Solution:

(i) Straight line $\hat{Y} = a + bX$ **:**

X	Y	XY	X^2	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
10	14	140	100	19.285	-5.285	27.9312
12	20	240	144	19.971	0.029	0.0008
14	25	350	196	20.657	4.343	18.8616
16	28	448	256	21.343	6.657	44.3156
18	22	396	324	22.029	-0.029	0.0008
20	17	340	400	22.715	-5.715	32.6612
90	**126**	**1914**	**1420**	**126**	**0**	**123.7714**

$$\hat{Y} = a + bX$$

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$b = \frac{6(1914) - (90)(126)}{6(1420) - (90)^2} = \frac{144}{420} = 0.343$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{126}{6} = 21; \quad \bar{X} = \frac{\sum X}{n} = \frac{90}{6} = 15$$

$$a = \bar{Y} - b\bar{X} = 21 - 0.343(15) = 15.855$$

Fitted straight line model: $\hat{Y} = 15.855 + 0.343X$

Sum of squares of residuals: $\sum (Y-\hat{Y})^2 = 123.7714$

(ii) Second degree curve $Y = a + bX + cX^2$ **:**

X	Y	XY	X^2	X^3	X^4	X^2Y
10	14	140	100	1000	10000	1400
12	20	240.	144	1728	2880	20736
14	25	350	196	2744	38426	4900
16	28	448	256	4096	65536	7168
18	22	396	324	5832	104976	7128
20	17	340	400	8000	160000	6800
90	**126**	**1914**	**1420**	**23400**	**399664**	**30276**

$$Y = a + bX + cX^2$$

Normal equations:

$$\sum Y = na + b\sum X + c\sum X^2$$
$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$
$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

$6a + 90b + 1420c = 126$ (i)

$90a + 1420b + 23400c = 1914$ (ii)

$1420a + 23400b + 399664c = 30276$ (iii)

Multiplying (i) by 15 and subtracting from (ii) and simplifying results in

$35b + 1050c = 12$ (iv)

$1575b + 47698c = 342$ (v)

Multiplying (iv) by 45 and subtracting from (v) results in

$448c = -198 \Rightarrow c = -0.44196$

Put value of c in (iv) gives b = 13.60179

Put value of "b" and "c" in (i) gives a = - 78.42857

Fitted parabolic model:

$\hat{Y} = -78.42857 + 13.60179X - 0.44196X^2$

X	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
10	14	13.39286	0.6071	0.3681
12	20	21.15000	-1.1500	1.3250
14	25	25.37143	-0.3714	0.1386
16	28	26.05714	1.9429	3.7702
18	22	23.20714	-1.2071	1.4607
20	17	16.82143	0.1786	0.0312
90	**126**	**126**	**0**	**7.0938**

Sum of squares of residuals: $\sum(Y-\hat{Y})^2 = 7.0938$

(iii) Exponential curve $Y = ab^X$:

$$Y = ab^X$$
$$\ln Y = \ln(ab^X) = \ln a + X \ln b$$
$$Y^* = a^* + b^* X$$
$$b^* = \frac{n\sum XY^* - \sum X \sum Y^*}{n\sum X^2 - (\sum X)^2}$$

X	Y	Y^*	XY^*	X^2
10	14	2.6391	26.3906	100
12	20	2.9957	35.9488	144
14	25	3.2189	45.0643	196
16	28	3.3322	53.3153	256
18	22	3.0910	55.6388	324
20	17	2.8332	56.6643	400
90	**126**	**18.1101**	**273.0219**	**1420**

$$b^* = \frac{6(273.0219) - (90)(18.1101)}{6(1420) - (90)^2} = \frac{8.2224}{420} = 0.0196$$

$$b^* = \ln b \Rightarrow 0.0196 = \ln b \Rightarrow b = 1.019793$$

$$\bar{Y}^* = \frac{\sum Y^*}{n} = \frac{18.1101}{6} = 3.01835; \quad \bar{X} = \frac{\sum X}{n} = \frac{90}{6} = 15$$

$$a^* = \bar{Y}^* - b^*\bar{X} = 3.01835 - 0.0196(15) = 2.7247$$

$$a^* = \ln a \Rightarrow 2.7247 = \ln a \Rightarrow a = 15.25184$$

Fitted straight line model: $\hat{Y} = 15.25297(1.019765)^X$

X	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
10	14	18.55046	-4.5507	20.7066
12	20	19.29099	0.7090	0.5027
14	25	20.06109	4.9389	24.3928
16	28	20.86194	7.13806	50.9520
18	22	21.69475	0.3053	0.0931
20	17	22.56081	-5.5608	30.9226
90	**126**	**126**		**127.5698**

Sum of squares of residuals: $\sum(Y-\hat{Y})^2 = 127.5698$

(iv) Power curve $Y = aX^b$:

X	Y	X^*	Y^*	X^*Y^*	X^{*2}
10	14	2.3026	2.6391	6.0768	5.3019
12	20	2.4849	2.9957	7.4440	6.1748
14	25	2.6391	3.2189	8.4950	6.9646
16	28	2.7726	3.3322	9.2388	7.6872
18	22	2.8904	3.0910	8.9342	8.3542
20	17	2.9957	2.8332	8.4874	8.9744
90	**126**	**16.0852**	**18.1101**	**48.6762**	**43.4572**

$$Y = aX^b$$

$$\ln Y = \ln(aX^b)$$

$$\ln Y = \ln a + b\ln X$$

$$Y^* = a^* + bX^*$$

$$b^* = \frac{n\sum X^*Y^* - \sum X^*\sum Y^*}{n\sum X^{*2} - \left(\sum X^*\right)^2} = \frac{6(48.6762) - (16.0852)(18.1101)}{6(43.4572) - (16.0852)^2} = \frac{0.7515084}{2.008146} = 0.37423$$

$$\bar{Y}^* = \frac{\sum Y^*}{n} = \frac{18.1101}{6} = 3.01835;\quad \bar{X}^* = \frac{\sum X^*}{n} = \frac{16.08524}{6} = 2.6809$$

$$a^* = \bar{Y}^* - b\bar{X}^* = 3.01835 - 0.37423(2.6809) = 2.0152$$

$$a^* = \ln a \Rightarrow a = 7.5013$$

Fitted straight line model: $\hat{Y} = 7.5013X^{0.3742}$

X	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
10	14	17.75687	-3.7569	14.1141
12	20	19.01072	0.9893	0.9787
14	25	20.13966	4.8603	23.6229
16	28	21.17163	6.8284	46.6270
18	22	22.12571	-0.1257	0.0158
20	17	23.01553	-6.0155	36.1866
90	**126**			**121.5451**

Sum of squares of residuals: $\sum(Y-\hat{Y})^2 = 121.5451$

(v) The curve $Y = a + b\frac{1}{X}$:

X	Y	X^*	X^*Y	X^{*2}
10	14	0.1000	1.4	0.01
12	20	0.0833	1.6667	0.0069
14	25	0.0714	1.7857	0.0051
16	28	0.0625	1.75	0.0039
18	22	0.0555	1.2222	0.0031
20	17	0.0500	0.85	0.0025
90	**126**	**0.4228**	**8.6746**	**0.0315**

$$b = \frac{n\sum X^*Y - \sum X^*\sum Y}{n\sum X^{*2} - \left(\sum X^*\right)^2} = \frac{6(8.6746) - (0.4228)(126)}{6(0.0315) - (0.4228)^2} = -117.3368$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{126}{6} = 21;\quad \bar{X}^* = \frac{\sum X^*}{n} = \frac{0.4228}{6} = 0.0705$$

$$a = \bar{Y} - b\bar{X}^* = 21 + 117.3368(0.0705) = 29.2687$$

Fitted straight line model: $\hat{Y} = 29.2687 - 117.3368\frac{1}{X}$

X	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
10	14	17.5350	-3.5350	12.4962
12	20	19.4906	0.5094	0.2595
14	25	20.8875	4.1125	16.9129
16	28	21.9351	6.0649	36.7827
18	22	22.7499	-0.7500	0.5624
20	17	23.4018	-6.4018	40.9835
90	**126**			**107.9972**

Sum of squares of residuals: $\sum(Y-\hat{Y})^2 = 107.9972$

As sum of squares for second degree (parabolic trend) is minimum therefore it is best choice for modeling this data set.

Multiple choice questions

A

1. Slope lies:
(a) -1 to +1 (b) 0 to 1 (c) $-\infty$ to ∞ (d) 0 to ∞

2. In regression, $\sum(Y-\hat{Y})=$:
(a) 0 (b) < 0 (c) >0 (d) $\neq 0$

3. Independent variable:
(a) Regressand (b) predictand (c) Regressor (d) explained

4. Dependent variable:
(a) Regressand (b) Regressor (c) explanatory (d) fixed

5. In the regression, $\hat{Y} = a + bX_i$
(a) $\sum X = \sum \hat{X}$ (b) $\sum Y = \sum \hat{Y}$ (c) $\sum Y = \sum X$ (d) $\sum \hat{Y} = \sum \hat{X}$

6. Provides basis for estimation:
(a) Dependent variable (b) independent variable
(c) Regressand (d) Random variable

7. Regression co-efficient is independent of:
(a) unit of measurement (b) change of origin (c) change of scale (d) both b and c

8. The sum of squares of residuals:
(a) e (b) Σe (c) Σe^2 (d) $\Sigma e^2/n$

9. In regression $\hat{Y} = a + bX_i$, X is ------ variable
(a) Dependent (b) independent (c) Qualitative (d) quantitative

10. Regression co efficient for Y on X:
(a) $\frac{\sum XY - \sum X \sum Y}{\sum X^2 - (\sum x)^2}$ (b) $\frac{n\sum XY - \sum X \sum Y}{\sum X^2 - (\sum x)^2}$
(c) $\frac{\sum XY - (\sum X \sum Y)/n}{\sum X^2 - (\sum x)^2/n}$ (d) $\frac{\sum XY - (\sum X \sum Y)}{n\sum X^2 - (\sum x)^2}$

11. The variable whose value is predicted:
(a) Independent variable (b) Cause (c) Regressand (d) Regressor

12. The difference between the actual and trend value:
(a) Slope (b) Residual (c) Intercept (d) Sum of residual

13. In regression line $\hat{Y} = a + bX_i$ the following is always true:
(a) $b = \frac{y_2 - y_1}{x_2 - x_1}$ (b) $\sum(X-\hat{X})^2 = 0$ (c) $\sum(Y-\hat{Y}) = 0$ (d) $\sum(Y-\hat{Y})^2 = 0$

14. If $\hat{Y} = 20 - 3X_i$ and $\hat{X} = 4 - 0.25Y_i$ then $r =$ ---
(a) 0 (b) 0.87 (c) -0.87 (d) 0.75

15. In regression $\hat{Y} = a + bX_i$, Y is ------ independent
(a) Independent (b) Dependent (c) Fix (d) Constant

16. $\hat{Y} = 5 + 2X_i$ If X = 5, then value of dependent variable:
(a) 15 (b) 5 (c) 0 (d) 10

17. The regression co-efficient of X on Y :
(a) $\frac{Cov(X,Y)}{\text{var}(X)}$ (b) $\frac{Cov(X,Y)}{\text{var}(Y)}$ (c) $\frac{Cov(X,Y)}{SD(Y)}$ (d) $\frac{Cov(X,Y)}{SD(X)}$

18. Co-variance (X, Y):
(a) $\frac{\sum(X-\bar{X})(Y-\bar{Y})}{n}$ (b) $\frac{\sum(X-\bar{X})^2}{n}$ (c) $\frac{\sum(Y-\bar{Y})^2}{n}$ (d) $\frac{\sum(X-\bar{X})^2(Y-\bar{Y})^2}{n}$

19. If $Cov(X,Y) = 0$, then correlation co-efficient is:
(a) Positive (b) Negative (c) Zero (d) 1

20. The term regression was used by:
(a) Pearson (b) Galton (c) Spearman (d) Newton

21. In simple linear regression one variable depends upon:
(a) One dependent variable (b) Two dependent variables
(c) One independent variable (d) Two independent variables

22. Dependence of variable on a single independent variable, regression is
(a) Simple (b) Partial (c) Multiple (d) multivariate

23. Galton defined regression as tendency towards:
(a) Average (b) Dispersion (c) skewness (d) kurtosis

24. If $\hat{Y} = a + bX_i$, b:
(a) Intercept (b) Slope (c) Regressor (d) Regressand

25. Correlation co-efficient can never:
(a) = 1 (b) > 1 (c) = 0 (d) < 1

26. The graph showing the paired points (X, Y):
(a) Historigram (b) Histogram (c) Scatter diagram (d) curve

27. Independent variable is assumed to be:
(a) Random (b) Constant (c) Fixed (d) Zero

28. Dependent variable is assumed to be:
(a) Random (b) Non-random (c) Fixed (d) Zero

29. In scatter diagram, clustering of points around a straight line, regression is
(a) Linear (b) Non-linear (c) Curvilinear (d) logarithmic

30. Regression co-efficient is zero, then two variables are:
(a) Dependent (b) Independent (c) Correlated (d) Associated

31. The line of regression always passes through:
(a) (X,Y) (b) $(\hat{X},\hat{Y})$ (c) $(\hat{X},Y)$ (d) $(\bar{X},\bar{Y})$

32. $\sum(Y-\hat{Y})^2$: (a) Zero (b) Least (c) Maximum (d) 1

33. In simple linear regression, number of independent variables:
(a) 0 (b) 1 (c) 2 (d) 3

34. In method of Least Square (L.S) estimation:
(a) Minimize the sum of residuals (b) Maximize the sum of residuals

(c) Minimize the sum of squares of residuals
(d) Maximize the sum of squares of residuals

35. $Y_i = \alpha + \beta X_i + \varepsilon_i$, which of the following statement is true for β:
(a) It is intercept on y-axis (b) It is rate of increase in y per unit increase in x
(c) It is rate of increase in x per unit increase in y (d) It is an error term

36. $\hat{Y} = 10 - 2X_i$, slope of line:
(a) 10 (b) -2 (c) 2 (d) 8

37. In $\hat{Y} = a + bX_i$, $\hat{Y}$ is:
(a) Independent variable (b) dependent variable
(c) response (d) predicted response

38. Regression is concerned with all except
a) Modeling the relationship between two quantitative variables.
b) Estimating the response.
c) Modeling the relationship between more than two quantitative variables.
d) Testing the equality of average values of two sets.

39. Deterministic relation:
a) Relationship between age and weight of a person.
b) Relationship between education and income.
c) Relationship between eye colour and hair colour.
d) Relationship between area and radius of a circle.

40. The regression coefficient between two variables is 1.7, it means
a) The average value of Y change by 1.7 when X is increased by 1.
b) The average value of X change by 1.7 when Y is increased by 1.
c) The X explains 17% of the variation in Y.
d) The X explains 1.7% of the variation in Y.

41. Relationship between two variables is linear, it means
a) All ordered pairs lie on a second degree curve.
b) All ordered pairs approximately lie on a curve
c) Scatter plot looks like a straight line d) Scatter plot has no shape

42. Two variables have perfect relationship
a) r = 0 b) r = 0.5 c) r = +1 d) r = ∞

B

1. If $r_{yx} = 0.75$, the correlation co-efficient between u = 1.5x and v = 2y:
(a) zero (b) 0.75 (c) -0.75 (d) 1.5

2. Co-efficient of correlation "r" lies between:
(a) 0 to 1 (b) -1and +1 (c) -1 and 0 (d) -0.5 and +0.5

3. r^2 is 0.49, then coefficient of correlation:
(a) 0.47 (b) 0.7 (c) 0.07 (d) 0.94

4. r is given by: (a) $\frac{s_{xy}}{s_x^2}$ (b) $\frac{s_{xy}}{s_y^2}$ (c) $\frac{s_{xy}}{s_x s_y}$ (d) $\frac{s_{xy}}{s_y^2 s_x^2}$

5. "r" is equal to:
(a) $\frac{b_{xy} b_{yx}}{2}$ (b) $b_{xy} \times b_{yx}$ (c) $\sqrt{b_{yx} \times b_{xy}}$ (d) $\sqrt{b_{xy} + b_{yx}}$

6. If $b_{yx} = -1$ and $b_{xy} = -0.64$ then r is equal to:
(a) +1 (b) -0.8 (c) 0.64 (d) 0.5

7. If $\hat{Y} = 20 - 3X$ and $\hat{X} = 4 - 0.25Y$ then r = - - - -
(a) 0 (b) 0.83 (c) -0.87 (d) 0.75

8. For 10 pairs of observation $\sum X = 385, \sum Y = 574, b_{yx} = -0.69$, then co-efficient of correlation may be:
(a) +0.69 (b) +0.98 (c) +0.50 (d) -0.88

9. The unit of co-efficient of correlation calculated from height (in inches) and weight (in pounds):
(a) Inches (b) Pounds (c) Both inches and pounds (d) No unit

10. If $r = 0.6, b_{xy} = 1.2$ then b_{yx} = - - -
(a) 0.20 (b) 0.72 (c) 0.30 (d) 0.36

11. "r" cannot be greater than:
(a) -1 (b) +1 (c) Zero (d) 0.5

12. If a constant is added to the values of a variable, "r" is:
(a) Negative (b) Positive (c) Zero (d) Remain unchanged

13. If $r_{xy} = -0.84$ then r_{yx} = - - is:
(a) -0.84 (b) 0.84 (c) 0.42 (d) 0.48

14. If $Cov(X, Y) = 0$, then correlation co-efficient is:
(a) Positive (b) Negative (c) Zero (d) 1

15. The value of correlation co-efficient can never:
(a) = 1 (b) > 1 (c) = 0 (d) < 1

16. Which r^2 indicates the weakest relationship between variables?
a) 0.80 b) -0.80 c) 0.15 d) -0.25

17. Which r indicates the strongest relationship between variables?
a) 0.80 b) -0.85 c) 0.25 d) -0.25

18. $\hat{Y} = 1.22 + 0.9X$, then correlation coefficient may be
a) -0.95 b) 1.95 c) 0.88 d) 0

19. r is -0.60 which of the following statement is correct?
a) The X variable explain 36% variability in Y.
b) The X variable explain 60% variability in Y.
c) The X variable explain 40% variability in Y.
d) The X variable explain -60% variability in Y.

20. If one variable tends to increase as other variable decrease then "r":
a) 0 b) 1 c) Positive d) Negative

21. The correlation between two variables is 0.75, it means
a) The average value of Y change by 0.75 when X is increased by 1.
b) The average value of X change by 0.75 when Y is increased by 1.
c) The X explains 75% of the variation in Y.
d) The X explains 56.25% of the variation in Y.

C

1. **The range of multiple correlation coefficients:**
a) 0 to +1 b) -1 to +1 c) < -1 d) > +1

2. **Multiple regression equation exists when independent variables are:**
a) uncorrelated b) correlated c) dependent d) interrelated

3. **For multiple regression analysis, $\Sigma(Y-\bar{Y})^2 = 50$ and $\Sigma(Y-\hat{Y})^2 = 20$ multiple coefficient of determination R^2 is:**
a) 0.70 b) 0.60 c) 0.50 d) 0.40

4. **Coefficient of multiple determination r^2:**
a) $\frac{SS\,Regression}{SSError}$ b) $\frac{SS\,Residual}{SSTotal}$ c) $\frac{SS\,Regression}{SSTotal}$ d) $\frac{SSTotal}{SSerror}$

5. **Correlation among the independent variables is termed**
a) homoscedasticity b) linearity
c) multi-collinearity d) adjusted coefficient of determination

6. **In a multiple regression model, the error term *e* is not assumed to**
a) Have a mean of 0 b) Have a variance of zero
c) Have a standard deviation of 1 d) Be normally distributed

7. **Purpose of multiple regression:**
a) to predict one response for one independent variable
b) to predict two responses for one independent variable
c) to predict one response for two various independent variables
d) to predict independent variable for one dependent variable.

8. **A model provides a poor fit:**
a) Sum of the squares of residual is large. b) Standard error of estimate is small
c) Sum of residual is zero d) Coefficient of determination is large.

9. **Standard error of regression measures:**
a) variability of independent variable relative to its mean
b) variability of dependent variable relative to its mean
c) variability of dependent variable relative to regression line
d) variability of independent variable relative to regression line

10. **Total variation explained by regression model is given by:**
a) intercept term b) coefficient of determination
c) regression coefficient d) partial regression coefficient

11. **If observed value is 45 and predicted value is 48 then error is**
a) 3 b) -3 c) 1.5 d) -1.5

12. **Correct relation between SSR, SSE and SST:**
a) SSR = SSE + SST b) SSE = SST + SSR
c) SST = SSR + SSE d) SST = SSR – SSE

D

1. **If 7X -10Y +87 = 0, then the values of "a" and "b":**
a) 8.7, 3 b) 8.7, 0.7 c) 8.7, 0.87 d) 0.7, 8.7

2. **Term is not appropriate for studying relationship between two variables**
(a) scatter plot (b) correlation (c) pie chart (d) regression

3. A variable Y behave exponentially:

a) $Y_{x+1} - Y_x$ is constant
b) $Y_{x+2} - Y_x$ is constant
c) Y_{x+1} / Y_x is constant
d) Y_{x+1} / Y_x is constant

4. Process of finding an approximating model:

a) Curve fitting b) correlation c) Dispersion d) Probability

5. Method of estimation that minimize sum of square residual

a) OLS b) MOM c) MLE d) Bayesian

6. Equations obtained by OLS estimation method:

a) Linear b) Nonlinear c) Normal d) Quadratic

7. If second difference of response is constant:

a) Linear b) exponential c) Power d) Quadratic

8. If X Changes by AP and Y change by GP:

a) Linear b) exponential c) Power d) Quadratic

9. If X and Y both changes by GP:

a) Linear b) exponential c) Power d) Quadratic

10. Exponential curve is also:

a) Power b) Logarithmic c) Semi Logarithmic d) Inverse

11. Power curve is also:

a) Linear b) Quadratic c) Semi Logarithmic d) Geometric

key

A

Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans
1	c	2	a	3	c	4	a	5	b	6	b
7	b	8	c	9	b	10	c	11	c	12	b
13	c	14	c	15	b	16	b	17	b	18	a
19	c	20	b	21	c	22	a	23	a	24	b
25	b	26	c	27	c	28	a	29	a	30	b
31	d	32	b	33	b	34	c	35	b	36	b
37	d	38	d	39	d	40	a	41	c	42	c

B

Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans
1	b	2	b	3	b	4	c	5	c	6	b
7	c	8	d	9	d	10	c	11	b	12	d
13	a	14	c	15	b	16	c	17	b	18	c
19	a	20	d	21	d						

C

Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans
1	a	2	a	3	b	4	c	5	c	6	b
7	c	8	a	9	c	10	b	11	b	12	c

D

Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans	Sr. No	Ans
1	b	2	c	3	d	4	a	5	a	6	c
7	d	8	b	9	c	10	c	11	d		

Exercise A

Q No. 10.1: (a) Explain the followings: Regression, linear regression, independent and dependent variable, deterministic and probabilistic model, standard error of estimate, scatter diagram and Ordinary method of least square.
(b) What is linear regression model? And explain the assumptions underlying the linear regression model.
(c) What are the properties of regression line obtained by OLS method?

Q No. 10.2: (a) Explain the following terms
Correlation, coefficient of correlation, coefficient of determination and rank correlation.
(b) Write down the properties of coefficient of correlation.

Q No. 10.3: Fit the regression line between wing lengths and age of sparrows. Interpret the value of b.

Age (days)	3.0	4.0	5.0	6.0	8.0	9.0	10.0
Wing length (cm)	1.4	1.5	2.2	2.4	3.1	3.2	3.2

Predict wing length of sparrow of age 7.0 days and 12.0 days.

Q No. 10.4: Fit the regression line between blood pressure and age of person. Interpret the value of regression coefficient.

Age(years)	30	40	50	60	70
Blood pressure (m/m Hg)	108	125	132	148	162

Estimate the blood pressure of a person having age 45 years.

Q No. 10.5: Fit the regression line between temperature and oxygen consumption. Interpret the value of regression coefficient.

Temperature (C^0)	-10	-5	0	5	10
Oxygen consumption (ml/g/hr)	4.5	3.6	3.4	3.1	2.7

Q No. 10.6: Compute correlation coefficient between wing length and tail length of a bird of a particular species. Interpret the value.

Wing length (cm)	10.4	10.8	10.2	10.7	11.2	11.4
Tail length (cm)	7.4	7.6	7.1	7.4	7.7	8.3

Q No. 10.7: Measurement of serum cholesterol (mg/100ml) and arterial calcium deposition (mg/ 100 g dry weight of tissue) were made on 12 animals. The data are as follows. Find correlation coefficient.

Calcium	52	42	59	24	40	32	57
Cholesterol	303	233	287	245	265	233	290

Q No. 10.8: Fit the regression line.

Nitrogen	0	10	20	30	40	50	60
Corn yield	18	25	40	45	70	65	100

Q No. 10.9: Yield of potatoes Y (pounds) and level of fertilizer application X (pounds), are given below.

X_i	1	1.5	2	2.5	3	3.5	4	4.5
Y_i	25	31	27	28	36	35	32	34

a) Construct scatter diagram
b) Is it linear trend?
c) How many pounds of potatoes would you expect from a plot to which 3.7 pounds of fertilizer has been applied.

Q No. 10.10: A breeder of horses wishes to model the relationship between gestation period and the length of life of horse. The breeder believes that two variables may follow a liner trend. Data is as follow;

Gestation period (days)	416	279	298	307	356	403	265
Life length (year)	24	25.5	20	21.5	22	23.5	21

According to your least square line approximately how long would you expect a horse to live whose gestation period was 400 days?

Q No. 10.11: Consider the model $\hat{Y} = 10.9 + 0.23X_i$; Y_i is foot length (cm) and X_i is height (inches) of a person.
(i) Estimate foot length of a person 70 inches tall.
(ii) Find error if his actual foot length is 28 cm.
(iii) Estimate the average foot length for a person 74 inches tall.

Q No. 10.12: Weight and systolic blood pressure of 8 males selected at random from a specific age group are given. Model a relationship between weight and blood pressure.

Weight	140	150	160	170	180	190	200	210
BP	122	120	135	150	155	155	148	160

Q No. 10.13: Fit the regression line between X: height of leaf above the ground (cm) and Y: concentration of leaf stomata/mm^2.

X_i	21.4	21.7	22.3	22.9	23.2	23.8	24.8	25.4	28.0
Y_i	4.5	4.4	4.6	4.7	4.5	4.4	4.5	4.2	3.4

Q No. 10.14: marks of students along with daily hours of study are given. Fit the regression line between these two variables and interpret the results also find correlation coefficient.

Study hours	1	1.5	2	2.5	3	3.5	4	4.5
Marks	52	55	60	60	64	70	80	92

Q No. 10.15: Calculate correlation co-efficient (r) for the following heights in inches of father (X) and their sons (Y). Also fit regression line.

X:	65	66	67	67	68	69	70	72
Y:	67	68	65	68	72	72	69	71

Q No. 10.16: Find regression lines and compute correlation co-efficient. And prove that correlation coefficient is geometric mean between both regression coefficients.

Husband's age	23	27	28	28	29	30	31	33	35	36
Wife's age	18	20	22	27	21	29	27	29	28	29

Q No. 10.17: Total sum of square and error sum of square that results from a regression equation fitted between length of right palm (dependent) and left palm (independent) of 50 persons selected at random from a population are 85.2 and 10.7 respectively. Find coefficient of determination and correlation coefficient between right palm and left palm and interpret both values.

Q No. 10.18: Find correlation coefficient between height of leaf above the ground (cm) and concentration of leaf stomata/mm^2.

X	21.4	21.7	22.3	22.9	23.2	23.8	24.8	25.4	28.0
Y	4.5	4.4	4.6	4.7	4.5	4.4	4.5	4.2	3.4

Q No. 10.19: Find correlation coefficient and rank correlation coefficient between weights of twins.

I	70.4	68.2	77.3	61.2	72.3	74.1	71.1
II	71.3	67.4	75.2	66.7	74.2	72.9	69.5

Q No. 10.20: Following data is about relative importance of eight ecological factors in the success of two particular species of bird in their habitat.

Factor	A	B	C	D	E	F	G	H
I	1	2	3	4	5	6	7	8
II	1	2	3	7	8	6	5	4

Find rank correlation coefficient.

Q No. 10.21: Plot the data in the form of scatterplot and then fit appropriate regression model between time and number of universities in Pakistan.

Year	2011	2012	2013	2014	2015	2016	2017
Universities	135	139	147	161	163	163	185

Q No. 10.22: Fit the regression line for corn yield on fertilizer

Fertilizer	0	10	20	30	40	50	60
Corn yield	18	30	37	48	72	82	95

[GCUF BS Computer Science and Urdu 2019]

Exercise B

Q No. 10.23: Define the followings
Multiple regression and correlation, curve fitting, exponential curve and geometric curve.

Q No. 10.24: The following means, SD and correlations are found for :
X_1 = Seed-hay crop in cwts per acre.; X_2 = Spring rainfall in inches.
X_3 = Accumulated temperature above 42 degree in spring in a certain district of Pakistan during 20 years.
$\bar{X}_1 = 28.02, S_1 = 4.42, r_{12} = 0.80;\ \bar{X}_2 = 4.91, S_2 = 1.10, r_{13} = -0.40;$
$\bar{X}_3 = 594, S_3 = 85, r_{23} = -0.56$
Find the partial correlation and regression equation of hay-crop on spring rain fall and accumulated temperature. **[STT-421 BS MATH GCUF]**

Q No. 10.25: An instructor of mathematics wish to determine the relationship of grade on a final examination to grade on two quizzes given during the semester. Calling X_1, X_2 and X_3 the grades of a student on the first quiz, second quiz and final examination respectively, he made the following computations for a total of 120 students.
$\bar{X}_1 = 6.8, S_1 = 1.0, r_{12} = 0.60;\ \bar{X}_2 = 7.0, S_2 = 0.8, r_{13} = 0.70;\ \bar{X}_3 = 74, S_3 = 9.0, r_{23} = 0.65$
Find the least squares regression equation of X_3 on X_1 and X_2. Also estimate the final grade of two students who scored respectively (1) 7 & 9 and (2) 5 & 7 on the two quizzes. **[STT-421 BS MATH GCUF 2018]**

Q No. 10.26: (i) From the data given below, fit the regression line of Y on X_1 and X_2.

Y	51	72	53	83	57	66
X_1	17	29	17	30	15	17
X_2	2	5	3	10	7	11

(ii) What is the predicted value of Y when $X_1 = 16$ and $X_2 = 4$.
(iii) Calculate standard error of estimate and coefficient of multiple determination.

Q No. 10.27: Find multiple coefficient of correlation $R_{1.23}$ and partial correlation coefficient $r_{12.3}$ When $r_{12} = 0.91$, $r_{13} = -0.82$ and $r_{23} = -0.81$.

Q No. 10.28: Find multiple coefficient of correlation $R_{3.12}$ and partial correlation coefficient $r_{13.2}$ When $r_{12} = 0.852$, $r_{13} = 0.917$ and $r_{23} = 0.890$.

Q No. 10.29: From the data given below from 10 observations, fit the regression line X_3 on X_1 and X_2
$\sum X_1 = 441; \sum X_2 = 147; \sum X_3 = 272,\ \sum X_1X_2 = 6485; \sum X_2X_3 = 4013; \sum X_1X_3 = 12005$
$\sum X_1^2 = 19461; \sum X_2^2 = 2173; \sum X_3^2 = 7428$

Q No. 10.30: From the data given below fit the regression line X_2 on X_1 and X_3
$\sum x_1x_2 = -33.5; \sum x_2x_3 = 42.5; \sum x_1x_3 = -47.83,\ \sum x_1^2 = 34.83; \sum x_2^2 = 53.5; \sum x_3^2 = 74.83$
$\bar{X}_1 = 6.3;\ \bar{X}_2 = 8.5;\ \bar{X}_3 = 11.7$

Exercise C

Q No. 10.31:

X_i	1	10	20	30	40
Y_i	1	100	400	600	1200

Plot the above data in the form of scatter plot. Judge the model through this plot and fit the same one.

Q No. 10.32: Fit a curve of the form $Y = ab^X$ to following data in which Y represents the number of bacteria per unit volume existing in a culture at the end of X hours. To support your answer make the graph.

X_i	1	2	3	4	5	6	7
Y_i	101	112	131	162	212	282	382

Q No. 10.33: Fit the straight line and quadratic equation and select the best one.

X_i	0	1	2	3	4
Y_i	10	13	20	30	35

Q No. 10.34: Fit the exponential equation of the form $Y = ae^{bx}$.

X_i	1	3	5	7	9	11
Y_i	8	13	10	14	20	23

Q No. 10.35: Fit the curve $Y = ab^X$.

X_i	1	2	3	4	5	6
Y_i	8	15	30	60	110	200

Q No. 10.36: Fit the curve $Y = ab^X$.

X_i	1.0	1.5	2.0	2.5	3.0	3.5
Y_i	5.5	5.35	5.25	5.17	5.13	5.09

Q No. 10.37: Fit the curve $Y = aX^b$ between weight and pulse rate of various mammals.

X_i	0.2	0.3	2	5	30	50
Y_i	420	300	205	120	85	70

Q No. 10.38: The table shows the atomic number X_i and the melting point Y_i (in degrees Celsius) for the alkali metals.

Alkali metal	Atomic number X	Melting point Y
Lithium	3	180.5
Sodium	11	97.8
Potassium	19	63.7
Rubidium	37	38.9
Cesium	55	28.5

a. Draw a scatter plot of Y_i versus X_i. Is a power model a good fit for the original data?

b. Find a power model for the original data.

c. One of the alkali metals, francium, is not shown in the table. It has an atomic number of 87. Using your model, predict the melting point of francium.s

Solution

Q#	Calculation and answers
10.3	$\Sigma X = 45, \Sigma Y = 17.1, \Sigma XY = 122.1, \Sigma X^2 = 331, \Sigma Y^2 = 44.99, \hat{Y} = 0.59 + 0.29X$
10.4	$\Sigma Y = 250, \Sigma Y = 675, \Sigma XY = 35060, \Sigma X^2 = 13500, \Sigma Y^2 = 92851,$ $\hat{Y} = 69.5 + 1.31X$
10.5	$\Sigma X = 0, \Sigma Y = 17.3, \Sigma XY = -20.5, \Sigma X^2 = 250, \Sigma Y^2 = 61.67, \hat{Y} = 3.46 - 0.082X$
10.6	$\Sigma X = 64.7, \Sigma Y = 45.5, \Sigma XY = 491.55, \Sigma X^2 = 698.73, \Sigma Y^2 = 345.87, r = 0.92$
10.7	$\Sigma X = 306, \Sigma Y = 1856, \Sigma XY = 82941, \Sigma X^2 = 14198, \Sigma Y^2 = 497106, r = 0.80$
10.8	$\Sigma X = 210, \Sigma Y = 383, \Sigma XY = 15450, \Sigma X^2 = 9100, \Sigma Y^2 = 26699,$ $\hat{Y} = 12.29 + 1.41X$
10.9	$\Sigma X = 22, \Sigma Y = 248, \Sigma XY = 707, \Sigma X^2 = 71, \Sigma Y^2 = 7800,$ $\hat{Y} = 24.45 + 2.38X$, (c) 33.26
10.10	$\Sigma X = 2324, \Sigma Y = 157.5, \Sigma XY = 52526.5, \Sigma X^2 = 793320, \Sigma Y^2 = 3565.75,$ $\hat{Y} = 18.89 + 0.011X$, 23.24
10.11	(i) 27, (ii) -1, (iii) 27.92
10.12	$\Sigma X = 1400, \Sigma Y = 1145, \Sigma XY = 202730, \Sigma X^2 = 249200, \Sigma Y^2 = 165563,$ $\hat{Y} = 45 + 0.56X$
10.13	$\Sigma X = 213.5, \Sigma Y = 39.2, \Sigma XY = 924.59, \Sigma X^2 = 5099.43, \Sigma Y^2 = 171.92,$ $\hat{Y} = 7.99 - 0.15X$
10.14	$\Sigma X = 22, \Sigma Y = 533, \Sigma XY = 1575.5, \Sigma X^2 = 71, \Sigma Y^2 = 36789,$ $\hat{Y} = 37.88 + 10.45X, r = 0.95$
10.15	$\Sigma X = 544, \Sigma Y = 552, \Sigma XY = 37560, \Sigma X^2 = 37028, \Sigma Y^2 = 38132,$ $\hat{Y} = 23.67 + 0.67X, r = 0.60$
10.16	$\Sigma X = 300, \Sigma Y = 250, \Sigma XY = 7623, \Sigma X^2 = 9138, \Sigma Y^2 = 6414,$ $\hat{Y} = -1.74 + 0.89X, r = 0.82, \hat{X} = 11.25 + 0.75Y$
10.17	$r^2 = 0.87, r = 0.93$
10.18	$\Sigma X = 213.5, \Sigma Y = 39.2, \Sigma XY = 924.59, \Sigma X^2 = 5099.43, \Sigma Y^2 = 171.92, r = -0.83$
10.19	$\Sigma X = 494.6, \Sigma Y = 497.2, \Sigma XY = 35219.2, \Sigma X^2 = 35101.44, \Sigma Y^2 = 35380.68,$ $d = 1, 0, 0, 0, 1, -1, -1, \Sigma d = 0, \Sigma d^2 = 4, r = 0.88, r_s = 0.93$
10.20	$d = 0, 0, 0, 3, 3, 0, -2, -4, \Sigma d = 0, \Sigma d^2 = 38, r_s = 0.55$
10.21	$\Sigma X = 21, \Sigma Y = 1093, \Sigma XY = 3493, \Sigma X^2 = 91, \Sigma Y^2 = 172439$ Taking 2011 as 0(origin), $\hat{Y} = 133.2 + 7.6X$.
10.22	$\Sigma X = 210, \Sigma Y = 382, \Sigma XY = 15160, \Sigma X^2 = 9100, \Sigma Y^2 = 25830, \hat{Y} = 14.93 + 1.32X$
10.24	$r_{123} = 0.76, r_{213} = -0.436, r_{312} = 0.097, \hat{X}_1 = 9.22 + 3.37X_2 + 0.004X_3$
10.25	(1) 83, (2) 66, $\hat{X}_3 = 16.07 + 4.36X_1 + 4.04X_2$
10.26	(i) $\hat{y} = 24.36 + 1.37X_1 + 1.71X_2$, (ii) 53, (iii) $S_{12} = 0.54$ and $R^2_{y.12} = 1$

10.27	$R_{1.23} = 0.92, r_{12.3} = 0.73$				
10.28	$R_{1.12} = 0.94, r_{12.3} = 0.664$				
10.29	$\Sigma x_1x_2 = 2.3, \Sigma x_2x_3 = 9.8, \Sigma x_2x_3 = 14.6, \Sigma x_1^2 = 12.9, \Sigma x_2^2 = 12.1,$ $\hat{X}_1 = -13.67 + 0.56X_2 + 1.10X_3$				
10.30	$\hat{X}_2 = 22.33 - 1.49X_1 - 0.38X_3$				
10.31	Scatter plot of the data is given at the end, shows that trend is in power form. $\hat{Y} = 1.07X^{1.92}$				
10.32	$\hat{Y} = 71.84(1.25)^X$	10.33	$\hat{Y} = 8.2 + 6.7X, \hat{Y} = 9.2 + 4.7X + 0.5X^2$	10.38	$\hat{Y} = 397.61X^{-0.64}$, 22.92
10.34	$\hat{Y} = 7.59e^{0.099X}$	10.35	$\hat{Y} = 4.21(1.92)^X$		
10.36	$\hat{Y} = 5.62(0.97)^X$	10.37	$\hat{Y} = 230.5X^{-0.31}$		

Chapter 11

ANOVA AND EXPERIMENTAL DESIGNS

Sir Ronald A. Fisher is pioneer of experimental design. He developed it for agricultural experiment.

11.1 Analysis of variance (ANOVA)

Partitioning the total variation of a set of observations into parts due to particular factors, and comparing variances (mean squares) by way of F-test, so that differences between means can be assessed is known as analysis of variance.

11.2 One-way and two-way ANOVA

The one-way ANOVA, in which "n" subjects are allocated, usually at random, to the different levels of a single factor/treatments.

The two-way ANOVA, in which "n" subjects are allocated, usually at random, to the different levels of two factors/treatments.

11.3 Testing of hypothesis about equality of more than two population means:

Hypothesis: $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

H_1: At least two means are not equal

Formulae for calculation:

X_{01}	X_{02}	..	X_{0k}	
X_{11}	X_{21}	..	X_{k1}	
X_{12}	X_{22}	..	X_{k2}	
.	.	.	.	
.	.	.	.	
.	.	.	.	
X_{1r}	X_{2r}	..	X_{kr}	
T_1	T_2	..	T_k	T

r = No. of observations in each sample and k = No. of samples / treatments

$CF = \frac{T^2}{n}$ (CF stands for correction factor)

$Total\ SS = \sum_{i=1}^{k}\sum_{j=1}^{r} X_{ij}^2 - CF$ (SS stands for sum of square)

$Between\ SS = \frac{\sum_{i=1}^{k} T_i^2}{r} - CF$

$Within\ SS = Total\ SS - Between\ SS$

ANOVA Table

S.O.V	df	SS	MS	F-ratio	F-table
Between treatments	$k-1$	$\frac{\sum T_i^2}{r} - CF$	$S_b^2 = \frac{SS}{k-1}$	$F = \frac{S_b^2}{S_w^2}$	$F_{\alpha(k-1,\, n-k)}$
Within/ Error	$n-k$	By difference	$S_w^2 = \frac{SS}{n-k}$		
Total	$n-1$	$\sum\sum X_{ij}^2 - CF$			

Conclusion: If F-ratio > F-table then H_0 may not be accepted, otherwise accepted.

Example 11.1: following data is about damage plants out of 600 samples, to different crops by porcupine in three different sides with respect to LBD canal. Test the hypothesis that damage in all three areas are same.

Left Side	Right Side	Middle
129	301	292
14	26	19
43	166	222
62	232	263

Solution:

$H_0: \mu_L = \mu_R = \mu_M$

H_1: At least average damage in two sides are different

Left Side	Right Side	Middle	
129	301	292	
14	26	19	
43	166	222	
62	232	263	
T_i	248	725	796	$T = 1769$

$$CF = \frac{(T)^2}{12} = \frac{(1769)^2}{12} = 260780.1$$

$$TSS = \sum\sum Y_{ij}^2 - CF = 399265 - 260780.1 = 138484.9$$

$$Between\ SS = \frac{\sum T_i^2}{r} - CF = \frac{(248)^2 + (725)^2 + (796)^2}{4} - 260780.1$$

$$= 44406.15$$

$$Error\ SS = Total\ SS - Between\ SS = 94078.75$$

ANOVA

S.O.V	SS	df	MS	F	Critical value of F
Between sides	44406.15	2	22203.08	2.1240	4.2564
Error	94078.75	9	10453.19		
Total	138484.9	11			

As calculated value of F is 2.1240 that is less than critical value of F, so we may accept $H_0: \mu_L = \mu_R = \mu_M$.

11.4 Experiment

An experiment is a process of conducting a test or series of test that observes the change (variation) in response by changing the input variables sequentially.

11.5 Experimental design

Experimental design is a procedure to assign the treatments to the experimental units in such a way that the valid inference about the effects of these treatments can be drawn. It is also called design of experiment or designed experiment.
It is an integral part of almost all scientific investigations and researches.

11.6 Terminology

11.6.1 Response: Response is a result obtained from experimental unit after the application of a specific treatment.

11.6.2 Factor: A factor is a parameter or entity which is suspected to have influence on the response.

11.6.3 Treatment: A treatment / level is the setting of the factor at a particular category / value, whose effects is to be measured.

11.6.4 Experimental units: The smallest division of experimental materials for which treatments are applied are called experimental units.

11.7 Layout of the experiment

A plan to assign the treatments to the experimental units in an experimental design is called layout of the experiment.

11.8 Experimental Error: The difference among the results if the experiment is repeated several times by same treatment is known as experimental error.

11.8.1 Causes:

1) It is caused by the uncontrollable factors.
2) Use of inappropriate design of experiment
3) Incorrect and unclear measurements.

11.9 Blocking: Collection of similar experimental units is called blocking.

11.10 Format of the experimental design

Factor			
Treatment/level 1	Treatment/level 2	.....	Treatment/level k
.	.	.	.
.	.	.	.
.	Response	.	.
.	.	.	.
.	.	.	.

Replications

11.11 Necessity for an experimental design

An experiment is planned

1) To get maximum information in minimum cost and time.
2) To avoid systematic errors.
3) To evaluate the outcomes critically and logically.
4) To ignore spurious effect, if any
5) To identify the relationship between cause and effect.
6) To facilitate the understanding of interactions among the factors.
7) To optimize the performance of experiment by fixing the levels of factor.
8) To minimize the experimental error.

11.12 Basic Principles of Experimental Design

There are three basic principles of experimental designs

(i) Randomization (ii) Replication (iii) Local control

11.12.1 Randomization

A process of assigning the treatments / levels randomly to the experimental units is known as randomization. Through randomization, every experimental unit will have the same chance of receiving any treatment. It is helpful in reducing experimental error.

11.12.2 Replication

Replication is the repetition of experiment under identical conditions. The single replication is called a replicate. The greater the number of replications, greater is the precision in the experiment. In short it serves two objectives; (i) experimental error may be reduced and (ii) more precise estimate of average effect is obtained.

11.12.3 Local control

When all extraneous sources of variation are not removed by randomization and replication then we use local control, which involves balancing, blocking and grouping of the experimental units for the refinement in experimental techniques.

11.13 Basic Experimental designs

There are three basic experimental designs.

1) Completely Randomized (CR) Design
2) Randomized Complete Block (RCB) Design
3) Latin Square (LS) Design

11.13.1 Completely Randomized (CR) Design

A completely randomized (CR) design is the simplest experimental design, in terms of data analysis and convenience. A completely randomized design (CRD) is one for which the treatments are assigned completely at random to experimental units. In this design each experimental unit has the same chance of receiving each treatment. In case of experiment involved one factor, CR design is used.

11.13.1.1 Layout

The experimental layout for a CR design may be done using random number table. For example CR design's layout for four treatments A, B, C and D may look like

C	A	B	D	E
C	B	C	A	B
A	D	D	B	E
E	C	A	D	B

11.13.1.2 Statistical model

Let Y_{ij} is the response of i^{th} observation on j^{th} treatment, then the linear model for CR design is $Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$ $\begin{cases} i = 1,2,3,\ldots,r \\ j = 1,2,3,\ldots,k \end{cases}$

Where

μ = over all mean

τ_j = effect of j^{th} treatment

ε_{ij} = error term and $\varepsilon_{ij} \sim N(0, \sigma^2)$

11.13.1.3 Statistical analysis

To test $H_0: \tau_j = 0$ against $H_1: \tau_j \neq 0$ ANOVA table is as follows

S.O.V	df	SS	MS	F-ratio
Treatments	$k-1$	$\frac{\sum_{j=1}^{k} T_j^2}{r} - CF$	$S_b^2 = \frac{SS}{k-1}$	$F = \frac{S_b^2}{S_w^2}$
Error	$n-k$	By difference	$S_w^2 = \frac{SS}{n-k}$	
Total	$n-1$	$\sum_{i=1}^{r}\sum_{j=1}^{k} X_{ij}^2 - CF$		

11.13.1.4 Critical Region and Conclusion:

Critical region for these particular hypotheses is $F > F_{(\alpha, k-1, n-k)}$

Example 11.2: Perform the complete analysis of the following data obtained by CR design.

Factor			
T_1	T_2	T_3	T_4
3	4	10	9
5	3	9	2
7	1	11	5
6	5	7	1
9	2	8	3

Solution:

$H_0: \tau_j = 0$ $H_1: \tau_j \neq 0$ OR

H_0: *All treatments means are equal* H_1: *At least two treatments means are not equal*

	Factor				
	T_1	T_2	T_3	T_4	
	3	4	10	9	
	5	3	9	2	
	7	1	11	5	
	6	5	7	1	
	9	2	8	3	
T_j	**30**	**15**	**45**	**20**	$T = 110$

$$CF = \frac{T^2}{n} = \frac{(110)^2}{20} = 605$$

$$TSS = \sum_i \sum_j X_{ij}^2 - CF = 3^2 + 4^2 + \ldots + 3^2 - 605 = 185$$

$$Tr.SS = \frac{\sum T_j^2}{r_j} - CF = \frac{(30)^2}{5} + \ldots + \frac{(20)^2}{5} - 605 = 105$$

$$Error.SS = T.SS - Tr.SS$$

ANOVA table

S.O.V	df	SS	MS	F-ratio
Treatments	3	105	35	7.00
Error	16	80	5	
Total	19	185		

Critical value: $F_{(\alpha,f_1=k-1,f_2=4)} = F_{(5\%,3,9)} = 3.24$

Critical Region: $F > 3.24$

As calculate value of F (7.00) lies in CR so we may reject H_0.

11.13.1.5 Advantages of CR design

1) Its analysis is very simple.
2) Its layout is very easy.
3) Any number of treatments and replication may be used.
4) Degree of freedom for error sum of square is maximum.

11.13.1.6 Disadvantages of CR design

1) It is useful when experimental units are homogeneous.
2) Maximum variation of experimental units may enter into error sum of square due to not restricted randomization in any direction.

11.13.2 Randomized complete block (RCB) design

The randomized complete block design (RCBD) is one in which blocks are of equal size, each of which contains all the treatments. It is the most widely used experimental designs in forestry research, especially in field experiments where the number of treatments is not large. In case of experiment involved two factor, RCB design is used.

11.13.2.1 Layout:

The experimental layout for a RCB design for four treatments A, B, C and D may look like

Block I	C	A	B	D
Block II	B	C	D	A
Block III	D	A	B	C

11.13.2.2 Statistical model

Let Y_{ij} is the response of i^{th} observation on j^{th} treatment, then the linear model for CR design is $Y_{ij} = \mu + B_i + \tau_j + \varepsilon_{ij} \begin{cases} i = 1,2,3,\ldots,r \\ j = 1,2,3,\ldots,k \end{cases}$

where

μ = over all mean

B_i = effect of i^{th} block

τ_j = effect of j^{th} treatment

ε_{ij} = error term and $\varepsilon_{ij} \sim N(0, \sigma^2)$

11.13.2.3 Statistical analysis

To test $H_0: \tau_j = 0$ against $H_1: \tau_j \neq 0$ and $H_0': B_i = 0$ against $H_1': B_i \neq 0$

ANOVA table is as follows

S.O.V	Df	SS	MS	F-ratio
Treatments	$k-1$	$\frac{\sum_{j=1}^{k} T_j^2}{r} - CF$	$S_t^2 = \frac{SS}{k-1}$	$F = \frac{S_t^2}{S_e^2}$
Blocks	$r-1$	$\frac{\sum_{i=1}^{r} T_i^2}{k} - CF$	$S_B^2 = \frac{SS}{k-1}$	$F = \frac{S_B^2}{S_e^2}$
Error	$(k-1)(r-1)$	By difference	$S_e^2 = \frac{SS}{(k-1)(r-1)}$	
Total	$rk-1$	$\sum_{i=1}^{r}\sum_{j=1}^{k} X_{ij}^2 - CF$		

11.13.2.4 Critical Region and Conclusion:

Critical region for hypothesis

(i) $H_0: \tau_j = 0$ is $F > F_{(\alpha, k-1, (k-1)(r-1))}$ and

(ii) $H_0': B_i = 0$ is $F > F_{(\alpha, r-1, (k-1)(r-1))}$

Example 11.3: perform the complete analysis of the following data obtained by RCB design.

Blocks	Treatments			
	T_1	T_2	T_3	T_4
B_1	8	5	16	10
B_2	13	7	15	7
B_3	9	6	11	4

Solution:

$H_0: \tau_j = 0$ and $H_1: \tau_j \neq 0$ OR

H_0: *All treatments means are equal*

H_1: *All treatments means are not equal*

Blocks	Treatments				
	T_1	T_2	T_3	T_4	$T_{i.}$
B_1	8	5	16	10	**39**
B_2	13	7	15	7	**42**
B_3	9	6	11	4	**30**
$T_{.j}$	**30**	**18**	**42**	**21**	$T_{..} = 111$

$$CF = \frac{T_{..}^2}{n} = \frac{(111)^2}{12} = 1026.75$$

$$TSS = \sum_i \sum_j X_{ij}^2 - CF = 8^2 + 5^2 + \ldots + 4^2 - 1026.75 = 164.25$$

$$Tr.SS = \frac{\sum T_{.j}^2}{r_j} - CF = \frac{(30)^2}{3} + \ldots + \frac{(21)^2}{3} - 1026.75 = 116.25$$

$$Blocks\ SS = \frac{\sum T_{i.}^2}{b_i} - CF = \frac{(39)^2}{4} + \ldots + \frac{(30)^2}{4} - 1026.75 = 19.5$$

$$Error\ SS = T.SS - Tr.SS - Block\ SS$$

S.O.V	df	SS	MS	F-ratio
Treatments	3	116.25	38.75	8.158
Blocks	2	19.5	9.75	2.053
Error	6	28.5	4.75	
Total	11	164.25		

Critical value: $F_{(\alpha,k-1,(k-1)(r-1))} = F_{(0.05,3,6)} = 4.76$

$$F_{(\alpha,r-1,(k-1)(r-1))} = F_{(0.05,2,6)} = 5.14$$

Critical Region: $F_1 > 4.76$; $F_2 > 5.14$

As calculated value of F_1 (8.158) for treatments lies in CR so we may not accept H_0, and F_2 (2.053) for blocks does not lie in CR so we may accept H_0.

11.13.2.5 Advantages of RCB design

1) Its analysis is very simple.
2) Its layout is very easy.
3) Any number of treatments and replication may be used in this design.
4) Missing observations is easily adjustable.
5) Blocking reduced experimental error.
6) It is most frequently used experimental design.

11.13.2.6 Disadvantages of RCB design

1) It is used when experimental units are homogeneous.
2) Randomization is restricted in only one direction.

11.13.3 Latin square (LS) design

Latin square designs differ from randomized complete block designs in that the experimental units are grouped in blocks in two different ways, that is, by rows and columns. In case of experiment involved three factors, LS design is used.

11.13.3.1 Layout:

The experimental layout for a LS design for four treatments A, B, C and D may look like

Rows	Columns			
	I	II	III	IV
I	A	B	C	D
II	B	C	D	A
III	C	D	A	B
IV	D	A	B	C

11.13.3.2 Statistical model:

Let Y_{ij} is the response of i^{th} observation on j^{th} treatment, then the linear model for CR design is $Y_{ijh} = \mu + R_i + C_j + \tau_h + \varepsilon_{ijh}$ $\begin{cases} i = 1,2,3,\ldots,k \\ j = 1,2,3,\ldots,k \\ h = 1,2,3,\ldots,k \end{cases}$

Where

μ = over all mean; R_i = effect of i^{th} row

C_j = effect of j^{th} column τ_h = effect of h^{th} treatment

ε_{ijh} = error term and $\varepsilon_{ijh} \sim N(0, \sigma^2)$

11.13.3.3 Statistical analysis:

To test $H_0: \tau_h = 0$ against $H_1: \tau_h \neq 0$, $H_0': R_i = 0$ against $H_1': R_i \neq 0$ and $H_0'': C_j = 0$ against $H_1'': C_j \neq 0$

ANOVA table is as follows

S.O.V	df	SS	MS	F-ratio
Treatments	$k-1$	$\frac{\sum_{h=1}^{k} T_h^2}{k} - CF$	$S_T^2 = \frac{SS}{k-1}$	$F = \frac{S_T^2}{S_e^2}$
Rows	$k-1$	$\frac{\sum_{i=1}^{k} T_i^2}{k} - CF$	$S_R^2 = \frac{SS}{k-1}$	$F = \frac{S_R^2}{S_e^2}$
Columns	$k-1$	$\frac{\sum_{j=1}^{k} T_j^2}{k} - CF$	$S_C^2 = \frac{SS}{k-1}$	$F = \frac{S_C^2}{S_e^2}$
Error	$(k-1)(k-2)$	By Difference	$S_e^2 = \frac{SS}{(k-1)(k-2)}$	
Total	k^2-1	$\sum\sum\sum x^2 - CF$		

11.13.3.4 Critical Region and Conclusion

Critical region for hypothesis

(i) $H_0: \tau_h = 0$ is, $F > F_{(\alpha, k-1, (k-1)(k-2))}$

(ii) $H_0': R_i = 0$ is $F > F_{(\alpha, k-1, (k-1)(k-2))}$ and

(iii) $H_0'': C_j = 0$ is $F > F_{(\alpha, k-1, (k-1)(k-2))}$

Example 11.4: Perform the complete analysis of the following data obtained by LS design.

Rows	Columns			
	C_1	C_2	C_3	C_4
R_1	7 A	3 B	5 C	9 D
R_2	4 B	1 C	7 D	8 A
R_3	6 C	0 D	8 A	10 B
R_4	3 D	8 A	4 B	13 C

Solution:

$H_0: \tau_i = 0$ $H_1: \tau_i \neq 0$ OR

H_0: All treatments means are equal H_1: At least two treatments means are not equal

$H_0': C_j = 0$ $H_1': C_j \neq 0$ OR

H_0': All Columns means are equal H_1': At least two columns means are not equal

$H_0'': R_i = 0$ $H_1'': R_i \neq 0$ OR

H_0'': All rows means are equal H_1'': At least two rows means are not equal

Rows	Columns				
	C_1	C_2	C_3	C_4	$T_{i.}$
R_1	7 A	3 B	5 C	9 D	24
R_2	4 B	1 C	7 D	8 A	20
R_3	6 C	0 D	8 A	10 B	24
R_4	3 D	8 A	4 B	13 C	28
$T_{.j}$	20	12	24	40	$T_{..} = 96$

$$CF = \frac{T_{..}^2}{r^2} = \frac{(96)^2}{16} = 576,\ TSS = \sum_i \sum_j X_{ij}^2 - CF = 7^2 + 3^2 + \ldots + 13^2 - 576 = 176$$

$$CSS = \frac{\sum T_{.j}^2}{r} - CF = \frac{(20)^2}{4} + \ldots + \frac{(40)^2}{4} - 576 = 104$$

$$RSS = \frac{\sum T_{i.}^2}{r} - CF = \frac{(24)^2}{4} + \ldots + \frac{(28)^2}{4} - 576 = 8$$

A	B	C	D
31	21	25	19

$$Tr.SS = \frac{\sum T_{i.}^2}{r} - CF = \frac{(31)^2}{4} + \ldots + \frac{(19)^2}{4} - 576 = 21$$

S.O.V	df	SS	MS	F-ratio
Treatments	3	21	7.000	0.97
Rows	3	8	2.667	0.37
Columns	3	104	34.667	4.83
Error	6	43	7.167	
Total	15	176	176	

ritical value: $F_{(\alpha, \lambda-1, n-\lambda)} = F_{(0.05,3,6)} = 4.76$

ritical Region: $F_1 > 4.76$; $F_2 > 4.76$; $F_3 > 4.76$

s calculate value of F_1 (0.97) for treatments does not lie in CR so we may accept H_0, F_2(0.372) for rows does not lie in CR so we may accept H_0' and F_3 (4.83) for columns es in CR so we may not accept H_0''.

11.13.3.5 Advantages of LS design

1) Its analysis is very simple.
2) It is more efficient than CR and RCB designs.
3) It provides efficient results for 5 to 10 treatments.
4) Two way blocking results in reduction of experimental error.

11.13.3.6 Disadvantages of LS design

1) It is less flexible than RCB and CR designs.
2) Replication is costly.
3) For small number of treatments it does not provide sufficient number of replicates.
4) Its layout is difficult for experimentation in agriculture.

11.14 The Least significant difference (LSD) test

A procedure used for pairwise comparison of treatments means with equal size, when H_0 is not accepted, is known as least significant difference (LSD) test.

Example 11.5: In example 11.3 hypothesis about equality of treatments is not accepted. Means for these treatments are $\bar{T}_1=10;\bar{T}_2=6;\bar{T}_3=14;\bar{T}_4=7$. Check which treatments are different by least significant difference (LSD) method?

Solution: $\bar{T}_1=10;\bar{T}_2=6;\bar{T}_3=14;\bar{T}_4=7$

Arrange the means in ascending order of magnitude

$\bar{T}_2$	$\bar{T}_4$	$\bar{T}_1$	$\bar{T}_3$
6	7	10	14

$LSD=t_{(\alpha/2,\nu)}\sqrt{\frac{2MSE}{r}}$; where MSE is mean square error, r= Number of observations in treatments under consideration and ν is degree of freedom of error.

$$LSD=t_{(\alpha/2,\nu)}\sqrt{\frac{2MSE}{r}}=t_{(0.025,8)}\sqrt{\frac{2(4.75)}{3}}=2.447(1.7795)\Rightarrow LSD=4.35$$

Test the difference between them, if the difference is less than LSD, means are not significant (in other words means are equal)

$\bar{T}_2$	$\bar{T}_4$	$\bar{T}_1$	$\bar{T}_3$
6	7	10	14

Those pair of treatments which have line under them are same. According to this

Same pairs	Different pairs
T_1,T_2	T_2,T_3
T_3,T_4	T_4,T_3
T_1,T_4	
T_1,T_3	

11.14.1 Conditions for the use of LSD test

We should use LSD test only when number of observations are same for each treatment and null hypothesis for equal means is rejected.

Multiple Choice Questions

1) If you were testing, how well different dish soaps cleaned grease, the independent (experimental) variable would be the:

(a) Dishes (b) water (c) dish soap (d) types of grease

2) F-ratio in CR design is

(a) $\frac{MST}{MSE}$ (b) $\frac{MSE}{MST}$ (c) $\frac{MSTr}{MST}$ (d) $\frac{MSTr}{MSE}$

3) The critical value of F test with 8 numerator and 6 denominator degrees of freedom at 0.05 level of significance is

(a) 3.58 (b) 4.88 (c) 4.15 (d) None of these

4) The ANOVA is for testing whether or not the

(a) means of two samples are equal (b) means of more than two samples are equal
(c) Proportions of two populations are equal (d) Independence of attributes.

5) There are 5 levels of a factor, where each level contains 9 observations. The degrees of freedoms for the critical value of F are

(a) 5 numerator and 9 denominator degrees of freedom
(b) 4 numerator and 8 denominator degrees of freedom
(c) 45 degrees of freedom (d) 4 numerator and 40 denominator degrees of freedom

6) In the ANOVA, treatment refers to

(a) experimental units (b) different levels of a factor
(c) a factor (d) none of the above

7) The mean square is the sum of squares divided by

(a) the total number of observations (b) its corresponding degrees of freedom - 1
(c) its corresponding degrees of freedom (d) none of the above

8) An experimental design where the experimental units are randomly assigned to the treatments is known as

a) factor block design b) random factor design
c) completely randomized design d) none of the above

9) In ANOVA with 4 groups and total samples 44, df for total sum of square is
a) 3 b) 40 c) 43 d) 44

10) ANOVA is used a for testing:
a) Equality of two sample means b) Equality of more than two sample means
c) equality of more than two population means d) independence of attributes

11) In ANOVA with 4 groups and total samples 44, df for between groups SS is
a) 3 b) 40 c) 43 d) 44

12) In ANOVA with 4 groups and total samples 44, df for within SS is
a) 3 b) 40 c) 43 d) 44

13) If MSE decreases then, value of F will
a) decrease b) increase c) no effect d) None

14) If SST = 150 AND SSTr = 110 then SSE = -----
a) - 40 b) 40 c) 260 d) 130

15) "factor" refers to
a) independent variable b) dependent variable
c) different levels of a treatment d) critical value of F

16) In ANOVA with 3 treatments with 10 observations in each treatment Error sum of square is 400, then MSE = ---
a) 13.79 b) 14.81 c) 44.44 d) 133.33

17) In ANOVA with 3 treatments with 10 observations in each treatment df for numerator is
a) 2 b) 9 c) 27 d) 29

18) In ANOVA with 3 treatments with 10 observations in each treatment df for denominator is
a) 2 b) 9 c) 27 d) 29

19) In ANOVA with 4 treatments with 8 observations in each treatment critical value of F is
a) 2.947 b) 8.623 c) 3.838 d) 4.347

20) In ANOVA procedure ------ test is used
a) t b) Z c) chi-square d) F

21) which component of ANOVA is not additive
a) SS b) df c) MS d) F ratios

22) In two way ANOVA df for error SS is
a) r-1 b) c-1 c) rc-1 d) $(r-1)(c-1)$

23) In one way ANOVA df for error SS is
a) k-1 b) n-1 c) nk-1 d) $n-k$

24) Two way ANOVA involves ------ independent factors
a) 1 b) 2 c) 3 d) 4

25) One way ANOVA involves ------ independent factors
a) 1 b) 2 c) 3 d) 4

26) ANOVA is based on
a) mean ratio b) probability ratio c) variance ratio d) error ratio

27) In one way ANOVA number of F ratios are
a) 1 b) 2 c) 3 d) 4

28) In two way ANOVA number of F ratios are
a) 1 b) 2 c) 3 d) 4

29) In experimental design different categories of an independent variable
a) factors b) levels c) blocks d) experimental units

30) Error SS in CR design(one way ANOVA) is given as
a) SSTotal + SSTreatment b) SSTotal – SSTreatment
c) SSTreatment + SSTotal d) SSTreatment - SSTotal

key

Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans
1	c	2	d	3	a	4	b	5	d	6	b
7	c	8	c	9	c	10	c	11	a	12	b
13	b	14	b	15	a	16	b	17	a	18	c
19	b	20	d	21	c	22	d	23	d	24	b
25	a	26	c	27	a	28	b	29	b	30	b

Exercise

Q No. 11.1: (a) Define the experimental design and explain the basic principles of an experimental design. (b) Discuss CR, RCB and LS design.

Q No. 11.2: Layout CR, RCB and LS design for testing five seeds variety, also make ANOVA table for these cases.

Q No. 11.3: Perform analysis to the following data obtained by the experiment conducted by Completely randomized (CR) design

FACTOR			
T_1	T_2	T_3	T_4
5	7	13	7
8	9	10	10
10	15	14	11
9	6	12	15
4	11	9	14

Q No. 11.4: Perform analysis to the following data obtained by the experiment conducted by Completely randomized (CR) design

FACTOR				
T_1	T_2	T_3	T_4	T_5
2.5	3.3	3.1	2.9	3.8
3.3	3.5	3.1	3.6	3.0
2.5	3.7	3.5	3.7	3.5
3.3	4.1	2.9	3.3	
3.5		3.4	3.4	
		2.9		

Q 11.5: Perform analysis to the following data obtained by the experiment conducted by Randomized Complete Block (RCB) design

FACTOR II	FACTOR I			
	T_1	T_2	T_3	T_4
T_1	50	30	30	60
T_2	65	60	75	45
T_3	80	90	90	70

Q No. 11.6: Complete this ANOVA table and answer the questions given at the end.

One-Way ANOVA Summary

Source	SS	df	MS	F
Between		3		
Within	23.5			
Total	55.2	19		

(i) Find sample size.
(ii) How many treatments are compared?
(iii) Test null hypothesis of equal means against alternative that there is difference in means of treatments.

Q No. 11.7: Following data is about porcupine burrows density in three different areas with respect to LBD canal and railway track. Test for their equality on three sides.

Left	Right	Middle
4	2	4
2	1	5
1	1	2
2	2	3
1	1	4
3	1	
	1	

(Data by Jahan Zeb Ijaz under supervision of Dr. Muhammad Wajid)

Q No. 11.8: Perform analysis to the following data obtained by the experiment conducted by Latin square (LS) design. A, B, C, D and E are treatments.

Rows	Columns				
	1	2	3	4	5
1	10.5 A	11 B	13 C	13 D	12 E
2	12 B	15 C	12.5 D	9 E	10 A
3	12.5 C	11 D	8.5 E	14 A	14 B
4	10.5 D	12 E	15 A	14 B	13 C
5	11 E	13 A	11.5 B	13.5 C	13.5 D

Q No. 11.9: apply Least significant difference (LSD) test, in case if hypothesis of equality between treatments is rejected.

FACTOR			
T_1	T_2	T_3	T_4
2.7	3.6	3.3	4.4
3.7	2.7	2.7	2.6
3.5	2.5	2.6	3.2
3.7	3.3	2.2	3.6
2.8	3.1	2.4	3.4
3.4	2.3	2.1	3.2

Q No. 11.10: Perform analysis to the following data about effect of different levels of ascorbic acid on the growth of plant (shoot length) obtained by the experiment conducted by Abdul Razzaq in his MPhil thesis from University of Okara 2016, through Completely randomized (CR) design. Also apply LSD test to determine which level produced different result to others.

Ascorbic acid levels			
0	0.1mM	0.5mM	1.0mM
48.1	50.9	60.6	70.5
51.3	54.7	63.2	73.4
52.9	56.8	65.1	75.1
54.2	61.4	67.3	79.3
56.0	58.7	67.8	74.2

Q No. 11.11: An organizational psychologist was interested in whether individuals working in different sectors of a company differed in their attitudes toward the company. The results for the three people surveyed in development were 10, 12, and 11; for the three in the marketing department, 6, 6, and 8; for the three in accounting, 7, 4, and 4; and for the three in production, 14, 16, and 13 (higher numbers mean more positive attitudes). Was there a significant difference in attitude toward the company among employees working in different sectors of the company at the .05 level?

Solution

Q No. 11.3:

Totals	
T_1	36
T_2	48
T_3	58
T_4	57
T	199
CF=1980.05	
$\Sigma\Sigma X_{ij}^2$=2179	

ANOVA table						
S.O.V	df	SS	MS	F_{ratio}	F_{table}	Conclusion
Factors	3	62.55	20.85	2.446	3.239	Accept H_0,
Error	16	136.4	8.525			
Total	19	198.95				

Q No. 11.4:

Totals	
T_1	15.1
T_2	14.6
T_3	18.9
T_4	16.9
T_5	10.3
T	75.8
CF=249.81	
$\Sigma\Sigma X_{ij}^2$=253.22	

ANOVA table						
S.O.V	df	SS	MS	F_{ratio}	F_{table}	Conclusion
Factors	4	1.10	0.275	2.149	2.928	Accept H_0,
Error	18	2.31	0.128			
Total	22	3.41				

Q 11.5:

Totals			
T_1	195	T_1	170
T_2	180	T_2	245
T_3	195	T_3	330
T_4	175		
T	745		
CF=46252.08			
$\Sigma\Sigma X_{ij}^2$=50875			

ANOVA table						
S.O.V	df	SS	MS	F_{ratio}	F_{table}	Conclusion
Factor I	3	106.25	35.41	0.162	4.757	Accept H_0,
Factor II	2	3204.17	1602.08	7.324	5.143	Reject H_0,
Error	6	1312.15	218.75			
Total	11	4622.92				

Q No. 11.6:

Source	SS	df	MS	F
Between	***31.7***	3	***10.567***	***7.193***
Within	23.5	***16***	***1.469***	
Total	55.2	19		

(i) Find sample size. $n = 20$

(ii) How many treatments are compared? **There are 4 treatments**

(iii) Test null hypothesis of equal means against alternative that there is difference in means of treatments. **Critical value of F against degree of freedoms 3 and 16 is 3.24. As value of F = 7.193 is greater than critical value F = 3.24, so we may not accept H_0.**

Q No. 11.7:

Totals	
T_1	13
T_2	9
T_3	18
T	40
CF=88.89	
$\Sigma\Sigma X_{ij}^2$=118	

ANOVA table						
S.O.V	.Df	SS	MS	F_{ratio}	F_{table}	Conclusion
Factors	2	15.65	7.825	8.72	3.682	Reject H_0,
Error	15	13.46	0.897			
Total	17	29.11				

Q No. 11.8:

Totals			
T_1	56.5	T_1	59.5
T_2	62	T_2	58.5
T_3	60.5	T_3	60
T_4	63.5	T_4	64.5
T_5	62.5	T_5	62.5
T	305		
CF=3721			
$\Sigma\Sigma X_{ij}^2$=3792			

ANOVA table						
S.O.V	df	SS	MS	F_{ratio}	F_{table}	Conclusion
Columns	4	6	1.5	0.479	3.259	Accept H_0,
Rows	4	4.8	1.2	0.383	3.259	Accept H_0,
Treatment	4	22.6	5.65	1.805	3.259	Accept H_0,
Error	12	37.6	3.13			
Total	24	71				

Q No. 11.9:

Totals	
T_1	19.8
T_2	17.5
T_3	15.3
T_4	20.4
T	73
CF=222.04	
$\sum\sum x^2$=229.68	

ANOVA table						
S.O.V	Df	SS	MS	F_{ratio}	F_{table}	Conclusion
Between	3	2.72	0.907	3.69	3.098	Reject H_0
Error	20	4.92	0.246			
Total	23	7.64				

Q No. 11.10:

Totals	
T_1	262.5
T_2	282.5
T_3	324.0
T_4	372.5
T	1241.5
CF=77066.11	
$\sum\sum x^2$=76664.83	

ANOVA table						
S.O.V	df	SS	MS	F_{ratio}	F_{table}	Conclusion
Between	3	1422.84	474.28	43.16	3.239	Reject H_0
Error	16	175.88	10.99			
Total	19	1598.72				

$$LSD = t_{(\alpha/2,\nu)}\sqrt{\frac{2MSE}{r}} = t_{(0.025,16)}\sqrt{\frac{2(10.99)}{5}} = 2.12(2.097) = 4.44$$

$\bar{T}_1$ $\bar{T}_2$ $\bar{T}_3$ $\bar{T}_4$

52.5 56.5 64.5 74.5 only T_1 T_2 pair is same

Q No. 11.11:

Totals	
T_1	33
T_2	20
T_3	15
T_4	43
T	111
CF=1026.75	
$\sum\sum x_{ij}^2$=1203	

ANOVA table						
S.O.V	df	SS	MS	F_{ratio}	F_{table}	Conclusion
Between	3	160.92	53.64	27.996	4.066	Reject H_0
Error	8	15.33	1.916			
Total	11	176.25				

Chapter 12

INDEX NUMBERS

Statistics is barometer for economist

Learning Goals

- To explain an index number
- To Compute a simple and composite index number
- To identify different types of index numbers
- To Explain the uses of an index numbers

12.1 Index number

An index number is a descriptive measure that shows the change in a variable(s) relative to its level in a given base period, with respect to any characteristic. More simply Index number is a method that measures the changes in data with respect to time or place. Variable may be value, volume or price of a commodity.
In other words it compares the value in current period and base period.

It is calculated as $IN = \frac{Current\ period\ value}{Base\ period\ value} \times 100$

Indexes are very important, especially in economics. Federal governments published various indices regularly. Types of index number based on variable(s) are price index, quantity index and value index etc.

12.1.1 Price index number:

Price index number is a method that measures the change in data of prices with respect to time or place.

12.1.2 Quantity index number:

Quantity index number is a method that measures the change in data of quantities with respect to time or place.

12.1.3 Value index number:

Value index number is a method that measures the change in value of commodities with respect to time or place.

12.2 Steps in the construction of an index number:

Followings are the important steps in the construction of an index number.

12.2.1 Scope:

The first and most important step in the construction of the index numbers is the purpose of the index numbers. Different index numbers are computed for specific

purposes and no single index number is 'all purpose' index number. A researcher should be clear about the purpose of the index number before its construction. Different purposes results in different types of index numbers e.g. consumer price index number (CPI), whole sale index (WPI) and producer price index (PPI)

12.2.2 Selection of base period:

In the construction of index numbers, selection of the base year is the first step.

12.2.2.1 Base period: The period with which prices in other periods are to be compared is called the base period. The base year should be a normal year (free from abnormal conditions like wars, famines, floods, political instability, etc. Base year can be selected in two ways (a) through fixed base method in which the base year remains fixed; and (b) through chain base method in which the base year goes on changing.

12.2.2.2 Fixed base method:

A method in which base period is fixed and does not change, but average of several year is also used as base period.

12.2.2.3 Chain base method:

A method in which base period changes for every year, and preceding year becomes base year for each year.

12.2.3 Selection of Commodities:

Number of commodities according to the scope of the index numbers to be computed are several, and all these commodities cannot be included, so only representative commodities should be selected. Types of index number based on number of commodities are simple index number and composite index number.

12.2.3.1 Simple index number:

Simple index number is a method that measures the change in a single variable with respect to time or place.

Examples for simple index numbers through fixed base method:

Example 12.1: Calculate index number, from the following currency rate of the USD in PKRs on 1st January every year, taking the year 2010 as base

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Price	85	86	91	98	106	101	106	109	111	139	155

Solution: : Index numbers (price relatives) are calculated by the formula $PR = \frac{P_n}{P_0} \times 100$ where $P_0 = 85$

Year	Price of USD	Index Number
2010	85	100
2011	86	101.18
2012	91	107.06
2013	98	115.29
2014	106	124.71
2015	101	118.82
2016	106	124.71
2017	109	128.24
2018	111	130.59
2019	139	163.53
2020	155	182.35

Example 12.2: Calculate index number, from the following GDP of Pakistan in millions of USD, taking average of first three years.

Year	2011	2012	2013	2014	2015
GDP USD (millions)	213.587	224.384	231.219	244.361	270.556
Year	2016	2017	2018	2019	
GDP USD (millions)	278.655	304.567	314.568	278.222	

Solution: Index numbers (price relatives) are calculated by the formula $PR = \frac{P_n}{P_0} \times 100$, where P_0 = Average of first three years = (213.587+224.384+231.219)/3 = 223.063

Year	Price of USD	Index Number
2011	213.587	95.75
2012	224.384	100.59
2013	231.219	103.66
2014	244.361	109.55
2015	270.556	121.29
2016	278.655	124.92
2017	304.567	136.54
2018	314.568	141.02
2019	278.222	124.73

Examples for simple index numbers through chain base method:

Example 12.3: Calculate chain indices, from the following wheat price in Rs / kg in Pakistan.

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019
Price	30	24	32	33	22	18	19	25	29

Solution: For calculating chain indices, link relatives should be calculated at first by the formula $LR = \frac{P_n}{P_{n-1}} \times 100$. At second step chain indices are computed.

Year	Price	Link Relatives	Chain indices
2011	30	-	-
2012	24	80	80
2013	32	133.33	$\frac{80 \times 133.33}{100} = 106.66$
2014	33	103.13	$\frac{106.66 \times 103.13}{100} = 110$
2015	22	66.67	73.34
2016	18	81.82	60.01
2017	19	105.56	63.34
2018	25	131.58	83.34
2019	29	116	96.67

12.2.3.2 Composite index number:

Composite index number is a method that measures the change in two or more variables with respect to time or place.

12.2.4 Selection of Average:

Selection of a suitable average for index number calculation is an important step. Theoretically, geometric mean is the best for this purpose, but generally arithmetic mean is used.

Composite index number may be computed by two methods (i) simple average of relatives and (ii) simple aggregative method

Simple aggregative index number: In this method the sum of the given year prices of all commodities is divided by the sum of the base year prices of the same commodities and the result is expressed in percentage, i.e. $P_{0n} = \frac{\sum P_n}{\sum P_0} \times 100$

Examples for composite index numbers using various averages.

Example 12.4 (a): Calculate index numbers, from the following data about prices / kg of various commodities, taking year 2014 as base year.

Year	2014	2015	2016	2017	2018	2019
Wheat	33	22	18	19	25	29
Barley	18	13	12	10	13	17
Rice	47	42	39	40	49	57
Maize	21	18	16	17	17	23

www.indexmundi.com

Solution: Index numbers (price relatives) are calculated by the formula $PR = \frac{P_n}{P_0} \times 100$, where P_0 = 33, 18, 47 and 21 for wheat, barley, rice and maize respectively.

Year	Wheat	Barley	Rice	Maize	Price relatives				I. No. (mean)
					Wheat	Barley	Rice	Maize	
2014	33	18	47	21	100	100	100	100	100
2015	22	13	42	18	66.67	72.22	89.36	85.71	78.49
2016	18	12	39	16	54.54	66.67	82.98	76.19	70.10
2017	19	10	40	17	57.58	55.56	85.11	80.95	69.80
2018	25	13	49	17	75.76	72.22	104.26	80.95	83.30
2019	29	17	57	23	87.88	94.44	121.28	109.52	103.28

Example 12.4 (b): Calculate index numbers, from the following data about prices per 100 kg of various commodities, taking year 2001 as base year, using GM as an average.

Year	Chilies	Garlic	Ginger
2001	6600	3225	3250
2002	4312	2275	2762
2003	5725	2200	4000
2004	5215	4300	7400
2005	3205	4500	4750
2006	8450	6133	3650
2007	10250	3404	4900

Solution: Index numbers (price relatives) are calculated by the formula $PR = \frac{P_n}{P_0} \times 100$, where P_0 = 6600, 3225 and 3250 for chilies, garlic and ginger respectively.

Year	Chilies	Garlic	Ginger	Price relatives			I. No. (GM)
				Chilies	Garlic	Ginger	
2001	6600	3225	3250	100	100	100	100
2002	4312	2275	2762	65.33	70.54	84.98	73.16
2003	5725	2200	4000	86.74	68.22	123.08	89.97
2004	5215	4300	7400	79.02	133.33	227.69	133.87
2005	3205	4500	4750	48.56	139.53	146.15	99.67
2006	8450	6133	3650	128.03	190.17	112.31	139.84
2007	10250	3404	4900	155.30	105.55	150.77	135.20

Example 12.5: Calculate chain indices, from the following data about prices / kg of various commodities, taking year 2000 as base year and using median as an average.

Year	Bajra	Gram	Maize
2000	6.95	25.1	6.91
2001	8.62	29.8	7.63
2002	9.25	25.6	8.21
2003	8.37	16.1	7.85
2004	12.50	23.8	8.59
2005	13.62	21.1	9.50

Solution:

Year	Bajra	Gram	Maize	Wheat	Link relatives			
					Bajra	Gram	Maize	Wheat
2000	6.95	25.1	6.91	8.62	-	-	-	-
2001	8.62	29.8	7.63	7.90	124.03	118.73	110.42	91.65
2002	9.25	25.6	8.21	8.62	107.31	85.91	107.60	109.11
2003	8.37	16.1	7.85	10.67	90.49	62.89	95.62	123.78
2004	12.50	23.8	8.59	11.50	149.34	147.83	109.43	107.78
2005	13.62	21.1	9.50	11.10	108.96	88.66	110.59	96.52

Year	Link relatives				Median	C. I
	Bajra	Gram	Maize	Wheat		
2000	-	-	-	-	-	-
2001	124.03	118.73	110.42	91.65	114.58	114.58
2002	107.31	85.91	107.60	109.11	107.46	123.13
2003	90.49	62.89	95.62	123.78	93.06	114.58
2004	149.34	147.83	109.43	107.78	128.63	147.38
2005	108.96	88.66	110.59	96.52	102.74	151.42

Example 12.6: Calculate index numbers, from the following data about prices /100 kg of various commodities, taking year 2001 as base year, by the method of simple aggregative.

Year	Chilies	Garlic	Ginger
2001	6600	3225	3250
2002	4312	2275	2762
2003	5725	2200	4000
2004	5215	4300	7400
2005	3205	4500	4750
2006	8450	6133	3650
2007	10250	3404	4900

Solution:

Year	Chilies	Garlic	Ginger	Total	I. No.
2001	6600	3225	3250	13075	100
2002	4312	2275	2762	9349	71.50
2003	5725	2200	4000	11925	91.20
2004	5215	4300	7400	16915	129.37
2005	3205	4500	4750	12455	95.26
2006	8450	6133	3650	18233	139.45
2007	10250	3404	4900	18554	141.90

12.2.5 Selection of Weights: If all the commodities selected for the construction of index numbers are not of equal importance and different numerical values are assigned to the commodities according to their importance. These assigned numerical values are called weights. Assigning these weights to the commodities is called selection of weights. An index number calculated by attaching weights to the commodities involved

is called weighted index number. Where-as an index number calculated without attaching weights to the commodities is called unweighted index number.

12.2.5.1 Un-weighted index number:
Index number that measures changes in the data of a group of commodities when the relative importance of the commodities has not been taken into account are called un-weighted index number. Numerical example for this method is given in previous section.

12.2.5.2 Weighted index number:
An index number is called weighted index number if we use the relative importance(weights) of commodities while constructing the index number. Weighted aggregative index and weighted average of relative index are two methods for its calculation.

12.2.5.2.1 Weighted aggregative index number:
It is the percentage ratio of aggregate items for a given period to the aggregate of weighted items in the base period. Different formulae for this method are given below.

Laspeyere's price index number:
This index uses base year quantities as weights, for this reason it is called base year weighted index. It is defined as $P_{0n} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$

Paasche's price index number:
This index uses current year quantities as weights, for this reason it is called current year weighted index. It is defined as $P_{0n} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100$

Marshal Edgeworth price index number:
In this index sum of the base year and current year quantities is used as weights. It is defined as $P_{0n} = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100$

Fisher's ideal price index number:
In this index square root of the product of the base year and current year weighted index numbers. It is called ideal because it satisfies the time reversal and factor reversal test. it is defined as $P_{0n} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}} \times 100.$

Walsh's price index number:
In this index square root of the product of the base year and current year quantities is used as weights. It is defined as $P_{0n} = \frac{\sum p_n \sqrt{q_0 q_n}}{\sum p_0 \sqrt{q_0 q_n}} \times 100.$

Example 12.7: Calculate weighted price index numbers, from the following data for 2010 taking 2005 as base period using

(i) Laspeyere's method.
(ii) Paasche's method
(iii) Fisher's method and
(iv) Marshal Edgeworth method

Commodities	2005		2010	
	Price	Quantity	Price	Quantity
A	12	500	30	600
B	15	200	25	250
C	20	50	24	55
D	100	300	110	400

Solution: Calculate weighted price index numbers, from the following data for 2010 taking 2005 as base period using

Com.	2005		2010		p_0q_0	p_1q_0	p_0q_1	p_1q_1
	p_0	q_0	p_1	q_1				
A	12	500	30	600	6000	15000	7200	18000
B	15	200	25	250	3000	5000	3750	6250
C	20	50	24	55	1000	1200	1100	1320
D	100	300	110	400	30000	33000	40000	44000
Total					**40000**	**54200**	**52050**	**69570**

(i) Laspeyere's price index number $P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{54200}{40000} \times 100 = 135.5$

(ii) Paasche's method $P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{69570}{52050} \times 100 = 133.66$

(iii) Fisher's method

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{\frac{54200}{40000} \times \frac{69570}{52050}} \times 100 = 134.58$$

(iv) Marshal Edgeworth method

$$P_{01} = \frac{\sum p_1 q_2 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100 = \frac{54200 + 69570}{40000 + 52050} \times 100 = 134.46$$

12.2.5.2.2 Weighted average of relatives index number:

In this method each relative is multiplied by its corresponding weight and then sum of the products of relatives and weights is divided by total of the weights. It is given as

$$P_{0n} = \frac{\sum IW}{\sum W}$$

Special cases:

Laspeyere's price index number:

It is defined as $P_{0n} = \frac{\sum\left(\frac{p_n}{p_0} \times 100\right)W}{\sum W}$; where $W = p_0 q_0$

Paasche's price index number:

It is defined as $P_{0n} = \frac{\sum\left(\frac{p_n}{p_0} \times 100\right)W}{\sum W}$; where $W = p_0 q_n$

Palgrave's price index number:

It is defined as $P_{0n} = \frac{\sum\left(\frac{p_n}{p_0} \times 100\right)W}{\sum W}$; where $W = p_n q_n$

Example 12.8: Calculate weighted index numbers, from the following data about prices per kg of various commodities, taking year 2014 as base year. Weights for wheat, barley, rice and are 10, 3, 7 and 5 respectively.

Year	2014	2015	2016	2017	2018	2019
Wheat	33	22	18	19	25	29
Barley	18	13	12	10	13	17
Rice	47	42	39	40	49	57
Maize	21	18	16	17	17	23

$$I.No = \frac{\sum IW}{\sum W}$$

Solution:

Year	Wheat	Barley	Rice	Maize	Price relatives				I. No.
					Wheat	Barley	Rice	Maize	
					10	3	7	5	
2014	33	18	47	21	100	100	100	100	100
2015	22	13	42	18	66.67	72.22	89.36	85.71	77.49
2016	18	12	39	16	54.54	66.67	82.98	76.19	68.29
2017	19	10	40	17	57.58	55.56	85.11	80.95	69.71
2018	25	13	49	17	75.76	72.22	104.26	80.95	84.35
2019	29	17	57	23	87.88	94.44	121.28	109.52	102.35

12.3 Consumer price index number(CPI):

Consumer price index number abbreviated as CPI is defined as a measure that study the weighted average of prices of a specific basket of goods and services including food, medical care and transportation etc. It is also known called household-budget, retail price or cost of living index number. Official agencies such as ministry of labour and federal bureau of statistics etc. are main responsible for its compilation.

Main reason for compiling a CPI is

(i) To compensate wage-earners for inflation by adjusting their wage rates in proportion to the percentage change in the CPI. Therefore specifically it is known as a compensation index.

(ii) To index pensions and social security benefits.

(iii) To index other payments, such as interest payments or rents, or the prices of bonds.

(iv) To use as a proxy for the general rate of inflation.

(v) To deflate household consumption expenditures or other items in national accounts.

12.3.1 Methods for computation of consumer price index number(CPI):

Consumer price index number (CPI) can be computed by (i) aggregate expenditure method and (ii) household budget method

12.3.1.1 Aggregate expenditure method:

Aggregate expenditure method is defined as $P_{0n} = \frac{\sum p_n q_n}{\sum p_0 q_0} \times 100$

12.3.1.2 Household budget method:

Household budget method also known as family budget method is defined as

$P_{0n} = \frac{\sum IW}{\sum W}$; where W are weights pre-defined or $W = p_0 q_0$ and $I = \frac{p_n}{p_0} \times 100$

Example 12.9 AJ & K university, BA/BSc A/2012

The following table gives average annual prices of 8 commodities for the year 1990 and 1995. Compute price index number for 1996 on the basis of 1990, using

(i) Aggregate expenditure method (ii) Household budget method

Commodity	Quantity consumed	Unit of price	Price	
			1990	1994
Wheat	20 Kg	Rs. Per Kg	120	150
Rice	8 Kg	Rs. Per Kg	1300	1400
Sugar	1 Kg	Rs. Per Kg	800	1000
Ghee	1 Kg	Rs. Per Kg	1000	1200
Milk	25 kg	Rs per Kg	10	15
Vegetable	16 Kg	Rs. Per Kg	5	8
Mutton	5 Kg	Rs. Per Kg	150	180
Fuel	4 maund	Rs. Per maund	200	300

1 maund = 40 Kg

Solution

The following table gives average annual prices of 8 commodities for the year 1990 and 1996. Compute price index number for 1996 on the basis of 1990, using

(i) Aggregate expenditure method (ii) Household budget method

Commodity	q_0	Price		p_0q_0 W	p_1q_0	I	IW
		p_0	p_1				
Wheat	20	120	150	2400	3000	125	300000
Rice	8	1300	1400	10400	11200	107.69	1119976
Sugar	1	800	1000	800	1000	125	100000
Ghee	1	1000	1200	1000	1200	120	120000
Milk	25	10	15	250	375	150	37500
Vegetable	16	5	8	80	128	160	12800
Mutton	5	150	180	750	900	120	90000
Fuel	4	200	300	800	1200	150	120000
Total				16480	19003		1900276

(i) Aggregate expenditure method:

$$P_{01} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100 = \frac{19003}{16480} \times 100 = 115.31$$

(ii) Household budget method:

$$P_{0n} = \frac{\sum IW}{\sum W} = \frac{1900276}{16480} = 115.31$$

where W are weights pre-defined or $W = p_0q_0$

10.4 Value index number:

Value index number is a method that measures the change in value of commodities with respect to time or place. It is given as

$$V_{0n} = \frac{\sum p_nq_n}{\sum p_0q_0} \times 100$$

Example 12.10 AJ & K. university BA/BSc A/2013

The following table gives prices and quantities of four commodities produced by a farmer taking 2005 as base year. Compute simple aggregative value index for year 2008.

Commodities	Quantity (kg)		Price (rupees)	
	2005	2008	2005	2008
A	130	150	50.50	55.00
B	110	98	75.25	80.25
C	85	105	100.50	110.45
D	72	80	58.44	60.50

Solution:

Comd	q_0	q_1	p_0	p_1	p_1q_1	p_0q_0
A	130	150	50.50	55.00	8250	6565
B	110	98	75.25	80.25	7864.5	8277.5
C	85	105	100.50	110.45	11597.25	8542.5
D	72	80	58.44	60.50	4840	4207.68
Total					**32551.75**	**27592.68**

$$V_{0n} = \frac{\sum p_nq_n}{\sum p_0q_0} \times 100 = \frac{32551.75}{27592.68} \times 100 = 117.97$$

12.5 Uses of index number:

Some uses of index number are given below.

(i) It is used to measure purchasing power of the money.
(ii) It is used in business and industry.
(iii) It is used to compare variations for different periods and places.
(iv) It is used to measure changes in the trade of a country.
(v) It helps in forecasting the future economic trends.
(vi) It is used to measure the cyclical and seasonal variations in time series.
(vii) It is used in education for IQ comparison and effectiveness of teaching system.
(viii) It works like a barometer showing fluctuation in daily life cost of living and employment etc.

12.6 Limitations of index number:

Some limitations of index number are given below.

(i) Index numbers are usually based on a sample, therefore sampling errors are introduced.
(ii) Different methods of construction yield different results.
(iii) Comparisons over long periods are not reliable.
(iv) All index numbers are not suitable for all purposes.
(v) It is not possible to take into account all changes in production.

Multiple Choice Questions

1. Index number shows
a) average change b) relative change c) variation d) skewness

2. An index number with single variable
a) simple index b) weighted index c) unweighted index d) composite index

3. Base period in chain base method
a) fixed b) first c) constant d) changed

4. An index number with more than one variable
a) simple index b) weighted index c) unweighted index d) composite index

5. Base period in fixed base method
a) last b) first c) constant d) changed

6. The most suitable average for index number
a) AM b) GM c) HM d) median

7. Index number used for measuring change in price of commodities
a) price index b) quantity index c) volume index d) value index

8. Index number used for measuring change in quantity of commodities
a) price index b) price relative c) volume index d) value index

9. Index number used for measuring change in value of commodities
a) price index b) quantity index c) volume index d) value index

10. Weights used for measuring index numbers are equal
a) weighted index b) quantity index c) un-weighted index d) CPI

11. Weights used for measuring index numbers are un-equal
a) weighted index b) quantity index c) un-weighted index d) CPI

12. Index number for base period a) 0 b) 100 c) 1000 d) mean

13.Fisher's index number is ---- between Laspeyer's and Paasche's index number a) AM b) GM c) HM d) Median

14. Link relative
a) $\frac{P_0}{P_n} \times 100$ b) $\frac{P_n}{P_{n-1}} \times 100$ c) $\frac{P_0}{P_n} \times 100$ d) $\frac{P_{n-1}}{P_n} \times 100$

15. Index number plays important role in economics and
a) mathematics b) accounting c) business d) sociology

16. Index numbers are expressed in
a) percentage b) units c) square of unit d) thousands

17. Index number that serve many purposes
a) CPI b) PPI c) general purpose index d) Cost of living index

18. Another name of CPI
a) cost of living index b) retail price index c) both a and b d) None

19. Purchasing power of money may be computed using
a) Price index b) quantity index c) value index d) CPI

20. Laspeyer's index = 110 and Paasche's index = 114 then Fisher's Index
a) 110 b) 111 c) 111.98 d) 112

21.Weights used in Paasche's formula belongs to ---- period
(a) base (b) first (c) current (d) given

22.Weights used in Lespyer's formula belongs to ---- period:
(a) base (b) first (c) current (d) given

key

Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans
1	b	2	a	3	d	4	d	5	c	6	b
7	a	8	c	9	d	10	c	11	a	12	b
13	b	14	b	15	c	16	a	17	c	18	c
19	d	20	c	21	c	22	a				

Exercise

Q No. 12.1: Define index number, simple and composite index number.

Q No. 12.2: Differentiate between weighted and unweighted index number.

Q No. 12.3: Find index numbers from the following data of the production (000 tons) of apples in Pakistan, using 2011 as base period.

Year	2011	2012	2013	2014	2015	2016	2017	2018
Production	525.9	598.7	556.4	606.1	617.2	620.5	669.9	649.3

Q No. 12.4: Find index numbers from the following data of the production (000 tons) of bananas in Pakistan, using average of first three year as base.

Year	2011	2012	2013	2014	2015	2016	2017	2018
Production	141.2	98.2	140.6	120.4	118.8	132.2	134.9	135.1

Q No. 12.5 Find chain indices from the following data of the production (000 tons) of citrus in Pakistan.

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018
Production	2150	1982	2147	2002	2168	2396	2344	2180	2351

Q No. 12.6 Find chain indices from the following data of gasoline price.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Price	1.48	1.42	1.34	1.56	1.85	2.27	2.57	2.80	3.25

Q No. 12.7 Compute quantity index from the following data of automobiles production in Pakistan from 2011 to 2019.

Year	2011	2012	2013	2014	2015	2016
Production	153114	175184	134849	148746	229686	274536
Year	2017	2018	2019			
Production	285248	342575	298083			

Old paper's questions

1. Punjab university, Lahore BA/BSc A/2008

Given

Commodity	Quantity (units)		Value	
	2001	2006	2001	2006
A	100	150	600	1200
B	80	100	400	700
C	60	72	180	432
D	30	33	450	363

Compute the following:

(i) Fisher's quantity index number for 2006

(ii) Simple aggregative value index for 2006

2. Punjab university, Lahore BA/BSc A/2011

Calculate consumer price index number for 1994 on the basis of 1990, using

(i) Aggregate expenditure method (ii) Household budget method

Commodity	Quantity Consumed	Unit of price	Price	
			1990	1994
Wheat	20 Kg	Rs. Per 40 Kg	200	240
Rice	8 Kg	Rs. Per 40 Kg	800	880
Sugar	4 Kg	Rs. Per 40 Kg	400	480
Ghee	1 Kg	Rs. Per Kg	30	40
Milk	25 liters	Rs per liter	10	12
Vegetable	16 Kg	Rs. Per 40 Kg	200	240
Mutton	5 Kg	Rs. Per Kg	100	120
Fuel	100 Kg	Rs. Per 40 Kg	800	920

3. Punjab university, Lahore BA/BSc A/2012

The prices and quantities of three commodities during 1955 and 1965 are given below

Commodity	1955		1965	
	Price	Quantity	Price	Quantity
A	10	501	12	600
B	40	100	38	194
C	50	76	40	56

Using 1955 as base period and base period quantity as weights, compute the weighted aggregative price index and weighted average of relative price index for 1965.

4. Punjab university, Lahore BA/BSc A/2013

Calculate Fisher's index number for the year 2012 taking 2007 as base period.

Commodity	Price		Quantity	
	2007	2012	2007	2012
A	64	80	270	290
B	40	45	124	144
C	18	24	130	137
D	58	68	185	355

5. Punjab university, Lahore BA/BSc A/2014

Calculate Laspeyre's price index number taking 2010 as base period. Discuss your results.

Commodity	Units consumed		Price per unit	
	2010	2013	2010	2013
A	20	16	12	20
B	35	38	21	24
C	10	9	30	41
D	45	50	8	12

6. Punjab university, Lahore BA/BSc A/2015 and A/2016

Construct chain indices for the following years, taking 1940 as base and using simple average of relatives

Year	Price in Rs. Per Maund (37.2 Kg)		
	Wheat	Rice	Maize
1940	2.80	10.50	2.70
1941	3.40	10.80	3.20
1942	3.60	10.60	3.50
1943	4.00	11.00	3.80
1944	4.20	11.50	4.00

7. Punjab university, Lahore BA/BSc S/2015

From the data given below Compute the index number of prices, taking 2009 as base, using median as an average of price relatives.

Year	Commodities (Prices in Rs.)			
	A	B	C	D
2009	16.25	20.00	2.40	10.50
2010	17.22	22.40	2.64	12.50
2011	19.55	16.00	3.00	12.65
2012	18.70	20.00	3.80	14.65

8. Bahauddin Zakariya University., Multan BA/BSc A/2015 and Punjab university, Lahore MA Economics A/2016

Given the following, Construct Fisher's index number for

(i) 1964, taking 1957 as base year

(ii) 1957, taking 1964 as base year

Commodity	1957		1964	
	Price	Quantity	Price	Quantity
Rice	9.3	100	4.5	90
Wheat	6.4	11	3.7	19
Jawar	5.1	5	2.7	3

9. Punjab university, Lahore BA/BSc A/2016

The following table gives the average annual prices of five commodities for the year 2001 – 2005.

Year	Price			
	Wheat	Rice	Cotton	Sugar
2001	220	19	1200	20
2002	260	21	900	18
2003	300	30	1500	35
2004	350	32	1000	25
2005	400	38	1100	25

Taking 2001 = 100, calculate simple average of relatives using GM as an average.

10. Punjab university, Lahore BA/BSc S/2017

Find the chain indices from the following price relatives of four commodities, using geometric mean of the relatives for each year.

Year	Commodities			
	A	B	C	D
1991	81	77	119	55
1992	62	54	128	82
1993	104	87	111	100
1994	93	75	154	96
1995	90	43	165	88

11. Punjab university, Lahore BA/BSc A/2018

The prices and quantities of three commodities during 1997 and 2007 are given below

Commodity	Price		Quantity	
	1997	2007	1997	2007
A	12	10	501	600
B	38	50	100	194
C	40	40	56	76

Compute weighted aggregative price index for 1997 with 2007 = 100 by Paasche's method.

12. Punjab university, Lahore BA/BSc A/2017

From the data given below, compute index number of prices, taking 1980 as base. Use simple average of relatives.

Year	Commodities (Prices in Rs.)			
	A	B	C	D
1980	16.25	20.00	2.40	10.50
1981	17.22	22.40	2.64	12.50
1982	19.55	16.00	3.00	12.66
1983	18.70	20.00	3.80	14.65

13. B.I.S.E., Sahiwal intermediate A/2017

Find chain indices from the following data:

Year	Price A	Price B	Price C
1990	255	216	330
1991	186	162	384
1992	312	261	333
1993	279	225	462

14. B.I.S.E., Sahiwal intermediate A/2018

Construct index numbers from the following data by applying
(i) Laspeyer's method (ii) Paasche's method

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	8	55	2	50
B	4	105	4	115
C	6	65	8	55
D	12	35	14	19

15. Punjab university, Lahore MA Economics A/2013 and S/2014

The following table contains information from the raw material purchase records of small factory for the year 2002 and 2003.

Commodity	2002		2003	
	Price	Quantity	Price	Quantity
A	5	50	6	72
B	7	84	10	80
C	10	80	12	96
D	4	20	5	30

Calculate Fisher's ideal index number taking 2002 as base.

16. Bahauddin Zakariyya University, Multan B.Com I-A/2015

These data indicates the values (in dollars) of the principal product explored by a developing country. Determine un-weighted aggregative price index number for 1993 and 1995 based on 1991.

Commodity	1991	1993	1995
Coffee	834	1436	1321
Sugar	96	188	122
Copper	241	258	269
Zinc	142	125	106

17. Bahauddin Zakariyya University, Multan B.Com II-A/2017

Calculate Laspeyer's, Paasche's, Fisher's and Marshal Edgeworth index number for 1981 taking 1980 as base year from the following data.

Commodity	1980		1981	
	Price	Quantity	Price	Quantity
Wheat	30	110	32	112
RICE	40	100	38	110
Jawar	25	50	22	80
Maize	10	40	5	50
Sugar	20	80	18	88

18. AJ & K. University, BSc-A/2014

The following table gives the average prices of five commodities during the year of 2005 & 2008.

Commodity	Quantity (2005)	Unit of price	Prices	
			2005	2008
Wheat	20 Kg	Rs. Per 40 Kg	300	350
Dal	8 Kg	Rs. Per 40 Kg	350	410
Edible oil	1.5 Kg	Rs. Per 40 Kg	800	900
Fuel	160 Kg	Rs. Per 40 Kg	200	225
Clothing	22 yards	Rs per yard	450	550

Calculate consumer price index number for 2008 on the basis of 2005, using
(i) Aggregate expenditure method (ii) Household budget method

19. AJ & K. University, BSc-A/2015

Compute the index numbers of Marshal Edgeworth and Fisher's type.

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	7	70	5	49
B	5	27	7	28
C	10	35	9	29
D	9	50	4	42
E	3	16	10	25

20. AJ & K. University, BSc-A/2016

(a) The following table gives the index number of three commodities in 2008. Calculate:
(i) simple average & (ii) the weighted average of these index numbers, when food, fuel and light and clothing are given weights 5, 1 and 3 respectively.

Commodities	Food	Fuel and light	Clothing
Index number	111	105	106

(b) We are given Fisher's index number = 104.95 and Laspeyer's index number = 105.25. Find Paasche's index number.

21. AJ & K. University, BSc-A/2017

Compute cost of living index numbers from the following data.

Commodity	Weights	Base year price in Rs	Current year price in Rs
A	4	2	7
B	1	5	2
C	5	8	5
D	2	5	12
E	3	3	6

Solution

Q No. 12.3:

Year	2011	2012	2013	2014	2015	2016	2017	2018
I. No	100	113.84	105.8	115.25	117.36	117.99	127.38	123.46

Q No. 12.4:

Average of first three years is 126.67

Year	2011	2012	2013	2014	2015	2016	2017	2018
I. No	111.47	77.52	111.0	95.05	93.79	104.37	106.50	106.66

Q No. 12.5

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018
LR	-	92.19	108.32	93.25	108.29	110.52	97.83	93.00	107.84
I. No	-	92.19	99.86	93.11	100.83	111.44	109.02	101.39	109.34

Q No. 12.6

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
LR	-	95.95	94.37	116.42	118.60	122.70	113.22	108.95	116.07
I. No	-	96.95	91.49	106.51	126.32	154.99	175.48	191.18	221.90

Q No. 12.7

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019
I. No	100	114.39	88.07	97.15	150.01	179.30	186.30	223.74	194.68

TIME SERIES

Basic Terms used in statistics are well known to a common person

Learning Goals

(i) To provide introduction to time series.
(ii) To analyze and forecast the time series.

13.1 Time Series

An arrangement of data in accordance with the time of its occurrence is called time series. For example monthly registered patients of corona, annual sale of books or enrollment of students in a college annually. OR

A sequence of observations measured over successive periods of time, is called a time series. It is also called historical data. Time periods may be a day, a week, a month, a year or any other regular interval.

13.1.1 Historigram

The graph of a time series is called historigram. The variable time is taken along X-axis and observed value along Y-axis.

A time series plot, in which time is taken along X-axis and observed value along Y-axis, is called historigram.

13.1.2 Purpose of time series

The basic purpose of time series is to use it for forecasting.

13.2 Components of a time series

A time series has four basic variations which are the components of time series, , i.e. Secular trend (T), seasonal variations (S), cyclical fluctuations (C) and irregular movements (I).

13.2.1 Secular trend

It is a long term movement(change) that indicates the general direction of the time series. It represents smooth, steady and gradual movement in a time series. Its time period is not less than 10 years.

13.2.2 Seasonal variations

Seasonal variations are short term movement(change) that indicates the identical changes in a time series during the corresponding season. Main causes of these variations are seasons, religious festivals and social costumes.

13.2.3 Cyclical fluctuations

Cyclical fluctuations are long term oscillations or swings about the trend line or curve since the movements take the form of upward and downward swings, they are also called "cycles". The four bases of a business cycle are

(i) Period of prosperity(boom or peak)
(ii) Period of recession(contraction or decline)
(iii) Period of depression(trough or slump)
(iv) Period of recovery(expansion or revival)

13.2.4 Irregular movement

Randoms variations are caused by some unusual events such as earthquake, floods, droughts, wars, strikes or political events etc. these variations are also known as accidental, residual or erratic variations

13.3 Signal

Systematic component of variation in time series is called signal. Its value can be completely determined. First three components (Secular trend, Seasonal variation and cyclical fluctuation) are signals.

13.4 Noise

An irregular or random component of variation in time series is called noise. Its value cannot be completely determined.

13.5 Analysis of time series

The analysis of time series is the decomposition of a time series into its different components for their separate study.

To discover the pattern in time series, that helps in forecasting the future value is called analysis of the time series.

13.6 Assumptions/types of models used in analysis of time series

It is assumed that the components of time series follow two types of relationship or models (i) multiplicative model (ii) Additive model

13.6.1 Multiplicative model: In multiplicative model we assume that each observed value Y of a time series is the product of the effects of four components, i.e. Secular trend (T), seasonal variations (S), cyclical fluctuations (C) and irregular movements (I). systematically it is denoted as $Y = T \times S \times C \times I$.

13.6.2 Additive model: In additive model we assume that each observed value Y of a time series is the sum of the effects of four components, i.e. Secular trend (T), seasonal variations (S), cyclical fluctuations (C) and irregular movements (I). systematically it is denoted as $Y = T + S + C + I$.

13.7 Importance of analysis of time series

(i) Analysis of time series plays an important role to understand the business policy.
(ii) It helps in studying the past behaviour of the data.
(iii) It helps in forecasting.
(iv) It helps in making comparisons.

13.8 Measurement of Secular trend

Following methods are used for measuring the secular trend.

(i) Free hand curve method.
(ii) The method of semi averages
(iii) The method of moving averages
(iv) The method of least squares

13.8.1 Free hand curve method

In this method the data are plotted on a graph measuring the time units along the X-axis and values of the time series along Y-axis. Draw a free hand curve through these plotted points in such way that it shows the general trend of the time series.

13.8.2 The method of semi averages

In this method the data is divided into two equal parts. Average for each part is computed and placed against the center of each part.

Example 13.1 From the following data analyse the time series by the method of semi averages.

Time period	1	2	3	4	5	6	7	8	9	10	11	12
Value	17	21	19	23	18	16	20	18	22	20	15	22

Solution:

Time period (X)	Y	Semi averages Y	X	$\hat{Y} = 18.71 + 0.08X$	$e = Y - \hat{Y}$
1	17			18.79	-1.79
2	21			18.87	2.13
3	19	$\bar{Y}_1 = 114 \div 6 = 19$	$\bar{X}_1 = 3.5$	18.95	0.05
4	23			19.03	3.97
5	18			19.11	-1.11
6	16			19.19	-3.19
7	20			19.27	0.73
8	18			19.35	-1.35
9	22			19.43	2.57
10	20	$\bar{Y}_2 = 117 - 6 = 19.5$	$\bar{X}_2 = 9.5$	19.51	0.49
11	15			19.59	-4.59
12	22			19.67	2.33
Total	**231**			**231**	

Semi average trend line is:

$$\frac{Y-\bar{Y}_1}{\bar{Y}_2-\bar{Y}_1} = \frac{X-\bar{X}_1}{\bar{X}_2-\bar{X}_1} \Rightarrow \frac{Y-19}{19.5-19} = \frac{X-3.5}{9.5-3.5} \Rightarrow \frac{Y-19}{0.5} = \frac{X-3.5}{6} \Rightarrow 12(Y-19) = X-3.5$$

$$12Y - 228 = X - 3.5 \Rightarrow 12Y = 224.5 + X \Rightarrow \hat{Y} = 18.71 + 0.08X$$

13.8.3 The method of moving averages

In this method k-period simple moving averages are calculated by averaging first k-observations and repeating this process by discarding the first and including the next observation that has not been previously included. This process of averaging is continued till the inclusion of the last observation of the time series.

Example 13.2 From the following data analyse the time series by the method of 4-quarter moving averages.

Year	1				2				3			
Quarter	I	II	III	IV	I	II	III	IV	I	II	III	IV
Value	12.5	15.3	10.6	8.8	11.8	16.1	13.3	10.2	13.8	14.4	11.3	8.0

Solution:

Year	Quarter	Y_i	4 quarter moving Total	4 quarter moving Average	4 quarter centred moving Total	4 quarter centred moving Average
1	I	12.5				
	II	15.3	47.2	11.80	23.43	11.72
	III	10.6	46.5	11.63	23.46	11.73
	IV	8.8	47.3	11.83	24.33	12.17
2	I	11.8	50	12.50	25.35	12.68
	II	16.1	51.4	12.85	26.2	13.10
	III	13.3	53.4	13.35	26.28	13.14
	IV	10.2	51.7	12.93	25.36	12.68
3	I	13.8	49.7	12.43	24.31	12.16
	II	14.4	47.5	11.88		
	III	11.3				
	IV	8.0				

Example 13.3 From the following data analyse the time series by the method of 3-year moving averages.

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Value	55	85	70	55	40	70	95	90	50	30	60

Solution:

Year	Y_i	3-year moving	
		Total	Average
2001	55		
2002	85	210	70
2003	70	210	70
2004	55	165	55
2005	40	165	55
2006	70	205	68.33
2007	95	255	85
2008	90	235	78.33
2009	50	170	56.67
2010	30	140	46.67
2011	60		

Example 13.4 Following data is about annual sale in millions of a shop. Analyze the data by the method of 5-year moving averages.

Year	Sale	Year	Sale	Year	Sale	Year	Sale
2005	116	2009	139	2013	186	2017	205
2006	129	2010	175	2014	214	2018	209
2007	155	2011	215	2015	186	2019	199
2008	119	2012	197	2016	198	2020	217

Solution:

Year	Y_i	5-year moving	
		Total	Average
2005	116		
2006	129		
2007	155	658	131.6
2008	119	717	143.4
2009	139	803	160.6
2010	175	845	169
2011	215	912	182.4
2012	197	987	197.4
2013	186	998	199.6
2014	214	981	196.2
2015	186	989	197.8

2016	198	1012	202.4
2017	205	997	199.4
2018	209	1028	205.6
2019	199		
2020	217		

13.8.4 The method of least squares

This method minimizes the sum of squares of residuals for the analysis of a time series by using the time as independent variable (X) and observed values as dependent variable (Y). The best curve fitting among all the curves which can be drawn to the data is that which has the property of the method of least squares.

Here is the process of coding of the time variable

Coding of the variable:

Time	Taking origin at the beginning	Taking origin at the middle				
		Odd numbers of time		Even numbers of time		
					X in half year unit	X in one year unit
	X	Time	X	Time	X	X
T1	0	T1	-	T1	-	-
T2	1	T2	-	T2	-	-
T3	2	T3	-3	T3	-7	-3.5
T4	3	T4	-2	T4	-5	-2.5
T5	4	T5	-1	T5	-3	-1.5
T6	5	T6	0	T6	-1	-0.5
T7	6	T7	1	T7	1	0.5
T8	7	T8	2	T8	3	1.5
T9	8	T9	3	T9	5	2.5
T10	9	T10	-	T10	7	3.5
-	-	-	-	-	-	-
-	-	-	-	-	-	-

Example 13.5 Because of establishment of new university, enrollment in degree class of public college in the area is decreasing. Following data shows the enrollment in college from 2015 to 2021. Develop the linear trend for this time series, also find trend values too. Forecast the enrollment for the year 2022.

Year	Enroll	Year	Enroll	Year	Enroll	Year	Enroll
2015	91	2017	155	2019	55	2021	75
2016	214	2018	185	2020	122		

Solution:

Year	Y_i	X_i	X_iY_i	X_i^2	$\hat{Y}_i$
2015	91	0	0	0	163.71
2016	214	1	214	1	151.85
2017	155	2	310	4	139.99
2018	185	3	555	9	128.13
2019	55	4	220	16	116.27
2020	122	5	610	25	104.41
2021	75	6	450	36	92.55
Total	897	21	2359	91	

$$\bar{X} = \frac{\sum X}{n} = \frac{21}{7} = 3 \quad \text{and} \quad \bar{Y} = \frac{\sum Y}{n} = \frac{897}{7} = 128.14$$

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{7(2359) - (21)(897)}{7(91) - (21)^2} = \frac{-2324}{196} = -11.86$$

$$a = \bar{Y} - b\bar{X} = 128.14 - (-11.86)(3) = 163.71$$

$$\hat{Y}_i = 163.71 - 11.86X_i$$

For estimating the enrollment for 2022, put x = 7:

$$\hat{Y}\,|_{x=7} = 163.71 - 11.86(7) = 80.69$$

Multiple Choice Questions

1. Data arranged according to the time of occurrence is called:
(a) grouped data (b) primary data (c) time series (d) secondary data

2. A time series consists of --- components: (a) 2 (b) 3 (c) 4 (d) 5

3. Graph of a time series is called:
(a) Power curve (b) Ogive (c) Histogram (d) Histogram

4. Secular trend is analysed by the method of semi average when
(a) trend is linear (b) trend is non-linear
(c) irregular movement exists (d) Number of the years is even

5. Increase in the sale of books during the month of April:
(a) Secular trend (b) Seasonal variation
(c) cyclical fluctuation (d) irregular movement

6. Seasonal variation in a time series occurs within a period of:
(a) 1 year (b) 5 years (c) 10 years (d) a quarter

7. Decline in crop production due to flood:
(a) secular trend (b) seasonal variation (c) irregular movement (d) None

8. Mathematical model used mostly for time series is:
(a) Additive (b) Multiplicative (c) Linear (d) Mixed

9. A time series consists of ------- model(s):
(a) 1 (b) 2 (c) 3 (d) 4

10. In linear equation $Y_i = a + bX_i$, a is:
(a) Intercept (b) Slope (c) error term (d) Regression coefficient

11. In linear equation $Y_i = a + bX_i$, b is:
(a) X-Intercept (b) Slope (c) error term (d) Y-intercept

12. In semi averages method, data is divided into --- groups:
(a) 1 (b) 2 (c) 3 (d) 4

13. Decomposition of time series is called:
(a) de-seasonalisation (b) de-trending
(c) analysis of time series (d) historigram

14. Increase in admission of students in colleges:
(a) Secular trend (b) Seasonal variation
(c) cyclical fluctuation (d) Irregular movement

15. Multiplicative model for time series is:
(a) $Y = TSCI$ (b) $Y = T + S + C + I$ (c) $Y = T + S \times C + I$ (d) None

16. Additive model for time series is:
(a) $Y = TSCI$ (b) $Y = T + S + C + I$ (c) $Y = T + S \times C + I$ (d) None

17. A business cycle has --- stages:
(a) 1 (b) 2 (c) 3 (d) 4

18. Free hand curve method based on:
(a) moving averages (b) Parabola (c) Graph (d) Irregular movements

19. Excess of marriage ceremonies before and after the month of Ramadan:
(a) Secular trend (b) Seasonal variation
(c) Cyclical fluctuation (d) irregular movements

20. Movements in secular trend are:
(a) Sudden (b) short term (c) Smooth (d) Abruptly

21. In time series, irregular movements are caused by:
(a) Floods (b) Lockdown (c) Epidemics (d) All

22. Components of a time series:
(a) 1 (b) 2 (c) 3 (d) 4

23. The best fitted trend is that for which sum of square of residuals is:
(a) 0 (b) least (c) maximum (d) maxima

24. Decomposition of a time series is called:
(a) Analysis (b) Historigram (c) Signal (d) Noise

Key

Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans	Sr.	Ans
1	c	2	C	3	d	4	a	5	b	6	a	7	c
8	b	9	B	10	a	11	b	12	b	13	c	14	a
15	a	16	B	17	d	18	c	19	b	20	c	21	d
22	d	23	B	24	a								

Exercise

Q No. 13.1: Define time series, secular trend, seasonal variation, cyclical fluctuation and irregular movement.

Q No. 13.2: Describe the method of semi average, method of moving average for the analysis of a time series.

Q No. 13.3: Annual Revenue of a certain company for last ten years is given below.

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Revenue	8.5	10.8	13.0	14.1	16.3	17.2	18.4	18.5	18.4	18.7

Use method of semi average to analyze the time series.

Q No. 13.4: Consider the following time series, analyze it using 4-quarter central moving average method.

Quarter	Year			
	2017	2018	2019	2020
1	71	68	62	60
2	49	41	49	50
3	58	60	53	55
4	78	81	72	74

Q No. 13.5: Consider the following time series, about price of a commodity during the month of Ramzan, analyze it using 3-decade moving average method.

Decade Of Ramzan	Year			
	2017	2018	2019	2020
1	110	125	150	200
2	95	105	125	150
3	100	90	120	120

Old papers

1. A J & K university, BSc. A/2012

Compute five days moving average for the following records of attendances

Weak	Monday	Tuesday	Wednesday	Thursday	Friday
1	24	55	22	48	52
2	27	52	32	53	53
3	28	54	34	45	55

2. A J & K university, BSc. A/2013

Given the following records of attendances, determine the linear trend using least square method by fitting a straight line to the given data.

Weak	Sun	Mon	Tue	Wed	Thur	Fri	Sat
1	24	55	29	48	52	55	61
2	27	52	32	43	53	56	65

3. A J & K university, BSc. A/2015

From the following records of attendances, Fit a straight line using least square method and find the trend values.

Weak	Mon	Tue	Wed	Thur	Fri
I	55	22	48	52	59
II	52	32	43	53	56
III	59	36	42	50	55

4. Peshawar university, BSc. A/2011

Determine the linear trend from the following data showing sales of cement (in thousands tons) by a cement agency. records of attendances, also find the trend values.

Year	Quarters			
	I	II	III	IV
1974	36	25	39	48
1975	37	34	36	45
1976	48	40	30	42

5. Peshawar university, BSc. A/2012

Given below are figures of production (in Lakh Kg) of a sugar factory. Determine the linear trend from the following data showing sales of cement (in thousands tons) by a cement agency, records of attendances, also find the trend values.

Year	1971	1972	1973	1974	1975	1976	1977
Production	40	45	46	42	47	50	46

Fit a straight line trend by the least squares method and tabulate the trend.

6. Peshawar university, BSc. A/2013

Given below are the enrollments in a city college in different years.

Year	2005	2006	2007	2008	2009	2010
No. of students	500	550	570	600	650	700

Fit a straight line trend by the least squares and find the trend values.

7. Peshawar university, BSc. A/2014

Given below are the estimated growth rates of a locality in different years.

Year	2005	2006	2007	2008	2009	2010
Growth rate	2.3	2.2	2.1	1.8	1.6	11.5

Fit a straight line to the data by the least squares and find the trend values.

Solution

Q No. 13.3: Semi Averages: $\bar{X}_1 = 2, \bar{X}_2 = 8, \bar{Y}_1 = 12.54, \bar{Y}_2 = 18.24.$

$$\frac{Y-\bar{Y}_1}{\bar{Y}_2-\bar{Y}_1}=\frac{X-\bar{X}_1}{\bar{X}_2-\bar{X}_1} \Rightarrow \frac{Y-12.54}{18.24-12.54}=\frac{X-2}{8-2} \Rightarrow \hat{Y}=10.64+0.95X$$

Q No. 13.4: 4 quarter central moving average is

63.625, 62.25, 61.5, 62.125, 61.75, 62, 62.125, 60.125, 58.75, 58.625, 59, 59.5,

Q No. 13.5 3 decade moving average

101.67, 106.67, 110, 106.67, 115, 121.67, 131.67, 148.33, 156.67, 156.67

Additional Exercise

Some theorems for BS Mathematics only

Measure of Central Tendency and Dispersion

Q 1: State and prove properties of arithmetic mean.

Properties of arithmetic mean

(i) AM of a constant is constant itself i.e. If $X = C$ then $\bar{X} = C$

Proof: $\bar{X} = \frac{\sum X}{n} = \frac{\sum C}{n} = \frac{nC}{n} = C$

(ii) Sum of the deviations of observations from their AM is zero. i.e. $\Sigma(X - \bar{X}) = 0$

Proof: $\sum(X - \bar{X}) = \sum(X) - \sum(\bar{X}) = \sum(X) - n\bar{X} = \sum(X) - \sum(X) = 0$

(iii) Sum of the squared deviations of the observations from their AM is least. i.e. $\Sigma(X - \bar{X})^2 < \Sigma(X - A)^2$

Proof:

$$\sum(X - A)^2 = \sum(X - A + \bar{X} - \bar{X})^2 = \sum\left([X - \bar{X}] + [\bar{X} - A]\right)^2$$

$$= \sum\left[[X - \bar{X}]^2 + [\bar{X} - A]^2 + 2[\bar{X} - A][X - \bar{X}]\right]$$

$$= \sum[X - \bar{X}]^2 + \sum[\bar{X} - A]^2 + \sum 2[\bar{X} - A][X - \bar{X}]$$

$$= \sum[X - \bar{X}]^2 + n[\bar{X} - A]^2 + 2[\bar{X} - A]\sum[X - \bar{X}]$$

$$= \sum[X - \bar{X}]^2 + n[\bar{X} - A]^2 + 2[\bar{X} - A](0)$$

$$= \sum[X - \bar{X}]^2 + n[\bar{X} - A]^2 \text{ Ignoring } n[\bar{X} - A]^2$$

$$\sum(X - A)^2 > \sum[X - \bar{X}]^2$$

(iv) AM is affected by both the change of origin and scale. If $Y = aX + b$ then $\bar{Y} = a\bar{X} + b$.

Proof:

$$\bar{Y} = \frac{\sum Y}{n} = \frac{\sum(aX + b)}{n} = \frac{a\sum X + nb}{n} = a\bar{X} + b$$

Q 2: Prove that $\sum(X - a)^2 = \sum(X - \bar{X})^2 + n(\bar{X} - a)^2$

Solution:

$$\sum(X - a)^2 = \sum(X - \bar{X} + \bar{X} - a)^2 = \sum\left[(X - \bar{X}) + (\bar{X} - a)\right]^2$$

$$= \sum\left[(X - \bar{X})^2 + (\bar{X} - a)^2 + 2(X - \bar{X})(\bar{X} - a)\right]$$

$$= \sum(X - \bar{X})^2 + n(\bar{X} - a)^2 + 2(\bar{X} - a)\sum(X - \bar{X})$$

$$= \sum(X - \bar{X})^2 + n(\bar{X} - a)^2 + 0 \qquad \text{Where } \sum(X - \bar{X}) = 0$$

$$= \sum(X - \bar{X})^2 + n(\bar{X} - a)^2$$

Q 3: If $x_1, x_2, \ldots, x_n$ are n positive values of a variable X with geometric mean G. Then

$$G^n = \prod_{i=1}^{n}(x_i)$$

Proof: We have

$$G = \left\{\prod_{i=1}^{n}(x_i)\right\}^{\frac{1}{n}} \text{ Taking power n to the both sides}$$

$$G^n = \left[\left\{\prod_{i=1}^{n}(x_i)\right\}^{\frac{1}{n}}\right]^n \qquad G^n = \prod_{i=1}^{n}(x_i)$$

Q 4: If $x_1, x_2, \ldots, x_n$ are n positive values of a variable x and a is constant, then (a) If $X = a$ then $G = a$ (b) If $Y = aX$ then $G_Y = aG_X$

Proof:

(a) $G=\left\{\prod_{i=1}^{n}(x_i)\right\}^{\frac{1}{n}}=\left\{\prod_{i=1}^{n}(a)\right\}^{\frac{1}{n}}=(a.a\ldots a)^{\frac{1}{n}}=\left(a^n\right)^{\frac{1}{n}}=a$

(b) $G_1=\left\{\prod_{i=1}^{n}(x_i)\right\}^{\frac{1}{n}}=\left\{\prod_{i=1}^{n}(ax_i)\right\}^{\frac{1}{n}}=\left\{a^n\prod_{i=1}^{n}(x_i)\right\}^{\frac{1}{n}}=a\left\{\prod_{i=1}^{n}(x_i)\right\}^{\frac{1}{n}}=aG_x$

$G_1=aG_x$

Q.5: If $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be two series of n positive values each, then

(a) If $Z_i=X_iY_i$ then $G_z=G_xG_y$

(b) If $Z_i=\frac{X_i}{Y_i}$ then $G_z=\frac{G_x}{G_y}$

Proof:

(a) $G_z=\left\{\prod_{i=1}^{n}(Z_i)\right\}^{\frac{1}{n}}=\left\{\prod_{i=1}^{n}(x_iy_i)\right\}^{\frac{1}{n}}=\left\{\prod_{i=1}^{n}(x_i)\prod_{i=1}^{n}(y_i)\right\}^{\frac{1}{n}}=\left\{\prod_{i=1}^{n}(x_i)\right\}^{\frac{1}{n}}\left\{\prod_{i=1}^{n}(y_i)\right\}^{\frac{1}{n}}=G_xG_y$

$G_z=G_xG_y$

That is, it makes no difference whether we take the geometric mean of the products or the product of the geometric means

(b) $G_z=\left\{\prod_{i=1}^{n}(Z_i)\right\}^{\frac{1}{n}}$

$=\left\{\prod_{i=1}^{n}\left(\frac{x_i}{y_i}\right)\right\}^{\frac{1}{n}}=\frac{\left\{\prod_{i=1}^{n}(x_i)\right\}^{\frac{1}{n}}}{\left\{\prod_{i=1}^{n}(y_i)\right\}^{\frac{1}{n}}}=\frac{G_x}{G_y}$ $\qquad G_z=\frac{G_x}{G_y}$

Q. 6 Recurrence formula for HM is $H_n=\left(\frac{1}{nx_n}+\frac{n-1}{n}H_{n-1}^{-1}\right)^{-1}$

If HM of 10 values is 19.4, and a new value 25 is added in the data set, then find the HM of these 11 values.

Solution: Recurrence formula for HM is

$$H_n=\left(\frac{1}{nx_n}+\frac{n-1}{n}H_{n-1}^{-1}\right)^{-1}$$

$$H_{11}=\left(\frac{1}{11\times25}+\frac{10}{11}(19.4)^{-1}\right)^{-1}=\left(\frac{1}{275}+\frac{10}{11}\times\frac{1}{(19.4)}\right)^{-1}$$

Q 7: Prove that $AM\geq GM\geq HM$

Proof: $AM=\frac{1}{n}\sum_i X_i$ *and* $\log AM=\log\left(\frac{1}{n}\sum_i X_i\right)$

$GM=\left\{\prod_i(X_i)\right\}^{\frac{1}{n}}$ *and* $\log GM=\left(\frac{1}{n}\sum_i\log X_i\right)$

$HM=\frac{1}{\frac{1}{n}\sum_i\left(\frac{1}{X_i}\right)}=\left\{\frac{1}{n}\sum_i\left(\frac{1}{X_i}\right)\right\}^{-1}$ *and* $\log HM=-\log\left(\frac{1}{n}\sum_i\left(\frac{1}{X_i}\right)\right)$

According to Jenson's inequality

$$\log AM=\log\left(\frac{1}{n}\sum_i X_i\right)\geq\frac{1}{n}\sum_i\log X_i=\log GM$$

$\Rightarrow AM\geq GM$

Similarly using Jenson's inequality

$$\log HM=-\log\left(\frac{1}{n}\sum_i\left(\frac{1}{X_i}\right)\right)\leq-\frac{1}{n}\sum_i\log\left(\frac{1}{X_i}\right)=\frac{1}{n}\sum_i\log X_i=\log GM$$

$\Rightarrow GM\geq HM$

Hence $\Rightarrow AM\geq GM\geq HM$

Q 8: Define arithmetic mean, geometric mean and harmonic mean; and prove that for any two positive numbers a and b, A.M. ≥ G.M. ≥ H.M.

Additional Exercise

Q 9: Find the arithmetic mean geometric mean and harmonic mean of the series

a) $1,2,4,\ldots,2^n$

b) $1,3,9,\ldots,3^n$

c) a, ar, ar^2, ar^3

Q 10: Show that the A.M of the first n natural numbers is $\frac{n+1}{2}$.

Q 11: Show that the weighted arithmetic mean of first n natural numbers when weights are equal to the corresponding numbers is $\frac{2n+1}{3}$.

Q 12: In a certain examination, the average grade of all students in class A is 68.4 and that of all students in class B is 71.2. If the average of both classes combined is 70, find the ratio of the number of students in class A to the number in class B.

Q 13: An incomplete distribution of families according to their expenditure per week is given below. The median and mode for distribution are Rs.25 and Rs.24 respectively. Calculate missing frequencies.

Expenditure	0-10	10-20	20-30	30-40	40-50
Frequency	14	?	27	?	15

Q 14: In a frequency table, the upper boundary of each class interval has a constant ratio to the lower boundary. Show that the Geometric mean (G) may be expressed by the formula $\log G = L_0 + \frac{L}{N}\sum f(i-1)$.

Where L_0 is the logarithm of the mid value of the first interval and L is the logarithm of the ratio between upper and the lower boundaries.

Q 15: If n_1 and n_2 are the sizes, G_1 and G_2 the geometric mean of two series respectively, then the geometric mean G of the combined series is given by $\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$

Q 16: Find mean of the series $a, a+d, a+2d, \ldots, a+(n-1)d$ (ii) 4, 9, 14, . . . 99

Additional Exercise

Q 17: State and prove the properties of variance

Solution: Properties of variance:

(i) Variance of a constant is zero. Var© = 0

Proof: $\text{var}(c) = \frac{\Sigma(c-\bar{c})^2}{n} = \frac{\Sigma(c-c)^2}{n} = 0$

(ii) Variance is not affected by change of origin, that is if $Y = X + b$ Then $\text{var}(Y) = \text{var}(X)$

Proof: $\text{var}(Y) = \frac{\Sigma(Y-\bar{Y})^2}{n} = \frac{\Sigma\{X+b-(\bar{X}+b)\}^2}{n} = \frac{\Sigma(X-\bar{X})^2}{n} = \text{var}(X)$

(iii) Variance is affected by change of scale, that is if $Y = aX$ then $\text{var}(Y) = a^2\,\text{var}(X)$

Proof: $\text{var}(Y) = \frac{\Sigma(Y-\bar{Y})^2}{n} = \frac{\Sigma(aX-a\bar{X})^2}{n} = \frac{a^2\Sigma(X-\bar{X})^2}{n} = a^2\,\text{var}(X)$

(iv) $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$ if X and Y are independent.

Proof: $\text{var}(X+Y) = \frac{\Sigma\{(X+Y)-(\bar{X}+\bar{Y})\}^2}{n} = \frac{\Sigma(X+Y-\bar{X}-\bar{Y})^2}{n}$

$= \frac{\Sigma\{(X-\bar{X})+(Y-\bar{Y})\}^2}{n} = \frac{\Sigma\{(X-\bar{X})^2+(Y-\bar{Y})^2+2(X-\bar{X})(Y-\bar{Y})\}}{n}$

$= \frac{\Sigma(X-\bar{X})^2+\Sigma(Y-\bar{Y})^2+2\Sigma(X-\bar{X})(Y-\bar{Y})}{n}$

$= \frac{\Sigma(X-\bar{X})^2}{n} + \frac{\Sigma(Y-\bar{Y})^2}{n} + \frac{2\Sigma(X-\bar{X})(Y-\bar{Y})}{n}$

$= \text{var}(X) + \text{var}(Y) + 2\,\text{cov}(X,Y)$

$= \text{var}(X) + \text{var}(Y) \quad \because \text{cov}(X,Y) = 0$

(v) $\text{var}(X-Y) = \text{var}(X) + \text{var}(Y)$ if X and Y are independent

Proof:

$\text{var}(X-Y) = \frac{\Sigma\{(X-Y)-(\bar{X}-\bar{Y})\}^2}{n} = \frac{\Sigma(X-Y-\bar{X}+\bar{Y})^2}{n}$

$$= \frac{\Sigma[(X-\bar{X})-(Y-\bar{Y})]^2}{n} = \frac{\Sigma[(X-\bar{X})^2+(Y-\bar{Y})^2-2(X-\bar{X})(Y-\bar{Y})}{n}$$

$$= \frac{\Sigma(X-\bar{X})^2+\Sigma(Y-\bar{Y})^2-2\Sigma(X-\bar{X})(Y-\bar{Y})}{n}$$

$$= \frac{\Sigma(X-\bar{X})^2}{n}+\frac{\Sigma(Y-\bar{Y})^2}{n}-\frac{2\Sigma(X-\bar{X})(Y-\bar{Y})}{n} = \text{var}(X)+\text{var}(Y)-2\text{cov}(X,Y)$$

$$= \text{var}(X)+\text{var}(Y) \quad \because \text{cov}(X,Y)=0$$

(vi) Combined variance:

If a distribution consists of k components with $n_1, n_2, \ldots, n_k$ observations with $\sum n_i = n$ having means $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ and variances $S_1^2, S_2^2, \ldots, S_k^2$, then the combined variance S^2 of all n observations is given by $S^2 = \dfrac{\sum_{i=1}^{k} n_i\left[S_i^2+(\bar{X}_i-\bar{X})^2\right]}{\sum_{i=1}^{k} n_i}$ OR

$S^2 = \dfrac{n_1[S_1^2+(\bar{X}_1-\bar{X})^2]+n_2[S_2^2+(\bar{X}_2-\bar{X})^2]+\ldots+n_k[S_k^2+(\bar{X}_k-\bar{X})^2]}{n_1+n_2+\ldots+n_k}$; Where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{X}_i$$

Proof: Let $X_{ij}, i=1,2,\ldots,n_i$ and $j=1,2,\ldots,k$ be the i^{th} observation in the j^{th} component consisting of n_i observations with mean and variance

$$\bar{X}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} X_{ij} \qquad S_i^2 = \frac{1}{n_i}\sum_{i=1}^{n_i}(X_{ij}-\bar{X}_i)^2$$

The combined variance of all $\sum n_i = n$ observations is

$$S^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X})^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j+\bar{X}_j-\bar{X})^2$$

$$= \frac{1}{n}\sum_{j=1}^{k}\left[\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2+\sum_{i=1}^{n_j}(\bar{X}_j-\bar{X})^2+2(\bar{X}_j-\bar{X})\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)\right]$$

$$= \frac{1}{n}\sum_{j=1}^{k}\left\{n_jS_j^2+n_j(\bar{X}_j-\bar{X})^2\right\} = \frac{1}{n}\sum_{j=1}^{k} n_j\left\{S_j^2+(\bar{X}_j-\bar{X})^2\right\} \Rightarrow S^2 = \frac{\sum_{j=1}^{k} n_j\left[S_j^2+(\bar{X}_j-\bar{X})^2\right]}{\sum_{j=1}^{k} n_j}$$

Since $\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2 = n_jS_j^2$, $\quad \sum_{j=1}^{k}(\bar{X}_j-\bar{X})^2 = n_j(\bar{X}_j-\bar{X})^2$ and

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2 = \sum_{j=1}^{k}(\bar{X}_j-\bar{X})^2 = 0$$

Q 18: If $\bar{X}_1$ and S_1^2 are mean and variance of n_1 observations and $\bar{X}_2$ and S_2^2 are mean and variance of n_2 observations then prove that variance of n_1+n_2 observations is $S^2 = \dfrac{n_1S_1^2+n_2S_2^2}{n_1+n_2}+\dfrac{n_1n_2(\bar{X}_1-\bar{X}_2)}{(n_1+n_2)^2}$

Proof: For $k=2$

$$S^2 = \frac{n_1[S_1^2+(\bar{X}_1-\bar{X})^2]+n_2[S_2^2+(\bar{X}_2-\bar{X})^2]}{n_1+n_2}$$

$$= \frac{n_1S_1^2+n_1(\bar{X}_1-\bar{X})^2+n_2S_2^2+n_2(\bar{X}_2-\bar{X})^2}{n_1+n_2}$$

$$= \frac{n_1S_1^2+n_2S_2^2}{n_1+n_2}+\frac{n_1(\bar{X}_1-\bar{X})^2+n_2(\bar{X}_2-\bar{X})^2}{n_1+n_2}$$

$$= \frac{n_1S_1^2+n_2S_2^2}{n_1+n_2}+\frac{n_1\left(\bar{X}_1-\frac{n_1\bar{X}_1+n_2\bar{X}_2}{n_1+n_2}\right)^2+n_2\left(\bar{X}_2-\frac{n_1\bar{X}_1+n_2\bar{X}_2}{n_1+n_2}\right)^2}{n_1+n_2}$$

$$= \frac{nS^2+nS^2}{n+n}+\frac{n\left[\frac{(n_1+n_2)\bar{X}_1-(n_1\bar{X}_1+n_2\bar{X}_2)}{n_1+n_2}\right]^2+n_2\left[\frac{(n_1+n_2)\bar{X}_2-(n_1\bar{X}_1+n_2\bar{X}_2)}{n_1+n_2}\right]^2}{n_1+n_2}$$

$$= \frac{nS^2+nS^2}{n+n}+\frac{n_1[(n_1+n_2)\bar{X}_1-(n_1\bar{X}_1+n_2\bar{X}_2)]^2+n_2[(n_1+n_2)\bar{X}_2-(n_1\bar{X}_1+n_2\bar{X}_2)]^2}{(n_1+n_2)^2}$$

$$= \frac{nS^2+nS^2}{n+n}+\frac{n_1(n_1\bar{X}_1-n_2\bar{X}_2)^2+n_2(n_1\bar{X}_1-n_1\bar{X}_2)^2}{(n_1+n_2)^2}$$

$$\frac{n_1\bar{X}_1+n_2\bar{X}_2}{n_1+n_2}=\frac{n_1\bar{X}_1+n_1\bar{X}_2-n_1\bar{X}_2+n_2\bar{X}_2}{n_1+n_2}=\frac{n_1(\bar{X}_1-\bar{X}_2)+n_2(\bar{X}_2-\bar{X}_2)}{n_1+n_2}=\bar{X}_2+\frac{n_1(\bar{X}_1-\bar{X}_2)}{n_1+n_2}$$

$$\frac{n_1S_1^2+n_2S_2^2}{n_1+n_2}=\frac{n_1S_1^2+n_1S_2^2-n_1S_2^2+n_2S_2^2}{n_1+n_2}=\frac{n_1(S_1^2-S_2^2)+(n_1+n_2)S_2^2}{n_1+n_2}=S_2^2+\frac{n_1(S_1^2-S_2^2)}{n_1+n_2}$$

Q 19: Prove that for two observations range is double of standard deviation.

Q 20: Prove that MD ≤ SD.

Q 21: Find variance and S.D of first n- natural number.

Q 22: The SD of 10 observations on a certain variable was calculated as 16.2. It was later discovered that one of the observation was wrongly recorded as 12.9 in fact it was 21.9, correct the SD.

Regression And Correlation

Q 23: Estimate the values of "a" and "b" in the simple linear regression $\hat{Y}=a+bX$ line using ordinary least square estimation (OLS) method.

Solution: Consider a set of n pairs of observations $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ to which we wish to fit a straight line $\hat{Y}=a+bX$ where "a" is intercept and "b" is slope of line. Let "S" donate the sum of squares of residuals, then

$$S=\sum_{i=1}^{n}\left(Y-\hat{Y}\right)^2=\sum_{i=1}^{n}(Y-a-bX)^2$$

For a minimum value of **"S"**, the partial derivatives $\frac{\partial S}{\partial a}$ and $\frac{\partial S}{\partial b}$ must be zero. Thus

$$\frac{\partial S}{\partial a}=\frac{\partial}{\partial a}\sum_{i=1}^{n}(Y-a-bX)^2=2\sum_{i=1}^{n}(Y-a-bX)(-1)=0$$

$$\Rightarrow\sum_{i=1}^{n}(Y-a-bX)=0\Rightarrow\sum_{i=1}^{n}Y-\sum_{i=1}^{n}a-\sum_{i=1}^{n}bX=0$$

$$\sum_{i=1}^{n}Y=na+\sum_{i=1}^{n}bX\quad\text{----(i)}$$

$$\frac{\partial S}{\partial b}=\frac{\partial}{\partial b}\sum_{i=1}^{n}(Y-a-bX)^2=2\sum_{i=1}^{n}(Y-a-bX)(-X)=0\Rightarrow\sum_{i=1}^{n}(YX-aX-bX^2)=0$$

$$\sum_{i=1}^{n}XY-a\sum_{i=1}^{n}X-b\sum_{i=1}^{n}X^2=0$$

$$\sum_{i=1}^{n}XY=a\sum_{i=1}^{n}X+b\sum_{i=1}^{n}X^2\quad\text{----(ii)}$$

Dividing the eq. (i) by n on both sides

$$\frac{\sum_{i=1}^{n}Y}{n}=a+b\frac{\sum_{i=1}^{n}X}{n}$$

$$\bar{Y}=a+b\bar{X}\Rightarrow a=\bar{Y}-b\bar{X}$$

It shows the least squares line passes through $(\bar{X},\bar{Y})$

By substituting the value of "a" in equation (ii)

$$\sum_{i=1}^{n}XY-(\bar{Y}-b\bar{X})\sum_{i=1}^{n}X-b\sum_{i=1}^{n}X^2=0$$

$$\sum_{i=1}^{n}XY-\bar{Y}\sum_{i=1}^{n}X+b\bar{X}\sum_{i=1}^{n}X-b\sum_{i=1}^{n}X^2=0\Rightarrow b\sum_{i=1}^{n}X^2-b\bar{X}\sum_{i=1}^{n}X=\sum_{i=1}^{n}XY-\bar{Y}\sum_{i=1}^{n}X$$

$$b\left\{\sum_{i=1}^{n}X^2-\bar{X}\sum_{i=1}^{n}X\right\}=\sum_{i=1}^{n}XY-\bar{Y}\sum_{i=1}^{n}X$$

$$b\sum_{i=1}^{n}(X-\bar{X})^2=\sum_{i=1}^{n}(X-\bar{X})(Y-\bar{Y})$$

$$b=\frac{\sum_{i=1}^{n}(X-\bar{X})(Y-\bar{Y})}{\sum_{i=1}^{n}(X-\bar{X})^2}$$

Hence $b=\frac{n\sum XY-\sum X\sum Y}{n\sum X^2-(\sum X)^2}$ and $a=\bar{Y}-b\bar{X}$

Additional Exercise

Q 24: Write the equation of a straight line through the origin and derive an expression for finding its slope by the principal of least squares.

Proof: The equation of straight line through the origin is

$$\hat{y} = bx$$

Let S denote the sum of squares of residuals

Then $S = \sum_{i=1}^{n}(y-\hat{y})^2 = \sum_{i=1}^{n}(y-bx)^2$

S will be minimum if $\frac{\partial S}{\partial b} = 0$

$$\frac{\partial S}{\partial b} = \frac{\partial}{\partial b}\left(\sum_{i=1}^{n}(y-bx)^2\right) = 0$$

$$2\sum_{i=1}^{n}(y-bx)(-x) = 0 \Rightarrow \sum_{i=1}^{n}(xy - bx^2) = 0$$

$$b\sum_{i=1}^{n}x^2 = \sum_{i=1}^{n}xy \Rightarrow b = \frac{\sum_{i=1}^{n}xy}{\sum_{i=1}^{n}x^2}$$

Q 25: If a variable X have to values x_1 and x_2 and Y have y_1 and y_2. Prove that $b = \frac{y_2 - y_1}{x_2 - x_1}$.

Q 26: Prove that $-1 \le r \le +1$

Proof: Consider $x = X - \bar{X}$ and $y = Y - \bar{Y}$

As $S_Y^2 = \frac{\sum(Y-\bar{Y})^2}{n} = \frac{\sum y^2}{n} \Rightarrow \frac{\sum y^2}{S_Y^2} = n$ Similarly $\frac{\sum x^2}{S_X^2} = n$

Also $r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}} = \frac{\sum xy}{\sqrt{\sum x^2\sum y^2}}$

Additional Exercise

$$r = \frac{\sum xy}{\sqrt{nS_X^2 nS_Y^2}} = \frac{\sum xy}{nS_XS_Y} \Rightarrow rn = \frac{\sum xy}{S_XS_Y}$$

Consider $\sum\left(\frac{x}{S_X} - \frac{y}{S_Y}\right)^2 > 0 \Rightarrow \sum\left(\frac{x^2}{S_x^2} + \frac{y^2}{S_y^2} - 2\frac{xy}{S_XS_Y}\right) > 0$

$$\frac{\sum x^2}{S_X^2} + \frac{\sum y^2}{S_Y^2} - 2\frac{\sum xy}{S_XS_Y} > 0 \Rightarrow n + n - 2rn > 0 \Rightarrow 2n - 2rn > 0$$

$$2n(1-r) > 0$$

$$2n > 0 \text{ so } (1-r) > 0 \Rightarrow 1 - r > 0 \Rightarrow r < 1 \text{ ---(i)}$$

Similarly by considering $\sum\left(\frac{x}{S_X} + \frac{y}{S_Y}\right)^2 > 0$ result can be obtained as $r > 1$ ---(ii)

By combining these two above equations (i) and (ii) we can write that $-1 \le r \le +1$

Q 27: Prove that $r_{uv} = r_{xy}$ where $u = \frac{X-A}{h}$ and $v = \frac{Y-B}{k}$

Proof: As $r_{xy} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}}$

If $u = \frac{X-A}{h}$ and $v = \frac{Y-B}{k}$ then $X = hu + A$ and $Y = kv + B$ also

$\bar{X} = h\bar{u} + A$ And $\bar{Y} = k\bar{v} + B$ So

$$r_{uv} = \frac{\sum(hu + A - h\bar{u} - A)(kv + B - k\bar{v} - B)}{\sqrt{\sum(hu + A - h\bar{u} - A)^2\sum(kv + B - k\bar{v} - B)^2}}$$

$$= \frac{\sum(hu - h\bar{u})(kv - k\bar{v})}{\sqrt{\sum(hu - h\bar{u})^2\sum(kv - k\bar{v})^2}} = \frac{\sum hk(u-\bar{u})k(v-\bar{v})}{\sqrt{\sum h^2(u-\bar{u})^2\sum k^2(v-\bar{v})^2}}$$

$$= \frac{hk\sum(u-\bar{u})(v-\bar{v})}{hk\sqrt{\sum(u-\bar{u})^2\sum(v-\bar{v})^2}} = \frac{\sum(u-\bar{u})(v-\bar{v})}{\sqrt{\sum(u-\bar{u})^2\sum(v-\bar{v})^2}} \Rightarrow r_{uv} = r_{xy}$$

Q 28: The correlation co-efficient is the geometric mean of two regression co-efficient

Proof:

The regression co-efficient are Written as $b_{xy} = r\frac{S_x}{S_y}$; $b_{yx} = r\frac{S_y}{S_x}$

$b_{xy} \times b_{yx} = r\frac{S_x}{S_y} \times r\frac{S_y}{S_x} = r^2$ Taking square root on both sides

$r = \pm\sqrt{b_{xy} \times b_{yx}}$

Sign of correlation coefficient depends on the Signs of regression co-efficient.

Q 29: Derive formula for the rank correlation coefficient.

Solution: Derivation of rank coefficient

Let a set of n objects be ranked with respect to character A $X_1, X_2, \ldots, X_{n-1}, X_n$ and according to character B as $Y_1, Y_2, \ldots, Y_{n-1}, Y_n$ assuming no tie in the ranks. Therefore

$$\sum X = \sum Y = 1+2+3+\ldots+n = \frac{n(n+1)}{2}$$

$$\sum X^2 = \sum Y^2 = 1^2+2^2+3^2+\ldots+n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum (X-\bar{X})^2 = \sum (Y-\bar{Y})^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{n} = \frac{n(n+1)(2n+1)}{6} - \frac{\left(\frac{n(n+1)}{2}\right)^2}{n}$$

$$\frac{n(n+1)}{2}\left\{\frac{(2n+1)}{3} - \frac{(n+1)}{2}\right\} = \frac{n(n+1)}{2}\left\{\frac{2(2n+1)-3(n+1)}{6}\right\}$$

$$\frac{n(n+1)}{2}\left\{\frac{4n+2-3n-3}{6}\right\} = \frac{n(n+1)}{2}\left\{\frac{n-1}{6}\right\} = \frac{n(n^2-1)}{12}$$

$$\sum d^2 = \sum (X-Y)^2 = \sum X^2 + \sum Y^2 - 2\sum XY$$

$$\frac{n(n+1)(2n+1)}{6} + \frac{n(n+1)(2n+1)}{6} - 2\sum XY = \frac{n(n+1)(2n+1)}{3} - 2\sum XY$$

$$\sum XY = \frac{n(n+1)(2n+1)}{3} - \sum d^2 = \sum XY = \frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum d^2$$

$$\sum (X-\bar{X})(Y-\bar{Y}) = \sum XY - \frac{\sum X \sum Y}{n} = \frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum d^2 - \frac{\frac{n(n+1)}{2}\frac{n(n+1)}{2}}{n}$$

$$= \frac{n(n+1)}{2}\left(\frac{(2n+1)}{3} - \frac{n(n+1)}{2}\right) - \frac{1}{2}\sum d^2$$

$$\sum (X-\bar{X})(Y-\bar{Y}) = \frac{n(n^2-1)}{12} - \frac{1}{2}\sum d^2$$

$$r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}} = \frac{\frac{n(n^2-1)}{12} - \frac{1}{2}\sum d^2}{\sqrt{\frac{n(n^2-1)}{12}\frac{n(n^2-1)}{12}}}$$

$$r = \frac{\frac{n(n^2-1)}{12} - \frac{1}{2}\sum d^2}{\frac{n(n^2-1)}{12}} = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Q 30: Prove that acute angle θ between two regression lines y on x and x on y is given as: $\tan(\theta) = \frac{r}{1-r^2}\left[\frac{|\sigma_x^2 - \sigma_y^2|}{\sigma_x \sigma_y}\right]$

Q 31: Give that x=4y+5, y=kx+4 are the regression lines, show that. If $0 < 4k < 1$, $k = \frac{1}{16}$ Find the mean of the two variables and the co-efficient of correlation.

Q 32: for two variables x and y with the same means, two regression equations are $Y = aX + b$ and $X = \alpha bY + \beta$. Show that $\frac{b}{\beta} = \frac{1-a}{1-\alpha b}$ Find also line of the two variables.

STATISTICAL

TABLES

z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)
-∞	to	-3.62	0.0001	-3.61	to	-3.49	0.0002		
-3.48	to	-3.39	0.0003	-3.38	to	-3.33	0.0004		
-3.32	to	-3.27	0.0005	-3.26	to	-3.22	0.0006		
-3.21	to	-3.18	0.0007	-3.17	to	-3.14	0.0008		
-3.13	to	-3.10	0.0009	-3.09	0.0010	-3.08	0.0010		
-3.07	to	-3.05	0.0011	-3.04	0.0012	-3.03	0.0012		
-3.02	to	-3.00	0.0013	-2.99	0.0014	-2.98	0.0014	-2.97	0.0015
-2.96	0.0015	-2.95	0.0016	-2.94	0.0016	-2.93	0.0017	-2.92	0.0018
-2.91	0.0018	-2.90	0.0019	-2.89	0.0019	-2.88	0.0020	-2.87	0.0021
-2.86	0.0021	-2.85	0.0022	-2.84	0.0023	-2.83	0.0023	-2.82	0.0024
-2.81	0.0025	-2.80	0.0026	-2.79	0.0026	-2.78	0.0027	-2.77	0.0028
-2.76	0.0029	-2.75	0.0030	-2.74	0.0031	-2.73	0.0032	-2.72	0.0033
-2.71	0.0034	-2.70	0.0035	-2.69	0.0036	-2.68	0.0037	-2.67	0.0038
-2.66	0.0039	-2.65	0.0040	-2.64	0.0041	-2.63	0.0043	-2.62	0.0044
-2.61	0.0045	-2.60	0.0047	-2.59	0.0048	-2.58	0.0049	-2.57	0.0051
-2.56	0.0052	-2.55	0.0054	-2.54	0.0055	-2.53	0.0057	-2.52	0.0059
-2.51	0.0060	-2.50	0.0062	-2.49	0.0064	-2.48	0.0066	-2.47	0.0068
-2.46	0.0069	-2.45	0.0071	-2.44	0.0073	-2.43	0.0075	-2.42	0.0078
-2.41	0.0080	-2.40	0.0082	-2.39	0.0084	-2.38	0.0087	-2.37	0.0089
-2.36	0.0091	-2.35	0.0094	-2.34	0.0096	-2.33	0.0099	-2.32	0.0102
-2.31	0.0104	-2.30	0.0107	-2.29	0.0110	-2.28	0.0113	-2.27	0.0116
-2.26	0.0119	-2.25	0.0122	-2.24	0.0125	-2.23	0.0129	-2.22	0.0132
-2.21	0.0136	-2.20	0.0139	-2.19	0.0143	-2.18	0.0146	-2.17	0.0150
-2.16	0.0154	-2.15	0.0158	-2.14	0.00162	-2.13	0.0166	-2.12	0.170
-2.11	0.0174	-2.10	0.0179	-2.09	0.0183	-2.08	0.0188	-2.07	0.0192
-2.06	0.0197	-2.05	0.0202	-2.04	0.0207	-2.03	0.0212	-2.02	0.0217
-2.01	0.0222	-2.00	0.0228	-1.99	0.0233	-1.98	0.0239	-1.97	0.0244
-1.96	0.0250	-1.95	0.0256	-1.94	0.0262	-1.93	0.0268	-1.92	0.0274
-1.91	0.0281	-1.90	0.0287	-1.89	0.0294	-1.88	0.0301	-1.87	0.0307
-1.86	0.0314	-1.85	0.0322	-1.84	0.0329	-1.83	0.0336	-1.82	0.0344
-1.81	0.0351	-1.80	0.0359	-1.79	0.0367	-1.78	0.0375	-1.77	0.0384
-1.76	0.0392	-1.75	0.0401	-1.74	0.0409	-1.73	0.0418	-1.72	0.0427
-1.71	0.0436	-1.70	0.0446	-1.69	0.0455	-1.68	0.0465	-1.67	0.0475
-1.66	0.0485	-1.65	0.0495	-1.64	0.0505	-1.63	0.0516	-1.62	0.0526
-1.61	0.0537	-1.60	0.0548	-1.59	0.0559	-1.58	0.0571	-1.57	0.0582
-1.56	0.0594	-1.55	0.0606	-1.54	0.0618	-1.53	0.0630	-1.52	0.0643
-1.51	0.0655	-1.50	0.0668	-1.49	0.0681	-1.48	0.0694	-1.47	0.0708
-1.46	0.0721	-1.45	0.0735	-1.44	0.0749	-1.43	0.0764	-1.42	0.0778
-1.41	0.0793	-1.40	0.0808	-1.39	0.0823	-1.38	0.0838	-1.37	0.0853
-1.36	0.0869	-1.35	0.0885	-1.34	0.0901	-1.33	0.0918	-1.32	0.0934
-1.31	0.0951	-1.30	0.0968	-1.29	0.0985	-1.28	0.1003	-1.27	0.1020
-1.26	0.1038	-1.25	0.1056	-1.24	0.1075	-1.23	0.1093	-1.22	0.1112
-1.21	0.1131	-1.20	0.1151	-1.19	0.1170	-1.18	0.1190	-1.17	0.1210

Table 1: Cumulative standardized random variable

z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)
-1.16	0.1230	-1.15	0.1251	-1.14	0.1271	-1.13	0.1292	-1.12	0.1314
-1.11	0.1335	-1.10	0.1357	-1.09	0.1379	-1.08	0.1401	-1.07	0.1423
-1.06	0.1446	-1.05	0.1469	-1.04	0.1492	-1.03	0.1515	-1.02	0.1539
-1.01	0.1562	-1.00	0.1587	-0.99	0.1611	-0.98	0.1635	-0.97	0.1660
-0.96	0.1685	-0.95	0.1711	-0.94	0.1736	-0.93	0.1762	-0.92	0.1788
-0.91	0.1814	-0.90	0.1841	-0.89	0.1867	-0.88	0.1894	-0.87	0.1922
-0.86	0.1949	-0.85	0.1977	-0.84	0.2005	-0.83	0.2033	-0.82	0.2061
-0.81	0.2090	-0.80	0.2119	-0.79	0.2148	-0.78	0.2177	-0.77	0.2206
-0.76	0.2236	-0.75	0.2266	-0.74	0.2296	-0.73	0.2327	-0.72	0.2358
-0.71	0.2389	-0.70	0.2420	-0.69	0.2451	-0.68	0.2483	-0.67	0.2514
-0.66	0.2546	-0.65	0.2578	-0.64	0.2611	-0.63	0.2643	-0.62	0.2676
-0.61	0.2709	-0.60	0.2743	-0.59	0.2776	-0.58	0.2810	-0.57	0.2843
-0.56	0.2877	-0.55	0.2912	-0.54	0.2946	-0.53	0.2981	-0.52	0.3015
-0.51	0.3050	-0.50	0.3085	-0.49	0.3121	-0.48	0.3156	-0.47	0.3192
-0.46	0.3228	-0.45	0.3264	-0.44	0.3300	-0.43	0.3336	-0.42	0.3372
-0.41	0.3409	-0.40	0.3446	-0.39	0.3483	-0.38	0.3520	-0.37	0.3557
-0.36	0.3594	-0.35	0.3632	-0.34	0.3669	-0.33	0.3707	-0.32	0.3745
-0.31	0.3783	-0.30	0.3821	-0.29	0.3859	-0.28	0.3897	-0.27	0.3936
-0.26	0.3974	-0.25	0.4013	-0.24	0.4052	-0.23	0.4090	-0.22	0.4129
-0.21	0.4168	-0.20	0.4207	-0.19	0.4247	-0.18	0.4286	-0.17	0.4325
-0.16	0.4364	-0.15	0.4404	-0.14	0.4443	-0.13	0.4483	-0.12	0.4522
-0.11	0.4562	-0.10	0.4602	-0.09	0.4641	-0.08	0.4681	-0.07	0.4721
-0.06	0.4761	-0.05	0.4801	-0.04	0.4840	-0.03	0.4880	-0.02	0.4920
-0.01	0.4960	0.00	0.5000	0.01	0.5040	0.02	0.5080	0.03	0.5120
0.04	0.5160	0.05	0.5199	0.06	0.5239	0.07	0.5279	0.08	0.5319
0.09	0.5359	0.10	0.5398	0.11	0.5438	0.12	0.5478	0.13	0.5517
0.14	0.5557	0.15	0.5596	0.16	0.5636	0.17	0.5675	0.18	0.5714
0.19	0.5753	0.20	0.5793	0.21	0.5832	0.22	0.5871	0.23	0.5910
0.24	0.5948	0.25	0.5987	0.26	0.6026	0.27	0.6064	0.28	0.6103
0.29	0.6141	0.30	0.6179	0.31	0.6217	0.32	0.6255	0.33	0.6293
0.34	0.6331	0.35	0.6368	0.36	0.6406	0.37	0.6443	0.38	0.6480
0.39	0.6517	0.40	0.6554	0.41	0.6591	0.42	0.6628	0.43	0.6664
0.44	0.6700	0.45	0.6736	0.46	0.6772	0.47	0.6808	0.48	0.6844
0.49	0.6879	0.50	0.6915	0.51	0.6950	0.52	0.6985	0.53	0.7019
0.54	0.7054	0.55	0.7088	0.56	0.7123	0.57	0.7157	0.58	0.7190
0.59	0.7224	0.60	0.7257	0.61	0.7291	0.62	0.7324	0.63	0.7357
0.64	0.7389	0.65	0.7422	0.66	0.7454	0.67	0.7486	0.68	0.7517
0.69	0.7549	0.70	0.7580	0.71	0.7611	0.72	0.7642	0.73	0.7673
0.74	0.7704	0.75	0.7734	0.76	0.7764	0.77	0.7794	0.78	0.7823
0.79	0.7852	0.80	0.7881	0.81	0.7910	0.82	0.7939	0.83	0.7967
0.84	0.7995	0.85	0.8023	0.86	0.8051	0.87	0.8078	0.88	0.8106
0.89	0.8133	0.90	0.8159	0.91	0.8186	0.92	0.8212	0.93	0.8238
0.94	0.8264	0.95	0.8289	0.96	0.8315	0.97	0.8340	0.98	0.8365

z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)
0.99	0.8389	1.00	0.8413	1.01	0.8438	1.02	0.8461	1.03	0.8485
1.04	0.8508	1.05	0.8531	1.06	0.8554	1.07	0.8577	1.08	0.8599
1.09	0.8621	1.10	0.8643	1.11	0.8665	1.12	0.8686	1.13	0.8708
1.14	0.8729	1.15	0.8749	1.16	0.8770	1.17	0.8790	1.18	0.8810
1.19	0.8830	1.20	0.8849	1.21	0.8869	1.22	0.8888	1.23	0.8907
1.24	0.8925	1.25	0.8944	1.26	0.8962	1.27	0.8980	1.28	0.8997
1.29	0.9015	1.30	0.9032	1.31	0.9049	1.32	0.9066	1.33	0.9082
1.34	0.9099	1.35	0.9115	1.36	0.9131	1.37	0.9147	1.38	0.9162
1.39	0.9177	1.40	0.9192	1.41	0.9207	1.42	0.9222	1.43	0.9236
1.44	0.9251	1.45	0.9265	1.46	0.9279	1.47	0.9292	1.48	0.9306
1.49	0.9319	1.50	0.9332	1.51	0.9345	1.52	0.9357	1.53	0.9370
1.54	0.9382	1.55	0.9394	1.56	0.9406	1.57	0.9418	1.58	0.9429
1.59	0.9441	1.60	0.9452	1.61	0.9463	1.62	0.9474	1.63	0.9484
1.64	0.9495	1.65	0.9505	1.66	0.9515	1.67	0.9525	1.68	0.9535
1.69	0.9545	1.70	0.9554	1.71	0.9564	1.72	0.9564	1.73	0.9573
1.74	0.9591	1.75	0.9599	1.76	0.9608	1.77	0.9616	1.78	0.9625
1.79	0.9633	1.80	0.9641	1.81	0.9649	1.82	0.9656	1.83	0.9664
1.84	0.9671	1.85	0.9678	1.86	0.9686	1.87	0.9693	1.88	0.9699
1.89	0.9706	1.90	0.9713	1.91	0.9719	1.92	0.9726	1.93	0.9732
1.94	0.9738	1.95	0.9744	1.96	0.9750	1.97	0.9756	1.98	0.9761
1.99	0.9767	2.00	0.9772	2.01	0.9778	2.02	0.9783	2.03	0.9788
2.04	0.9793	2.05	0.9798	2.06	0.9803	2.07	0.9808	2.08	0.9812
2.09	0.9817	2.10	0.9821	2.11	0.9836	2.12	0.9830	2.13	0.9834
2.14	0.9838	2.15	0.9842	2.16	0.9846	2.17	0.9850	2.18	0.9854
2.19	0.9857	2.20	0.9861	2.21	0.9864	2.22	0.9868	2.23	0.9871
2.24	0.9875	2.25	0.9878	2.26	0.9881	2.27	0.9884	2.28	0.9887
2.29	0.9890	2.30	0.9893	2.31	0.9896	2.32	0.9898	2.33	0.9901
2.34	0.9904	2.35	0.9906	2.36	0.9909	2.37	0.9911	2.38	0.9913
2.39	0.9916	2.40	0.9918	2.41	0.9920	2.42	0.9922	2.43	0.9925
2.44	0.9927	2.45	0.9929	2.46	0.9931	2.47	0.9932	2.48	0.9934
2.49	0.9936	2.50	0.9938	2.51	0.9940	2.52	0.9941	2.53	0.9943
2.54	0.9945	2.55	0.9946	2.56	0.9948	2.57	0.9949	2.58	0.9951
2.59	0.952	2.60	0.9953	2.61	0.9955	2.62	0.9956	2.63	0.9957
2.64	0.9959	2.65	0.9960	2.66	0.9961	2.67	0.9962	2.68	0.9963
2.69	0.9964	2.70	0.9965	2.71	0.9966	2.72	0.9967	2.73	0.9968
2.74	0.9969	2.75	0.9970	2.76	0.9971	2.77	0.9972	2.78	0.9973
2.79	0.9974	2.80	0.9974	2.81	0.9975	2.82	0.9976	2.83	0.9977
2.84	0.9977	2.85	0.9978	2.86	0.9979	2.87	0.9979	2.88	0.9980
2.89	0.9981	2.90	0.9981	2.91	0.9982	2.92	0.9982	2.93	0.9983
2.94	0.9984	2.95	0.9984	2.96	0.9985	2.97	0.9985	2.98	0.9986
2.99	0.9986	3.00	to	3.02	0.9987	3.03	0.9988		
3.04	0.9988	3.05	to	3.07	0.9989	3.08	0.9990		
3.09	0.9990	3.10	0.9990	3.11	to	3.13	0.9991		

Table 1: Cumulative standardized random variable

Z	Φ(Z)	Z	Φ(Z)	Z	Φ(Z)	Z	Φ(Z)	Z
3.14	to	3.17	0.9992	3.18	to	3.21	0.9993	
3.22	to	3.26	0.9994	3.27	to	3.32	0.9995	
3.33	to	3.38	0.9996	3.39	to	3.46	0.9997	
3.49	to	3.61	0.9998	3.62	to	3.89	0.9999	
3.90	to	∞	1.0000					

Table 2: Critical value of Student's t distribution

d f	Level of significance (α)						
	0.10	0.05	0.04	0.02	0.025	0.01	0.005
1	3.078	6.314	7.916	15.895	12.706	31.821	63.657
2	1.886	2.920	3.320	4.849	4.303	6.965	9.925
3	1.638	2.353	2.605	3.482	3.182	4.541	5.841
4	1.533	2.132	2.333	2.998	2.776	3.746	4.602
5	1.476	2.015	2.191	2.756	2.571	3.365	4.032
6	1.440	1.943	2.104	2.612	2.447	3.143	3.707
7	1.415	1.895	2.046	2.517	2.364	2.998	3.499
8	1.397	1.860	2.004	2.449	2.306	2.896	3.355
9	1.383	1.833	1.973	2.398	2.262	2.821	3.250
10	1.372	1.812	1.948	2.359	2.228	2.764	3.169
11	1.363	1.796	1.982	2.328	2.200	2.718	3.106
12	1.356	1.782	1.912	2.303	2.179	2.681	3.054
13	1.350	1.771	1.899	2.282	2.160	2.650	3.012
14	1.345	1.761	1.887	2.264	2.145	2.624	2.977
15	1.341	1.753	1.878	2.249	2.131	2.602	2.945
16	1.337	1.746	1.869	2.235	2.120	2.583	2.921
17	1.333	1.740	1.862	2.224	2.110	2.567	2.898
18	1.330	1.734	1.855	2.213	2.101	2.552	2.878
19	1.328	1.729	1.850	2.205	2.093	2.539	2.861
20	1.325	1.725	1.844	2.200	2.086	2.528	2.845
21	1.323	1.721	1.840	2.189	2.080	2.518	2.831
22	1.321	1.717	1.835	2.183	2.074	2.508	2.819
23	1.319	1.714	1.832	2.177	2.069	2.500	2.807
24	1.318	1.711	1.828	2.172	2.064	2.492	2.797
25	1.316	1.708	1.825	2.167	2.060	2.485	2.787
26	1.315	1.706	1.822	2.162	2.056	2.479	2.779
27	1.314	1.703	1.819	2.158	2.052	2.473	2.771
28	1.313	1.701	1.817	2.154	2.048	2.467	2.763
29	1.311	1.699	1.814	2.150	2.045	2.462	2.756
30	1.310	1.697	1.812	2.147	2.042	2.457	2.750
31	1.309	1.696	1.810	2.144	2.040	2.453	2.744
32	1.309	1.694	1.808	2.141	2.037	2.449	2.738
33	1.308	1.692	1.806	2.138	2.034	2.445	2.733
34	1.307	1.691	1.805	2.134	2.032	2.441	2.728
35	1.306	1.690	1.801	2.133	2.030	2.438	2.724
36	1.306	1.688	1.802	2.131	2.028	2.434	2.719

Table 2: Critical value of Student's t distribution

df	Level of significance (α)						
	0.10	0.05	0.04	0.02	0.025	0.01	0.005
37	1.305	1.687	1.800	2.129	2.026	2.431	2.715
38	1.304	1.686	1.799	2.127	2.024	2.429	2.711
39	1.304	1.685	1.798	2.125	2.023	2.426	2.708
40	1.303	1.684	1.796	2.123	2.021	2.423	2.704
41	1.303	1.683	1.795	2.121	2.020	2.421	2.701
42	1.302	1.682	1.794	2.120	2.018	2.418	2.698
43	1.302	1.681	1.793	2.118	2.017	2.416	2.695
44	1.301	1.680	1.792	2.116	2.015	2.414	2.693
45	1.301	1.679	1.791	2.115	2.014	2.412	2.690
46	1.300	1.679	1.790	2.114	2.013	2.410	2.687
47	1.300	1.678	1.789	2.112	2.012	2.408	2.685
48	1.299	1.677	1.788	2.111	2.011	2.407	2.682
49	1.299	1.677	1.788	2.110	2.010	2.405	2.680
50	1.299	1.676	1.787	2.109	2.009	2.403	2.678
51	1.298	1.675	1.786	2.108	2.008	2.402	2.676
52	1.298	1.675	1.786	2.107	2.007	2.400	2.674
53	1.298	1.674	1.785	2.106	2.006	2.399	2.672
54	1.297	1.674	1.784	2.105	2.005	2.397	2.670
55	1.297	1.673	1.784	2.104	2.004	2.396	2.668
56	1.297	1.673	1.783	2.103	2.003	2.394	2.667
57	1.297	1.672	1.782	2.102	2.002	2.393	2.665
58	1.296	1.672	1.782	2.101	2.002	2.392	2.663
59	1.296	1.671	1.781	2.100	2.001	2.391	2.662
60	1.296	1.671	1.781	2.099	2.000	2.390	2.660
61	1.296	1.670	1.780	2.099	2.000	2.389	2.659
62	1.295	1.670	1.780	2.098	1.999	2.388	2.657
63	1.295	1.669	1.779	2.097	1.998	2.387	2.656
64	1.295	1.669	1.779	2.096	1.998	2.386	2.655
65	1.295	1.669	1.778	2.096	1.997	2.385	2.654
66	1.295	1.668	1.778	2.095	1.996	2.384	2.652
67	1.294	1.668	1.778	2.095	1.996	2.383	2.651
68	1.294	1.668	1.777	2.094	1.995	2.382	2.650
69	1.294	1.667	1.777	2.093	1.995	2.382	2.649
70	1.294	1.667	1.776	2.093	1.994	2.381	2.648
71	1.294	1.667	1.776	2.092	1.994	2.380	2.647
72	1.293	1.666	1.776	2.092	1.993	2.379	2.646

Table 2: Critical value of Student's t distribution

df	Level of significance (α)						
	0.10	0.05	0.04	0.02	0.025	0.01	0.005
73	1.293	1.666	1.775	2.091	1.993	2.379	2.645
74	1.293	1.666	1.775	2.091	1.993	2.378	2.644
75	1.293	1.665	1.775	2.090	1.992	2.377	2.643
76	1.293	1.665	1.774	2.090	1.992	2.376	2.642
77	1.293	1.665	1.774	2.089	1.991	2.376	2.641
78	1.292	1.665	1.774	2.089	1.991	2.375	2.640
79	1.292	1.664	1.773	2.088	1.990	2.375	2.640
80	1.292	1.664	1.773	2.088	1.990	2.374	2.639
85	1.292	1.663	1.772	2.086	1.988	2.371	2.635
90	1.291	1.662	1.771	2.084	1.987	2.368	2.632
95	1.291	1.661	1.770	2.082	1.985	2.366	2.629
100	1.290	1.660	1.769	2.081	1.984	2.364	2.626
105	1.290	1.659	1.768	2.080	1.983	2.362	2.623
110	1.289	1.659	1.767	2.078	1.982	2.361	2.621
115	1.289	1.658	1.766	2.077	1.981	2.359	2.619
120	1.289	1.658	1.766	2.076	1.980	2.358	2.617
130	1.288	1.657	1.764	2.075	1.978	2.355	2.614
140	1.288	1.656	1.763	2.073	1.977	2.353	2.611
150	1.287	1.655	1.763	2.072	1.976	2.351	2.609
160	1.287	1.654	1.762	2.071	1.975	2.350	2.607
170	1.287	1.654	1.761	2.070	1.974	2.348	2.605
180	1.286	1.653	1.761	2.069	1.973	2.347	2.603
190	1.286	1.653	1.760	2.068	1.973	2.346	2.602
200	1.286	1.653	1.760	2.067	1.972	2.345	2.601
220	1.285	1.652	1.759	2.066	1.971	2.343	2.599
240	1.285	1.651	1.758	2.065	1.970	2.342	2.596
260	1.285	1.651	1.758	2.064	1.969	2.341	2.595
280	1.285	1.650	1.757	2.063	1.968	2.340	2.594
300	1.284	1.650	1.757	2.063	1.968	2.339	2.592
340	1.284	1.649	1.756	2.062	1.967	2.337	2.590
400	1.284	1.649	1.755	2.060	1.966	2.336	2.588
450	1.283	1.648	1.755	2.060	1.965	2.335	2.587
500	1.283	1.648	1.754	2.059	1.965	2.334	2.586
600	1.283	1.647	1.754	2.058	1.964	2.333	2.584
700	1.283	1.647	1.753	2.058	1.963	2.332	2.583
1000	1.282	1.646	1.752	2.056	1.962	2.330	2.581

Table 3: Critical Value of Chi Square distribution

df	Level of significance (α)					
	0.01	0.02	0.05	0.95	0.98	0.99
1	6.635	5.412	3.842	0.004	0.001	0.0002
2	9.210	7.824	5.992	0.103	0.040	0.020
3	11.345	9.837	7.815	0.352	0.185	0.115
4	13.277	11.668	9.488	0.711	0.429	0.297
5	15.086	13.882	11.071	1.146	0.752	0.554
6	16.812	15.033	12.592	1.635	1.134	0.872
7	18.457	16.622	14.067	2.167	1.564	1.239
8	20.090	18.168	15.507	2.733	2.032	1.647
9	21.666	19.679	16.919	3.325	2.532	2.088
10	23.209	21.161	18.307	3.940	3.059	2.558
11	24.725	22.618	19.675	4.575	3.609	3.054
12	26.217	24.054	21.026	5.226	4.178	3.571
13	27.688	25.472	22.362	5.892	4.765	4.107
14	29.141	26.873	23.685	6.571	5.368	4.660
15	30.578	28.260	24.996	7.261	5.985	5.229
16	32000	29.633	26.296	7.962	6.614	5.812
17	33.408	30.995	27.587	8.672	7.255	6.408
18	34.805	32.346	28.869	9.391	7.906	7.015
19	36.191	33.687	30.144	10.117	8.567	7.633
20	37.566	35.020	31.410	10.851	9.237	8.260
21	38.932	36.343	32.671	11.591	9.915	8.897
22	40.289	37.660	33.924	12.338	10.600	9.543
23	41.638	38.968	35.173	13.091	11.293	10.196
24	42.980	40.270	36.415	13.848	11.992	10.856
25	44.314	41.566	37.653	14.611	12.697	11.524
26	45.642	42.856	38.885	15.379	13.409	12.198
27	46.963	44.140	40.113	16.151	14.125	12.879
28	48.278	45.419	41.337	16.928	14.848	13.565
29	49.588	46.693	42.557	17.708	15.575	14.257
30	50.892	47.962	43.773	18.493	16.306	14.954

Denominator degree of freedom v_2	Numerator degree of freedom v_1						
	1	2	3	4	5	6	7
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91
13	4.67	3.81	3.41	3.18	3.02	2.92	2.83
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66
17	4.45	3.59	3.20	2.97	2.81	2.70	2.61
18	4.41	3.56	3.16	2.93	2.77	2.66	2.58
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51
21	4.33	3.47	3.07	2.84	2.69	2.57	2.49
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42
25	4.24	3.39	2.99	2.76	2.60	2.49	2.41
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20
75	3.97	3.12	2.73	2.49	2.34	2.22	2.13
99	3.94	3.09	2.70	2.46	2.31	2.19	2.10
149	3.90	3.06	2.67	2.43	2.27	2.16	2.07

Numerator degree of freedom v_1

Denominator degree of freedom v_2

	8	9	10	11	12	13	14
1	238.88	240.54	241.88	242.98	243.91	244.69	245.36
2	19.37	19.38	19.40	19.40	19.41	19.42	19.42
3	8.85	8.81	8.79	8.76	8.74	8.73	8.71
4	6.04	6.00	5.96	5.94	5.91	5.89	5.87
5	4.82	4.77	4.74	4.70	4.68	4.65	4.64
6	4.15	4.10	4.06	4.03	4.00	3.98	3.96
7	3.73	3.68	3.64	3.60	3.57	3.55	3.53
8	3.44	3.39	3.35	3.31	3.28	3.26	3.24
9	3.23	3.18	3.14	3.10	3.07	3.05	3.03
10	3.07	3.02	2.98	2.94	2.91	2.89	2.86
11	2.95	2.90	2.85	2.82	2.79	2.76	2.74
12	2.85	2.80	2.75	2.72	2.69	2.66	2.64
13	2.77	2.71	2.67	2.63	2.60	2.58	2.55
14	2.70	2.65	2.60	2.56	2.53	2.51	2.48
15	2.64	2.59	2.54	2.51	2.48	2.45	2.42
16	2.59	2.54	2.49	2.46	2.42	2.40	2.37
17	2.55	2.49	2.45	2.41	2.38	2.35	2.33
18	2.51	2.46	2.41	2.37	2.34	2.31	2.29
19	2.48	2.42	2.38	2.34	2.31	2.28	2.25
20	2.45	2.39	2.35	2.31	2.28	2.25	2.22
21	2.42	2.37	2.32	2.28	2.25	2.22	2.20
22	2.40	2.34	2.30	2.26	2.23	2.20	2.17
23	2.37	2.32	2.27	2.24	2.20	2.18	2.15
24	2.36	2.30	2.25	2.22	2.18	2.15	2.13
25	2.34	2.28	2.24	2.20	2.16	2.14	2.11
26	2.32	2.27	2.22	2.18	2.15	2.12	2.09
27	2.31	2.25	2.20	2.17	2.13	2.10	2.08
28	2.29	2.24	2.19	2.15	2.12	2.09	2.06
29	2.28	2.22	2.18	2.14	2.10	2.08	2.05
30	2.27	2.21	2.16	2.13	2.09	2.06	2.04
35	2.22	2.16	2.11	2.07	2.04	2.01	1.99
40	2.18	2.12	2.08	2.04	2.00	1.97	1.95
50	2.13	2.07	2.03	1.99	1.95	1.92	1.89
75	2.06	2.01	1.96	1.92	1.88	1.85	1.83
99	2.03	1.98	1.93	1.89	1.85	1.82	1.79
149	2.00	1.94	1.89	1.85	1.82	1.79	1.76

Numerator degree of freedom v_1

Denominator degree of freedom v_2

	15	20	25	30	50	99	199
1	245.95	248.01	249.26	250.10	251.77	253.03	253.67
2	19.40	19.45	19.46	19.47	19.48	19.49	19.49
3	8.70	8.66	8.63	8.62	8.58	8.55	8.53
4	5.86	5.80	5.77	5.75	5.70	5.66	5.63
5	4.62	4.56	4.52	4.50	4.44	4.41	4.37
6	3.94	3.87	3.84	3.81	3.75	3.71	3.67
7	3.51	3.45	3.40	3.38	3.32	3.28	3.23
8	3.22	3.15	3.11	3.08	3.02	2.98	2.93
9	3.01	2.94	2.89	2.86	2.80	2.76	2.71
10	2.85	2.77	2.73	2.70	2.64	2.59	2.64
11	2.72	2.65	2.60	2.57	2.51	2.46	2.40
12	2.62	2.54	2.50	2.47	2.40	2.35	2.30
13	2.53	2.46	2.41	2.38	2.31	2.26	2.21
14	2.46	2.39	2.34	2.31	2.24	2.19	2.13
15	2.40	2.33	2.28	2.25	2.18	2.12	2.07
16	2.35	2.28	2.23	2.19	2.12	2.07	2.01
17	2.31	2.23	2.18	2.15	2.08	2.02	1.96
18	2.27	2.19	2.14	2.11	2.04	1.98	1.92
19	2.23	2.16	2.11	2.07	2.00	1.94	1.88
20	2.20	2.12	2.07	2.04	1.97	1.91	1.84
21	2.18	2.10	2.05	2.01	1.94	1.88	1.81
22	2.15	2.07	2.02	1.98	1.91	1.85	1.78
23	2.13	2.05	2.00	1.96	1.89	1.82	1.76
24	2.11	2.03	1.98	1.94	1.86	1.80	1.73
25	2.09	2.01	1.96	1.92	1.84	1.78	1.71
26	2.07	1.99	1.94	1.90	1.82	1.76	1.73
27	2.06	1.97	1.92	1.88	1.81	1.74	1.71
28	2.04	1.96	1.91	1.87	1.79	1.73	1.69
29	2.03	1.94	1.89	1.85	1.77	1.71	1.67
30	2.02	1.93	1.88	1.84	1.76	1.70	1.62
35	1.96	1.88	1.82	1.79	1.70	1.64	1.56
40	1.92	1.84	1.78	1.74	1.66	1.59	1.51
50	1.87	1.78	1.73	1.69	1.60	1.52	1.44
75	1.80	1.71	1.65	1.61	1.52	1.44	1.34
99	1.77	1.68	1.62	1.57	1.48	1.39	1.28
149	1.73	1.64	1.58	1.54	1.44	1.35	1.22