

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/209729264>

Textbook of Computer applications and biostatistics

Chapter · January 2011

CITATIONS

0

READS

79,805

4 authors, including:



[Remeth J Dias](#)

Government Polytechnic, Jalgaon

88 PUBLICATIONS 595 CITATIONS

[SEE PROFILE](#)



[Kailas K Mali](#)

Adarsh College of Pharmacy, Vita

78 PUBLICATIONS 530 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Tamarind Gum [View project](#)



Liquisolid Compacts [View project](#)

TEXTBOOK OF COMPUTER APPLICATIONS AND BIOSTATISTICS

Dr. S. B. Bhise

M. Pharm PhD

Principal,

Singhad Institute of Pharmaceutical Sciences, Lonavala
satishbhise@gmail.com

Dr. R. J. Dias

M. Pharm PhD MBA

Professor,

Singhad Institute of Pharmaceutical Sciences, Lonavala
rjdias75@gmail.com

K. K. Mali

M. Pharm (Biopharm)

Associate Professor,

Satara College of Pharmacy, Satara
malikailas@gmail.com

P. H. Ghanwat

DIE, MCA

Visiting Lecturer,

Satara College of Pharmacy, Satara
pravin_hg@rediffmail.com



TRINITY PUBLISHING HOUSE

Serving Pharmacy Profession

Textbook of Computer Applications and Biostatistics

Published by

Mrs. Anita R. Dias
For Trinity Publishing House,
475/8, F-3, Suryanandan Apartments,
Near Hotel Suruban, Sadar Bazaar,
Satara - 415 001. India.
Mobile: +91 9850953955, +91 8087250235
E-mail: trinity.satara@gmail.com

© 2011 Trinity Publishing House

All rights reserved. No part and style of this book be reproduced or transmitted, in any form, or by any means- electronic, mechanical, photocopying, recording or otherwise, without prior permission of the publishers and authors.

Disclaimer: As new information becomes available, changes become necessary. The editors/authors/contributors and the publishers have, as far as it is possible, taken care to ensure that the information given in this book is accurate and up-to-date. In view of the possibility of human error or advances in medical science neither the editor nor the publisher nor any other party who has been involved in the preparation or publication of this work warrants that the information contained herein is in every respect accurate or complete. Readers are strongly advised to confirm. This book is for sale in India only and cannot be exported without the permission of the publisher in writing. Any disputes and legal matters to be settled under Mumbai jurisdiction only.

ISBN 978-81-920565-1-7

Rs. 350/-

Printed at
Vikram Printers Pvt. Ltd.
31 & 34, Parvati Industrial Estate,
Pune-Satara Road,
Pune- 411 009. India.
Phone: (020)24220890, 24228905.
www.vikram-printers.com

Designed by
Srushti Computers, Satara.
G-2, 'Venna', Adarshnagar, Khed, Satara.
srushti.graphics1@gmail.com.

Distributed by
Amit Book Company Pvt. Ltd.
B-3/16, Darja Ganj,
Near The Time of India,
New Delhi - 110 002.
Phone: 011 - 43538989.
abdkaushal@in.com.



PREFACE

We are very pleased to put forth the first edition of book, 'Textbook of Computer Applications and Biostatistics'. This book is intended to be an introduction to pharmacy students regarding applications of computers and biostatistics to pharmacy. The basic knowledge of computers and their applications is covered in details as it is essential to students in every walk of their lives. The procedures for operating MS-Office 2003 is discussed here as many colleges still use this version. However our second edition will have the procedures for operating latest versions of Windows. We regret for inconvenience caused to few readers, due to this.

The concepts of biostatistics are discussed here with minimum of maths, so as to drive away the maths phobia in pharmacy students. Moreover, most of the statistics can be handled through computers using excel and we have emphasized in every chapter on how to use computers for statistical needs. This will help students to handle the data and infer about their experiments easily.

This book is an sincere effort to bring statistical concepts in simple, understandable form so that every student will enjoy to learn them with ease. The learning objectives, summary, multiple choice questions and exercise in all twenty two chapters makes the book more interesting.

We acknowledge the help and co-operation extended by various persons in bringing out this book. We are highly indebted to the authors of the various books and articles mentioned in bibliography which became a major source of information for writing this book. We also thank the publishers and designers who graciously worked hard to publish this book in time.

Our request to all users of this book is to provide constructive criticism in improving further editions of the book. We sincerely hope that readers will certainly welcome the book.

Satara

January 15, 2011.

**SB Bhise
RJ Dias
KK Mali
PH Ghanwat**

CONTENTS

Chapter No.	Title of the Chapter	Page No.
1.	Introduction to Computers Introduction, History of computers, Evolution and generations of computers, Characteristics of computer, Types of computers, Applications of computers.	1
2.	Anatomy and Computer Peripherals Anatomy of computer system, Parts of computer system, Hardware, Software, Input devices, Output devices, Memory, Binary numbers in computers, Unit of size, Computer language.	13
3.	Operating System Introduction, Functions, Types of operating systems, MS-DOS, MS-Windows, Working in Paint, Wallpaper, Screensaver.	34
4.	Microsoft Word Introduction, Features of word processing, Opening Microsoft Word, Components of the screen in MS-Word, Creating, editing and saving the document, Formatting the text, Printing a document, Quitting Microsoft Word.	60
5.	Microsoft Excel Introduction, Features of Microsoft Excel, Opening of Spreadsheet, Components of an Excel workbook, Entering data and saving a new workbook, Mathematical calculations, Moving and copying data, Deleting and adding rows and columns, Aligning data, Changing the size of row and column, Creating a graph, Adding, renaming or deleting a sheet from the workbook, Closing the workbook, Quitting Microsoft Excel.	75
6.	Microsoft PowerPoint Introduction, Features of PowerPoint presentation, Starting PowerPoint presentation, Components of MS-PowerPoint, Exploring PowerPoint views, Creating and saving a PowerPoint presentation, Adding slides, chart, picture, text box to a presentation, Duplicating and deleting slides, Adding animation to a presentation, Making slide show, Closing a PowerPoint presentation, Quitting a PowerPoint presentation.	95
7.	Computer Networking and Internet Applications Introduction, Types of networking, Topology, TCP/IP protocol, Advantages of Networking, Internet, Internet connection, Requirements for connecting to the Internet, Internet services, WWW, DNS, URL's, E-mail, Intranet, Net surfing, Chatting, Computer virus.	110
8.	Applications of Computers to Pharmacy Use of computers in Manufacturing of drugs, Quality control, Quality assurance, Pharmaceutical analysis, Inventory control, Clinical research, Retail pharmacy, Drug information services, Marketing and sales, Hospital pharmacy, Clinical services, Bioequivalence testing, Basic research, etc.	141
9.	Introduction to Biostatistics Definition, Types of statistics, Applications and uses of Biostatistics, Types of variables, Identification of the type of variable.	148

CONTENTS

Chapter No.	Title of the Chapter	Page No.
10.	Presentation of Data Tabulation of data, Graphical presentation of categorical and metric data, Charting of data using MS-Excel.	157
11.	Shape of Distribution of Data Shapes of data, Bell shaped distribution, Skewness, Kurtosis.	187
12.	Measures of Central Tendency Measures of location: Mode, Median, Mean, Measures of spread: Percentile, Range, Interquartile range, Standard deviation, Use of Excel in measures of central tendency.	193
13.	Probability and Probability Distribution Classic probability, Probability of simple and composite event, Probability involving two variables and conditional probability, Probability distribution.	231
14.	Sampling Techniques Methods of Sampling, Precision, Accuracy and Bias.	244
15.	Estimation of Confidence Interval Concept of confidence interval, Standard error of the mean, Estimation of confidence interval.	251
16.	Hypothesis Testing Hypothesis testing, Decision rule, Types of errors, One-tailed and two-tailed test, Defining the critical region for statistical test.	259
17.	Choice of Statistical Tests Parametric and Non-parametric tests, Choice of statistical tests, Commonly used hypothesis tests.	266
18.	Hypothesis Testing for One Sample One sample z-test, One sample t-test.	271
19.	Hypothesis Testing for Two Samples Two samples z-test, Two samples t-test, Paired t-test, Use of MS-Excel in hypothesis testing for two samples	282
20.	One Way Analysis of Variance (ANOVA) Definition, Calculations of ANOVA by using definitional and computational formula, One way ANOVA using MS-Excel.	305
21.	Correlation Definition, Types of correlation, Calculation of correlation coefficient by definitional and computational formula, Correlation using MS-Excel.	327
22.	Linear Regression Definition, Meaning of regression, Regression coefficient, Linear regression using MS-Excel.	340
	Important Points & Formulaes At A Glance	351
	Appendices	354
	Bibliography	358

Chapter 8

APPLICATIONS OF COMPUTERS TO PHARMACY

Learning objectives

When we have finished this chapter, we should be able to:

1. Understand applications of computers to pharmacy.
2. Know various computer programmes used in different areas of pharmacy field.

Introduction

The utility of computer in collection, evaluation, organisation and dissemination of information has made their presence virtually in every walk of life. Their potential in every field of pharmacy has led to its extensive use encompassing research of drug, its manufacturing and till its proper usage. The following properties of computers have made them to bring biggest revolution of the twenty first century known as information technology. The properties are:

1. Large storage capacity
2. Speed and accuracy
3. Flexibility
4. Ease of dissemination and transmission
5. Multiple user capacity
6. Can do repetitive tasks.

Let us see the applications of computers to various fields of pharmacy as given below:

1. Use of computers for manufacturing of drugs

The manufacturing of drugs in various dosage forms requires sophisticated instruments and machinery which are now-a-day controlled by computers. The touch screen panels provided to these machines can be utilised for controlling various manufacturing variables thereby producing quality medicines. Automation brought in manufacturing area has increased the efficiency, quality and safety manifold.

Computer- aided manufacturing (CAM) is the use of computers to plan and control manufacturing process. A well designed CAM system allows manufacturers to become much more productive. Not only a greater number of products are produced, but also speed and quality is increased.

Softwares available: Marg pharmaceutical software for manufacturers, DMC Medical Manufacturing, Taylor Pharmaceutical Manufacturing, TGI Process manufacturing, MISys Manufacturing software, etc.

2. Use of computers in quality control

The quality of medicine is utmost important for any manufacturing plant and every batch has to pass stringent quality control norms. Computers play very important role in quality control department by analyzing and interpreting whether raw materials and finished products match the expected quality norms, as per the specifications given in official book.

Softwares available: Darwin LIMS software for QA/QC, DMC quality control software, Monark software for Pharma QC system, MasterControl QC software etc.

3. Use of computers in quality assurance

The computers are used in documentation of every process for assuring quality medicine. The preparation of standard operating procedures, validating their use, and ensuring their adherence according to guidelines are few areas where computers are necessary.

Softwares available: Qtor Pharmaceutical QA software, EtQ software, Marg QA, DMC QA software, etc

4. Use of computers in pharmaceutical analysis

Various analytical methods like high performance liquid chromatography (HPLC), Gas chromatography (GC), Gas chromatography with mass spectrometry (GC-MS), Infrared spectroscopy (IR), UV visible spectroscopy (UV) etc are used for analysing various drugs. Computers are useful in all these techniques for identifying and interpreting the compounds. For example, Gas liquid chromatography separates the individual components and MS identifies it. This operation generates hundreds of spectra in few minutes containing number of peaks. Computer when used for interpretation of this data, stores these spectra for some time and then represents it in the graphical form. For identification of mass spectrometry computer compares the spectrum of the given sample with the spectrum of pure compound.

Softwares available: DeWinter, Drug Testing Software Management Suit, DrugPak IPA Core Analysis, Assistant Pro 4.1, etc.

5. Use of computers in inventory control

The inventory control includes purchasing of raw materials as per the demand, supply of finished goods, distribution of raw materials to various departments, and maintaining the stocks effectively. Softwares can be utilized to provide reports on stock status, opening and closing balance and to prepare purchase orders.

Softwares available: mSupply, IMS Leon, Meditab IMS, CASI, DMC, etc.

6. Use of computers in clinical research

Computer is utilized for coding of data and statistical analysis of results of clinical trials. Computer process ensures uniformity, completeness and accuracy of data collected from various centres and helps in evaluating efficacy of design protocols on investigational drugs. Computers has been extensively used in multicentric clinical trials on investigational drugs.

Softwares available: OpenClinica, TEMPO, Cytel, InferMed, Clinplus, TrackWise, Metadata ClinAxy II, etc.

7. Use of computers in managing drug store (Retail pharmacy)

In retail pharmacy store, inventory control is done using computers. Computers also helps in sales analysis and purchasing of the medicines. Computers can be used for storage of patient records and the drug history profiles. The software are also available which shows drug interactions in prescribed medication and patient instructions to be given. This allows pharmacists to do better patient counseling job.

Software available: Dava plus, PharmaSoft, ApotheSoft Rx, PEPID, Essel, Abascus Pharmacy, MediVision, etc.

8. Use of computers in drug information services

The vast data is available on drugs and disease state and computerization of this data is needed to retrieve the information needed. The user oriented drug information systems are available whereby drug information guide for patients and full text database of drug is given as ready reckoner. Drug information centres using computers can answer drug related queries like how, why, what and when of medication immediately.

Softwares available: MicroMedeX

Websites: RxFacts .org, medind.nic.in, www.uiowa.edu, www.health.umd.edu, Lexicomp online, Drugs.com, Davi's Drug Guide online, etc.

9. Use of computers in marketing and sales

The use of computers in every arena of marketing and sales department have made the field more promising and lucrative. In this field computers are used in market research, consumer survey, advertizing, sales promotion campaigns, product management, developing dealer distribution network, sales analysis etc. E marketing is newer concept in which medicines are purchased or sold online eg. Dava Bazaar.

Softwares available: Medisno Pharma CRM, Metastorm, Marg online MR software, Marg ethical marketing software, etc.

10. Use of computers in hospital pharmacy

Use of computer in managing bigger hospitals have saved a lot of money and man hours. A complete hospital information system is a multiterminal database management system that provides facilities for integrated handling of information regarding registration of patients; investigation, treatment, follow up etc.

Some of the other areas covered under this system are inventory control and purchase of medicine, drug information database, adverse effects and drug interaction management, drug distribution within hospital, prescription monitoring and preparation of hospital formulary.

Softwares available: Meditab IMS, WinPharm, WorkPath, MediNovs, HMS-Leon, etc.

11. Use of computers in clinical services

Computer programmes have been designed to assist in the monitoring of patients with chronic diseases like diabetes, hypertension and asthma. Therapeutic drug monitoring based upon plasma concentration of a drug is possible using computer programmes.

Softwares available: NAMAHA, Pharmacy Plus, VirtualCare, MEDIPHOR, etc.

12. Use of computers in bioequivalence testing centres

Bioequivalence of generic drugs with reference innovator's drug is very important regulatory need for marketing off patented drugs. Various bioequivalence centres use computer programmes to accumulate, sort and use the data for bioequivalence testing.

Softwares available: WinNonlin, DDSolver, EquivTest, MONOLIX, Study Size 2.0, etc.

13. Use of computers in basic research

The softwares available today are playing important role in drug design. Computer aided drug design (CADD), Qualitative Structure Activity Relationship (QSAR), Molecular modeling are promising areas that help the researchers in doing their research with ease.

Available softwares: AMBER, CHARMM and GROMACS are widely used to carry out molecular modeling. FTDock and DARWIN has tremendous application in the rational drug design process.

Other softwares available: Prochemist, Tripos, Oxford Molecular QSAR, HYPERCHEM, MATLAB, DRAGON, RECKON, Spartan, etc.

14. Use of computers in Medicine

Many of the modern-day medical equipment have small, programmed computers. Many of the medical appliances of today work on pre-programmed instructions. The circuitry and logic in most of the medical equipment is basically a computer.

Computer software is used for diagnosis of diseases. It can be used for the examination of internal organs of the body. Advanced computer-based systems are used to examine delicate organs of the body. Some of the complex surgeries can be performed with the aid of computers. The different types of monitoring equipment in hospitals are often based on computer programming.

Medical imaging is a vast field that deals with the techniques to create images of the human body for medical purposes. Many of the modern methods of scanning and imaging are largely based on the computer technology. It has been possible to implement many of the advanced medical imaging techniques, due to the developments in computer science. Magnetic resonance imaging employs computer software. Computed tomography makes use of digital geometry processing techniques to obtain 3-D images. Sophisticated computers and infrared cameras are used for obtaining high-resolution images. Computers are widely used for the generation of 3-D images in medicine.

Softwares available: CliniScript, Clinichem, LIFEDATA EMR, LDRA, EasyDiagnosis, Diagnosis Pro, COMPUTER CLINIC, DICOM, 3D-DOCTOR, MIM, etc.

15. Use of computers in Biostatistics

Every data generated during the research is to be handled or manipulated carefully for drawing inference from it. Both the types, descriptive and inferential statistics can be applied by researcher using MS Excel or relevant softwares.

Softwares available: SigmaStat, GraphPad Prism, Minitab, SPSS, SAS, DesignExpert, SigmaXL, etc.

16. Use of computers in Patent searching

Now a days computers are widely used for searching patents available online on various countries official websites. The important sites available for the same are given in table below:

Important websites for patent search

Sr. No.	Title	Official websites
1.	US Patent and Trademark Office	www.uspto.gov
2.	UK Patent Office (IPO)	www.ipo.gov.uk
3.	European Patent Office	www.epo.org
4.	Japanese Patent Office	www.jpo.go.jp
5.	World Intellectual Property Organisation	www.wipo.int
6.	Indian Patent Office	www.ipindia.nic.in
7.	Australian Patent Office	www.ipaustralia.gov
8.	Singaporean Patent Office	www.ipos.gov.sg
9.	Chinese Patent Office	www.chinatrado.com
10.	Google Patent Search	www.google.com/patents
11.	free online Patent search	www.freepatentsonline.com
12.	Espace Patent database	http://wordwide.espacenet.com

17. Use of computers in pharmacology simulations

Now a days computers are widely used as alternative for animal experimentation. Various simulated pharmacology experiments are generated with the help of computers and are widely used for the learning of undergraduate students.

Softwares available: Biosoft, Neurosim, LabTutor, X-Cology, Cardiolab, MacDogLab, Ex-pharm, PCCAL, etc.

18. Use of computers in pharmacokinetics simulations

It is a simulation method used in determining the safety levels of a drug during its development. It gives an insight to drug efficacy and safety before exposure of individuals to the new drug that might help to improve the design of a clinical trial.

Simcyp Simulator and GastroPlus (from Simulations Plus) are simulators that take account for individual variabilities. GastroPlus is an advanced software program that simulates the absorption, pharmacokinetics, and pharmacodynamics for drugs in human and preclinical species.

Softwares available: NONLIN, KINPAK, ESTRIP, STRIPACT, PK solutions, KINETICS, SimBiology, etc.

19. High performance computing bioinformatics

The advent of high throughput technologies like genome sequencing, microarrays and proteomics has transformed biology into a data rich information sciences. The huge data generated needs to be organized in a structured manner to facilitate the use of data mining tools for extracting knowledge. The ultimate objective of these efforts is to improve our understanding of human health and thereby provide rationale solutions to overcome disease. The use of computers in this area has made performance computing possible.

Softwares available: Amber, Alchemy, Sybyl, MOE, Cerius, Bioconductor, ISYS v.1.35,

MetaCore, etc.

20. Literature storage and retrieval system

Computers have been utilized to offer bibliographic, indexing and abstracting services. the articles can be referred through keywords, titles, authors or journals. Automated on-line literature retrieval systems like Medline and Chemline are offered by National Library of Medicine, USA. International Pharmaceutical Abstracts (IPA) and Martindale's Extra Pharmacopoeia are also available on compact disks.

Databases available: Excerpta Medica, LIMS, AMA/NET Information base, etc.

Summary

Use of computer in various fields of Pharmacy is given in following table:

Sr. No.	Use of Computers in Pharmacy	Softwares Available
1.	Manufacturing of drugs	Marg, DMC, Taylor, TGI, MIsys
2.	Quality control	Darwin LIMS, DMC-QC, Maonark, MasterControl
3.	Quality assurance	Qtor, EtQ, Darwin LIMS, Marg, DMC-QA
4.	Pharmaceutical analysis	DeWinter, Drugpak, IPA Core, Assistant Pro, DTSMS
5.	Inventory control	mSupply, IMS Leon, Meditab IMS, CASI, DMC
6.	Clinical research	OpenClinica, Clinplus, Cytel, Metadata, TrackWise
7.	Retail drug store and wholesalers	PharmaSoft, Apothesoftware, PEPID, Medivision, Abacus
8.	Drug information services	MicroMedex, Lexicomp Platinum, Davi's Drug Guide Tarascon's Pharmacopoeia, DIT Drug Risk Navigator
9.	Marketing and sales	Marg Ethical Marketing, Pharma CRM, Metastorm
10.	Hospital management and pharmacy	WinPharm, WorkPath, MediNous, HMS-Leon
11.	Clinical services	PharmacyPlus, VirtualCare, TEICTDM, NAMA MEDIPHOR, MW/Pharm
12.	Bioequivalence testing centres	WinNonlin, DDSolver, EquivTest, MONOLIX
13.	Research and development	Prochemist, Tripos, AMBER, DRAGON, RECKON Spartan, CHARMM, GROMACS, MATLAB
14.	Biostatistics	SAS, SPSS, Minitab, SigmaStat
15.	Medical Diagnosis and Imaging	DICOM, 3D-DOCTOR, Easy Diagnosis, MIM, LDRA
16.	Patent search	www.uspto.gov, www.wipo.int, www.epo.org, www.ipo.gov.uk, www.ipindia.nic.in, etc
17.	Pharmacology simulations	LabTutor, X-Cology, Ex-Pharm, Biosoft, Neurosim
18.	Pharmacokinetic simulation	NONLIN, KINETICS, KINPAK, ESTRIP
19.	Bioinformatics	MetaCore, BioSpice, ISYS, 3Dslicer, Bioconductor
20.	Literature survey	Pubmed, Medline, Google, IPA, LIMS

Multiple Choice Questions:

1. _____ is used to plan and control manufacturing process.
a. CAM b. CADD c. QSAR d. NONLIN
2. _____ includes purchasing, supply, distribution and maintenance of stocks using computers.
a. Clinical research b. Inventory control
c. Quality control d. Pharmaceutical analysis
3. The following software is used in the field of Clinical Research.
a. PharmaSoft b. MicroMedex c. Prochemist d. OpenClinica
4. Automated on-line literature retrieval system offered by National Library of Medicine, USA is _____.
a. MEDIPHOR b. Medline c. IPA d. Minitab
5. Bioequivalence testing centre uses following software.
a. VirtualCare b. X-Cology c. KINETICS d. Equivtest
6. The following software is used for statistical analysis.
a. LabTutor b. MONOLIX c. SAS d. DRAGON
7. Which of the following website helps online patent search.
a. www.medind.nic.in b. www.health.umd.edu
c. www.wipo.int d. www.Rxfacts.org
8. _____ is programme that simulates pharmacokinetics of drug.
a. Expharm b. GastroPlus c. DavaPlus d. ClinPlus
9. The example of e-marketing, in which medicines are purchased or sold online is _____.
a. DavaPlus b. GastroPlus c. ClinPlus d. Dava Bazaar
10. The laboratory pharmacology simulations are given by following software.
a. LabTutor b. PharmacyPlus c. Workpath d. DD Solver

Exercise:

1. Give an account on application of computers to pharmacy.
2. Discuss the use of computers in basic research.
3. Give various softwares available for pharmaceutical manufacturing.
4. How literature storage and retrieval is possible with computers?
5. Give the use of computers in pharmacokinetics.

Answers:**Multiple Choice Questions**

1. a 2. b 3. d 4. b 5. d 6. c 7. c 8. b 9. d 10. a



Chapter 9

INTRODUCTION TO BIOSTATISTICS

Learning objectives

When we have finished this chapter, we should be able to:

1. Explain types of statistics with their components.
2. Explain the difference between nominal, ordinal, metric discrete and metric continuous variables.
3. Identify the type of a variable.

What is Biostatistics?

As defined by Daniel in 1978, “Biostatistics is a field of study concerned with the organisation and summarisation of data from health sciences and drawing of inferences about a body of data when only a part of the data are observed”.

In simpler terms, biostatistics can be defined as the branch of statistics applied to biological sciences whereby collection, classification, summarizing, analysis and interpretation of data is done.

Types of Statistics

All statistical procedures can be divided into general categories- descriptive or inferential.

Descriptive statistics

As the name implies, descriptive statistics describes data that we collect or observe (empirical data). They represent all of the procedures that can be used to organise, summarise, display, and categorise data collected for a certain experiment or event. It includes tabulation, graphical presentation, measures of central tendency, etc.

Inferential statistics

Inferential statistics represents a wide range of procedures (tests) that are used to infer or make predictions about a large body of information based on sample observations. The inferential statistics include z test, t test, analysis of variance, etc.

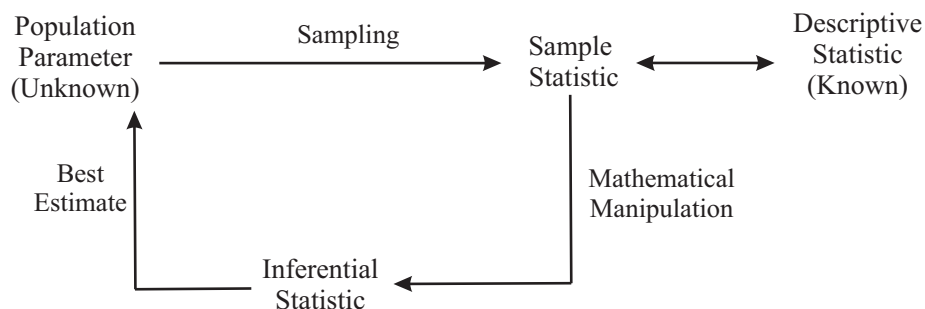
Statistical samples and population

Statistical data usually involve a relatively small portion of an entire population, and decisions and interpretations (inferences) are made about that population through numerical manipulation. The population may be defined as the entire number of observations that constitute a particular group. Samples are generally a relatively small group of observations that have been taken from a defined population. Parameters are characteristics of populations while statistic are characteristics of samples representing summary measures computed on observed sample values. Parameters or population values are usually represented by Greek symbols (e.g. μ , σ , ψ) and sample statistic are denoted by letters (e.g. \bar{X} , S^2 , r).

Table 9.1 Examples of populations and samples

Examples (task)	Population parameter	Sample statistic
Characterisation of the weights of tablets in a particular batch	All tablets that constitute the batch	100 tablets removed for weighing
Measurement of the incidence of heart disease in Maharashtra in patients over 45 years of age	All inhabitants of Maharashtra over the age of 45	300 patients attending GP clinics at specified geographical locations throughout Maharashtra
Evaluation of the incidence of asthma in a certain chemical company employing 500 workers	All employees of the company	50 named workers at the company

Samples, as given in above examples, are only a small subset of a much larger population and are used for nearly all statistical tests. By using various formulas, these descriptive sample results are manipulated to make predictions (inferences) about the population from which they are sampled.

**Figure 9.1** Descriptive statistics and inferential statistics

Applications and Uses of Biostatistics

1. Problem Solving

Often the research is conducted on a limited scale due to scarcity of resources. Biostatistics helps us in interpreting the data of whole population by taking a sample from that population.

2. Use in Clinical Trials

Biostatistics can be used in designing and analysing study whereby the relative potency of a new drug with respect to a standard drug can be found out.

3. Use in Manufacturing of Pharmaceuticals

The changes in manufacturing variables, machines and manpower to improve the efficiency of manufacturing of pharmaceuticals can be tested and confirmed by using statistical methods.

4. Use in Quality Control of Pharmaceuticals

The quality of pharmaceuticals can be controlled by performing various quality control tests on limited samples and interpreting the results for whole batches manufactured.

5. Use in Bioequivalence study

The bioavailability of test drug can be compared to innovators drug and the decision of its bioequivalence is based on passing the appropriate statistical test for the purpose.

6. Use in Research and Development of Drugs and Technology

Biostatistics can be applied to test the significance of any experiment in research and development of drugs and technology by comparing the obtained results with the result given by standard drug and technology.

7. Use in Anatomy and Physiology

Biostatistics is used in anatomy and physiology

- i. to define what is normal or healthy in a population and to find limits of normality in variables. e.g. weight and pulse rate.
- ii. to find the difference between means and proportions of normal at two places or in different periods. The mean height of boy in Maharashtra is less than the mean height in Punjab. Whether this difference is due to chance or a natural variation or because of some other factors such as better nutrition playing a part, has to be decided.

8. Use in Pharmacology

Biostatistics is used in pharmacology

- i. to find the action of drug by giving it to animals or humans to see whether the changes produced are due to the drug or by chance.
- ii. to compare the action of two different drugs or two successive dosages of the same drug.

9. Use in Medicine

Biostatistics is used in medicine

- i. to compare the efficiency of a particular drug, operation or line of treatment by comparing it with control groups.
- ii. to find an association between two attributes such as cancer and smoking.
- iii. to identify signs and symptoms of a disease or syndrome. Cough in typhoid is found by chance and fever is found in almost every case. The proportional incidence of one symptom or another indicates whether it is a characteristic feature of the disease or not.

10. Use in Community Medicine and Public Health

Biostatistics is used in medicine and public health

- i. to test usefulness of sera and vaccines in the field: Here the percentage of attacks or deaths among the vaccinated subjects is compared with that among the unvaccinated ones to find whether the difference observed is statistically significant.
- ii. In epidemiological studies: The role of causative factors can be statistically tested. Deficiency of iodine as an important cause of goitre in a community is confirmed only after comparing the incidence of goitre cases before and after giving iodised salt.

11. Use in Health and Vital statistics

Health and Vital statistics are essential tools in demography, public health, medical practice

and community services. Biostatistics as a science of figures can tell

- a. the leading causes of death,
- b. the important causes of sickness, severity of disease, its prevalence, etc,

12. Use in Biotechnology, Bioinformatics and Computational Biotechnology

It can be used in analysis of genomics data, for example from microarray or proteomics experiment, often concerning diseases or disease stages. Statistical methods are beginning to be integrated into Medical informatics, public health informatics, bioinformatics and computational biology.

Types of Variables

A variable is something whose value can vary. For example age, sex, and blood type are variables. Data are the values we get when we measure or observe a variable. There are two major types of variables- categorical variables and metric variables. Each of these can be further divided into two sub-types as shown in table 9.2.

Table 9.2 Types of Variables

Type of variables	Sub type	Characteristics	Unit
Categorical variables	Nominal	Values in arbitrary categories	No units
	Ordinal	Values in ordered categories	No units
Metric variables	Discrete	Integer values on numeric scale	Counted units
	Continuous	Continuous values on numeric scale	Measured units

Categorical variables

1. Nominal categorical variables

Consider the variable blood type, O, A, B and A/B. The variable 'blood type' is a nominal categorical variable. A typical characteristics of this variable are that they do not have any units of measurement, and the ordering of the categories is completely arbitrary. In other words, the categories cannot be ordered in any meaningful way. Therefore, we can easily write the blood type categories as A/B, B, O, A or B, O, A, A/B or B, A, A/B, O, or whatever.

2. Ordinal categorical variables

The Child Pugh Score is an ordinal categorical variable. This data too do not have any units of measurement as like that of nominal variables but the ordering of the categories is not arbitrary as it was with nominal variables. It is now possible to order the categories in a meaningful way.

Ordinal data are not real numbers. They cannot be placed on the number line. The reason is that the Child Pugh Score data, and the data of most other clinical scales, are not properly measured but assessed in some way, by the clinician working with the patient. This is a characteristic of all ordinal data.

As ordinal data are not real numbers, it is not appropriate to apply any of the rules of basic arithmetic to sort this data. We can not add, subtract, multiply or divide ordinal values. This limitation has marked implications for the analyses of such data.

Metric Variables

1. Continuous metric variables

The variable 'weight' is a metric continuous variable. With metric variables, proper measurement is possible and therefore these variables produce data that are real numbers, and can be placed on the number line. Some common examples of metric continuous variables include: Birth weight (g), blood pressure (mmHg), blood cholesterol ($\mu\text{g/ml}$), waiting time (minutes), body mass index (kg/m^2), peak expiry flow (l per min), and so on. These variables have units of measurement attached to them.

In contrast to ordinal values, the difference between any pair of adjacent values of continuous metric variables is exactly the same. The difference between birth weights of 3000 g and 3001 g is the same as the difference between 3001 g and 3002 g, and so on.

Metric continuous variables can be properly measured and have units of measurement.

2. Discrete metric variables

Continuous metric data usually comes from measuring while discrete metric data, usually comes from counting. For example, number of deaths, number of pressure sores, number of angina attacks, and so on, are all discrete metric variables. The data produced are real numbers, and are invariably integer (i.e. whole number). They can be placed on the number line, and have the same interval and ratio properties as continuous metric data. Metric discrete variables can be properly counted and have units of measurement- 'numbers of things'.

An aid to identify type of variable

The easiest way to identify whether data is metric or categorical, is to check whether it has units attached to it, such as: g, mm, $^{\circ}\text{C}$, $\mu\text{g/cm}^3$, number of pressure sores, number of deaths, and so on. If not, it may be ordinal or nominal and if the values can be put in any meaningful order then it is ordinal. Figure 9.2 is an aid to variable type recognition.

Importance of Identifying variable

In order to select the correct inferential test procedure, it is essential that as researchers, we should understand the variables involved with our data. The type and subtype of variable is very much important for selecting appropriate statistical test. The details of selection of statistical tests are given in *chapter 17*.

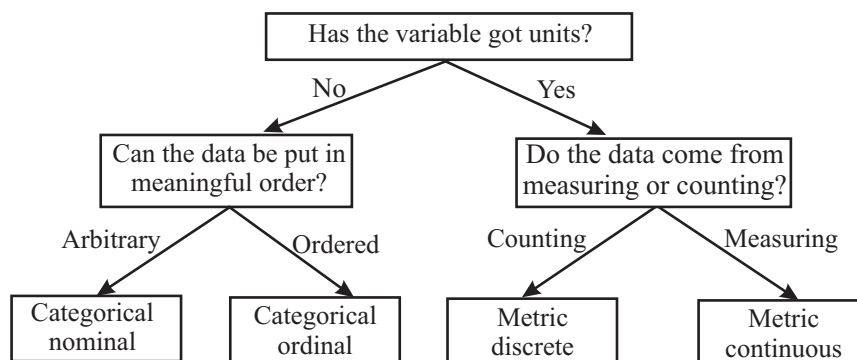


Figure 9.2 An algorithm to help identify variables type

Problem: Identify the type of the following variables:

1. Dosage form - tablet/ capsule / ointment
2. Bioavailability measurements (C_{max} , T_{max} , AUC)
3. Age (in years).
4. Hypertension -Mild, Moderate and Severe
5. Smoking history (cigarattes per day)
6. Hardness
7. Dissolution test- pass or fail criteria
8. Tablet weight
9. Male vs female subjects
10. Scores for patient responses to treatment

Solution:

Sr. No.	Variable	Type of Variable	Characteristics
1.	Dosage form - tablet/ capsule / ointment	Categorical, Nominal	No unit, No order
2.	Bioavailability measurements (C_{max} , T_{max} , AUC)	Metric, continuous	Unit present, can be measured
3.	Age (in years)	Metric, continuous	Unit present, can be measured
4.	Hypertension -Mild, Moderate and Severe	Categorical, Ordinal	No unit, ordered form
5.	Smoking history (cigarattes per day)	Metric, discrete	Unit- numbers of things counting can be done
6.	Hardness	Metric, continuous	Unit present, can be measured
7.	Dissolution test- pass or fail criteria	Categorical, Nominal	No unit, arbitrary
8.	Tablet weight	Metric, continuous	Unit present, can be measured
9.	Male vs female subjects	Categorical, Nominal	No unit, arbitrary
10.	Scores for patient responses to treatment	Categorical, Ordinal	No unit, ordered form

Summary**Biostatistics**

It can be defined as the branch of statistics applied to biological sciences whereby collection, classification, summarising, analysis and interpretation of data is done.

Types of Statistics

- 1. Descriptive Statistics:** It describes collected data.
- 2. Inferential Statistics:** It infers about a large body of information based on sample.

Types of Variables

- 1. Categorical Nominal :** No any units of measurement and values in arbitrary categories.
- 2. Categorical Ordinal:** No any units of measurement and values in ordered categories.
- 3. Metric Continuous:** It can be properly measured and have units of measurement.
- 4. Metric Discrete:** It can be properly counted and have units of measurement.

Multiple choice questions

1. Biostatistics is branch of statistics applied to _____ science whereby collection, classification, summarising, analysis and interpretation of data is done.
 - a. pharmaceutical
 - b. medicinal
 - c. biological
 - d. chemical
2. Descriptive statistics represents all of the procedures that can be used to _____.
 - a. organise, summarise
 - b. display and categorise
 - c. a & b
 - d. none of above
3. The population is _____.
 - a. the entire number of observations that constitutes a particular group.
 - b. the entire number of samples that constitutes a particular group.
 - c. a & b
 - d. None of above
4. Categorical variable can be divided into _____.
 - a. nominal & continuous
 - b. nominal & discrete
 - c. discrete & continuous
 - d. nominal & ordinal
5. Metric data can be divided into
 - a. nominal & continuous
 - b. nominal & discrete
 - c. discrete & continuous
 - d. nominal & ordinal
6. The goal of _____ is to focus on summarizing and explaining a specific set of data.
 - a. inferential statistics
 - b. descriptive statistics
 - c. none of the above
 - d. all of the above
7. Metric variable can be properly _____.
 - a. measured
 - b. counted
 - c. a & b
 - d. none of the above

8. In categorical variables, values are in _____ categories.
- a. arbitrary & counted
 - b. arbitrary & measured
 - c. arbitrary & ordered
 - d. counted & measured
9. Bioavailability measurement is a _____ variable.
- a. metric continuous
 - b. metric discrete
 - c. categorical continuous
 - d. categorical ordinal
10. Scores for patient responses to treatment is a _____ variable.
- a. metric continuous
 - b. metric discrete
 - c. categorical continuous
 - d. categorical ordinal

Exercise

1. Define biostatistics and enumerate applications of it.
2. Give various types of variables with their characteristics.
3. Give the importance of identifying type of variable in biostatistics.
4. Identify the type of variables associated with clinical trials of a drug given below,
 - a. Sex
 - b. Age
 - c. Height
 - d. Weight
 - e. Blood type (A, B, AB, O)
 - f. Blood pressure (Mild, Moderate, Severe)
 - g. Blood glucose level
 - h. Fed vs fasted state
 - i. Manufacturer (generic vs brand)
 - j. Smoking history (no of cigarettes per day)
5. Identify the types of variables given below, associated with manufacturing a batch of tablets of Ciprofloxacin.
 - a. Impurities- present or absent
 - b. Amount of active ingredient (content uniformity)
 - c. Disintegration time
 - d. Dissolution test- pass or fail criteria
 - e. Friability- pass or fail criteria
 - f. Hardness
 - g. Appearance (good, better, best)
 - h. Machine efficiency score (-5 to +5)
 - i. Weight variation test (pass or fail)
 - j. Human resources employed (No of persons)

6. Give various types of biostatistics.
7. Distinguish between parameter and statistic.
8. Distinguish between categorical and metric variables.
9. How will you identify the type of given variable?
10. What do you mean by sample and population? Explain.

Answers:**Multiple Choice Questions**

1. c 2. c 3. a 4. d 5. c 6. b 7. c 8. c 9. a 10. d

Exercise

4.
 - a. Sex- categorical nominal
 - b. Age- metric continuous
 - c. Height- metric continuous
 - d. Weight- metric continuous
 - e. Blood type (A, B, AB, O)- categorical nominal
 - f. Blood pressure (Mild, Moderate, Severe)- categorical ordinal
 - g. Blood glucose level- metric continuous
 - h. Fed vs fasted state- categorical nominal
 - i. Manufacturer (generic vs brand)- categorical nominal
 - j. Smoking history (no of cigarettes per day)- metric discrete

5.
 - a. Impurities- present or absent- categorical nominal
 - b. Amount of active ingredient (content uniformity)- metric continuous
 - c. Disintegration time- metric continuous
 - d. Dissolution test- pass or fail criteria- categorical nominal
 - e. Friability- pass or fail criteria- categorical nominal
 - f. Hardness- metric continuous
 - g. Appearance (good, better, best)- categorical ordinal
 - h. Machine efficiency score (-5 to +5)- categorical ordinal
 - i. Weight variation test (pass or fail)- categorical nominal
 - j. Human resources employed (No of persons)- metric discrete



Chapter 10

PRESENTATION OF DATA

Learning objectives

When we have finished this chapter, we should be able to:

1. Construct the tables of frequency, relative frequency, cumulative frequency and relative cumulative frequency.
2. Construct grouped frequency table and a cross-tabulation table.
3. Choose the most appropriate graph for the given data type.
4. Draw pie charts, bar charts, histograms, frequency polygons and ogives.
5. Interpret and explain what a table or graph reveals.

Introduction

Whenever the data is collected for some project, it is usually in the 'raw' form and not in a organised way. Descriptive statistics deals with sorting this raw data by putting it into a table or by presenting it in an appropriate chart or summarising it numerically.

An important consideration in sorting the raw data is the type of variable concerned. The data from some variables are best described with a table, some with a chart, and some with both. However, a numeric summary is more appropriate for some types of variable.

Tabulation of Data

Tabulation is the first step before the data is used for analysis or interpretation. Frequency distribution tables presents data in a relatively compact form, ready to use but certain information may be lost. The data can be reduced to manageable form using frequency tables.

The frequency table

The frequency table can have one or all the following parameters, depending on the type of data.

1. Frequency:

Frequency is the repetition of observations or actual number of subjects in each category.

2. Relative frequency:

Relative frequency is the frequency converted into percentage of the total number of observations.

$$\text{Relative frequency} = \frac{\text{Number of observations in category}}{\text{Total number of observations}} \times 100 \quad \dots 1$$

3. Cumulative frequency:

It is the cumulative total of frequencies and is obtained by adding the frequency of

observations at each level point to those frequencies of the preceding level (s).

4. Cumulative relative frequency:

It is cumulative frequency converted into the percentage of the total number of observations.

Let us take the examples of various types of data and construct the frequency table.

1. Frequency table for nominal data

Example 10.1

In blood group detection camp, 95 pharmacy students were sampled to have following blood groups.

Blood groups of 95 pharmacy students were as follows:

B, AB, A, A, B, A, AB, A, B, A, A, AB, B, A, A, B, A, AB, A, B, A, A, B, A, AB, A, B, A, B, B, A, AB, A, B, B, A, O, AB, B, A, A, AB, O, B, A, A, B, O, B, A, B, A, A, B, A, A, B, A, B, A, A, B, A, A, AB, A, A, AB, B, A, A, AB, B, A, A, B, A, A, B, A, B, A, B, A, B, A, A, AB, A, AB, A, A, O, AB, A, AB, A, A, B, A.

Solution

As we know, the ordering of nominal categories is arbitrary, and in this example they are shown by the number of students in each – largest first. The total frequency ($n = 95$), is shown at the top of the frequency column. This is helpful for the reader.

1. Frequency

Table 10.1 Frequency table showing the distribution of blood group of 95 pharmacy students

Category of Blood group	Tally marks	Frequency (number of students) $n=95$
A		49
B		27
AB		15
O		04

2. Relative frequency

Table 10.2 Relative frequency table showing the percentage of students in each blood group

Category of Blood group	Frequency (number of students) $n=95$	Relative Frequency (% of students in each category)
A	49	$(49/95)*100 = 51.6$
B	27	$(27/95)*100 = 28.4$
AB	15	$(15/95)*100 = 15.8$
O	04	$(04/95)*100 = 04.2$

3. Cumulative frequency

It makes no sense to calculate cumulative frequency for nominal data, because of the arbitrary category order. Hence, cumulative frequency is not calculated.

2. Frequency table for Ordinal Data

When the variable in question is ordinal, we can allocate the data into ordered categories.

Example 10.2

Let us take an example of 'level of satisfaction' of 60 final year students regarding infrastructure available in the college. The following data is given in numbered form for easy understanding: (4- very satisfied, 3- satisfied, 2- neutral, 1-dissatisfied, 0- very dissatisfied).

Data: 3, 0, 2, 1, 3, 4, 0, 3, 4, 0, 2, 3, 4, 1, 3, 2, 3, 4, 0, 1, 3, 4, 3, 4, 0, 3, 2, 1, 3, 4, 2, 1, 3, 3, 1, 4, 3, 1, 3, 4, 1, 3, 4, 3, 4, 0, 3, 2, 3, 4, 1, 3, 1, 0, 3, 4, 3, 2, 1, 0

Solution:

Level of satisfaction is clearly an ordinal variable. 'Satisfaction' cannot be properly measured, and has no units. But the categories can be meaningfully ordered, as they have been given here.

The frequency values indicate that more than half of the patients were happy with their infra structural facilities, 34 students (13+21), out of 60. Much smaller numbers expressed dissatisfaction.

Table 10.3 The frequency distributions for the ordinal variable 'level of satisfaction' with infrastructure available in college

Satisfaction with infrastructure	Tally marks	Frequency (number of students) n=60
Very satisfied (4)		13
Satisfied (3)		21
Neutral (2)		07
Dissatisfied (1)		11
Very dissatisfied (0)		08

Table 10.4 The relative frequency distributions for data

Satisfaction with infrastructure	Frequency (number of students) n=60	Relative Frequency (% of students in each level of satisfaction)
Very satisfied (4)	13	$(13/60)*100 = 21.7$
Satisfied (3)	21	$(21/60)*100 = 35$
Neutral (2)	07	$(07/60)*100 = 11.7$
Dissatisfied (1)	11	$(11/60)*100 = 18.3$
Very dissatisfied (0)	08	$(08/60)*100 = 13.3$

Table 10.5 The Cumulative and relative cumulative frequency distributions for data

Satisfaction with infrastructure	Frequency (number of students) n=60	Cumulative Frequency (Cumulative number of students)	Relative Cumulative Frequency (Cumulative % of students)
Very satisfied (4)	13	13	$(13/60)*100 = 21.7$
Satisfied (3)	21	$13+21= 34$	$(34/60)*100 = 57.7$
Neutral (2)	07	$34+07= 41$	$(41/60)*100 = 68.3$
Dissatisfied (1)	11	$41+11= 52$	$(52/60)*100 = 86.7$
Very dissatisfied (0)	08	$52+08= 60$	$(60/60)*100 = 100$

3. Frequency Table for Metric Continuous data

Organising raw metric continuous data into a frequency table is usually impractical, because there are such a large number of possible values. The most useful approach with metric continuous data is to group them first, and then construct a frequency distribution of the grouped data.

The construction of frequency distributions in case of metric continuous data requires following three steps:

1. Choosing of class intervals,
2. Tallying the data into these classes, and
3. Counting the tallies in each class (frequency)

While choosing the class intervals, the number of observations to be grouped are very important. We seldom use fewer than 6 or more than 15 classes. The general guide to decide number of intervals for various sample sizes (observations) are given in table below as per Sturges' rule.

Table 10.6 Number of intervals for various sample sizes using Sturges' Rule

Sample Size	K Intervals
23-45	6
46-90	7
91-181	8
182-363	9
364-726	10
727-1454	11
1455-2909	12

In a grouped frequency distribution.

- 1) all class intervals must be of same width, or size;
- 2) the intervals should be mutually exclusive and exhaustive;
- 3) the interval widths should be assigned so the lowest interval includes the smallest observed outcome and the top interval includes the largest specified outcome.

Let's see one example of grouping metric continuous data

Example 10.3

The following are the weights in kg of 60 final year pharmacy students.

50, 61, 70, 61, 78, 56, 71, 63, 66, 75, 77, 52, 80, 45, 56, 57, 58, 60, 62, 72, 78, 48, 50, 63, 51, 64, 67, 52, 53, 54, 55, 56, 57, 70, 71, 62, 72, 57, 73, 64, 65, 66, 67, 62, 63, 64, 65, 52, 60, 54, 56, 63, 58, 57, 61, 76, 58, 84, 46, 50,

Solution:

Here, the smallest value in the observations is 45 while the largest one is 84, the difference of 39. So, 8 groups of width 5 can be taken, to cover all observations.

Table 10.7 The frequency distribution table for metric continuous data

Weight (kg)	Tally marks	Frequency n=60
45-49		03
50-54		10
55-59		12
60-64		15
65-69		06
70-74		07
75-79		05
80-84		02

1. Class:

The group of observations is called as class. In this example there are 8 classes (45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84).

2. Class limits:

The minimum value that can be included in the class is lower class limit while the maximum value that can be included in the class is upper class limit. In this example, for class 45-49, the lower class limit is 45 while upperclass limit is 49.

3. Class boundaries:

Consider the classes 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84 in this example. 49 is the upper class limit for 45-49 class while 50 is the lower class limit for 50-54 class. Here, class limits are not continuous and therefore we have to subtract 0.5 from lower limit and add 0.5 from upper limit. Thus class become continuous as shown in table below and they are called as class boundaries. In case of classes 45-50, 50-55, 55-60, 60-65, 65-70, 70-75, 75-80, 80-85 the class limits are called continuous and hence class limits are called as class boundaries.

Table 10.8 Class boundaries for non continuous class limits

Class limits	Class Boundaries
45-49	44.5-49.5
50-54	49.5-54.5
55-59	54.5-59.5
60-64	59.5-64.5
65-69	64.5-69.5
70-74	69.5-74.5
75-79	74.5-79.5
80-84	79.5-84.5

Class interval:

The difference between class boundaries is called class interval or class width.

Table 10.9 The relative, cumulative and relative cumulative frequency distribution table

Weight (kg)	Frequency n=60	Relative frequency	Cumulative frequency	Relative Cumulative frequency
45-49	03	05	03	05
50-54	10	16.7	13	21.7
55-59	12	20	25	41.7
60-64	15	25	40	66.7
65-69	06	10	46	76.7
70-74	07	11.7	53	88.3
75-79	05	8.3	58	96.7
80-84	02	3.3	60	100

Class mark:

Class marks are simply the midpoints of the classes and they are found by adding lower and upper limits of a class (or its lowest and upper boundaries) and dividing by two.

$$\text{Class mark (midpoint)} = \frac{\text{Lower class boundary} + \text{upper class boundary}}{2} \quad \dots 2$$

Open Ended Groups

One problem arises when one or two values are a long way from the general mass of the data, either much lower or much higher. These values are called outliers. Their presence can mean having lots of empty or near-empty rows at one or both ends of the frequency table.

For example, if one student is having weight of 24 kg in above example, then the groups (class intervals) 40-44, 35-39, 30-34 and 25-29 will have empty cells. In this case open-ended group can be used, here <45 group can be incorporated and frequency of one can be recorded thus avoiding empty cells.

4. Frequency Table for Metric Discrete data

Constructing frequency tables for metric discrete data is generally easy as compared to continuous metric data, because the number of possible values which the variable can take is often limited

Example 10.4

The data given below are the number of times inhaler used in past 24 h by 53 children with asthma. Construct frequency, relative frequency, cumulative frequency, relative cumulative frequency table.

Data: 0, 2, 1, 3, 5, 2, 4, 6, 1, 2, 3, 1, 0, 2, 3, 1, 2, 1, 4, 5, 2, 0, 2, 3, 4, 0, 2, 1, 4, 2, 1, 2, 1, 2, 4, 1, 0, 1, 3, 1, 5, 3, 1, 2, 1, 7, 1, 3, 1, 0, 1, 3, 6.

Table 10.10 The frequency distribution table for metric discrete data

Number of times inhaler used in past 24 h	Tally marks	Frequency n=53
0		6
1		16
2		12
3		8
4		5
5		3
6		2
7		1

Table 10.11 The relative, cumulative and relative cumulative frequency distribution table

Number of times inhaler used in past 24 h	Frequency n=53	Relative frequency	Cumulative frequency	Relative cumulative frequency
0	06	11.32	06	11.32
1	16	30.18	22	41.50
2	12	22.64	34	64.14
3	08	15.09	42	79.23
4	05	09.43	47	88.66
5	03	05.66	50	94.32
6	02	03.77	52	98.09
7	01	01.88	53	100

Cross-tabulation

Each of the frequency tables above provides us with a description of the frequency distribution of a single variable. However, sometimes, we need to examine the association between two variables, within a single group of individuals. We can do this by putting the data into a table of

cross-tabulations, where the rows represent the categories of one variable, and the columns represent the categories of a second variable.

Example 10.5

A study was carried out on the degree of job satisfaction among doctors and nurses in rural and urban areas. To describe the sample a cross-tabulation was constructed which included the sex and the residence (rural or urban) of the doctors and nurses interviewed. This was useful because in the analysis the opinions of male and female staff had to be compared separately for rural and urban areas.

Table 10.12 Type of health worker by residence

Residence	Type of Health Worker		Total
	Doctors	Nurses	
Rural	10 (16%)	69 (38%)	79 (33%)
Urban	51 (54%)	113 (62%)	164 (67%)
Total	61 (100%)	182 (100%)	243 (100%)

Table 10.12 a shows that a higher percentage of nurses than of doctors work in rural areas, but that, overall, a greater proportion of staff works in urban areas (67%).

Graphical Presentation of Data

1. The Pie Chart

Graphical presentation of data with appropriate chart is a good idea for describing data effectively. Appropriate chart depends primarily on the type of data, as well as on what particular features of it we are looking for.

Graphical presentation of Nominal and Ordinal data

The pie chart is a diagram in which the frequencies of the groups are shown in a circle. Each segment (slice) of a pie chart should be proportional to the frequency of the category it represents.

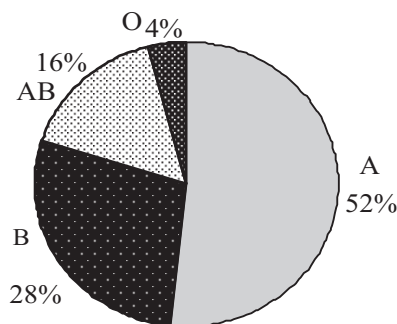


Figure 10.2 Pie chart of percentage blood group of pharmacy students

For example, Figure 10.2 is a pie chart of blood group of each of 95 pharmacy students

shown in Table 10.2. A disadvantage of a pie chart is that it can only represent one variable (in Figure 10.2, blood group). We will therefore need a separate pie chart for each variable we want to chart. A disadvantage of pie chart can lose clarity if it is used to represent more than four or five categories.

2. The Simple Bar Chart

An alternative to the pie chart for nominal data is the bar chart. This is a chart with frequency on the vertical axis and category on the horizontal axis. The simple bar chart is appropriate if only one variable is to be shown. Figure 10.3 is a simple bar chart of blood group of pharmacy students. Note that all the bars should be of the same width, and there should be equal spaces between bars. These spaces emphasise the categorical nature of the data.

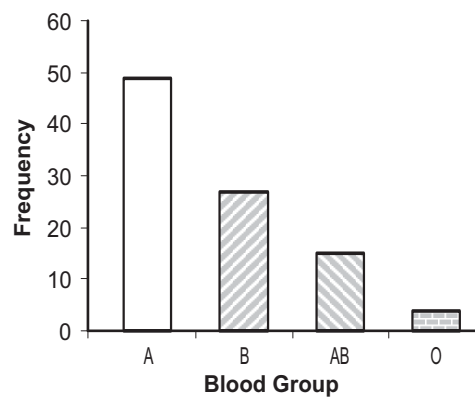


Figure 10.3 Simple bar chart of blood group of pharmacy students

3. The Clustered Bar Chart

If there are more than one group, we can use the clustered bar chart. Suppose we are knowing the sex of the students in the above example, then it will give us two sub-groups, boys and girls, with the data shown in Table 10.13. There are two ways of presenting a clustered bar chart. Figure 10.3 shows one, with blood group categories on the horizontal axis. This arrangement is helpful if we want to compare the relative sizes of the groups within each category.

Table 10.13 The frequency distribution table of blood group of 95 pharmacy students by sex

Category of Blood group	Frequency (number of Boys) n=48	Frequency (number of Girls) n=47
A	04	11
B	29	20
A	01	03
O	14	13

Alternatively, the chart can also be drawn with the categories boys and girls, on the

horizontal axis. This format is more useful if we want to compare category sizes within each group.

Example 10.6

Girls with blood group A, B, AB and O can be compared. Which chart is more appropriate depends on what aspect of the data we want to examine.

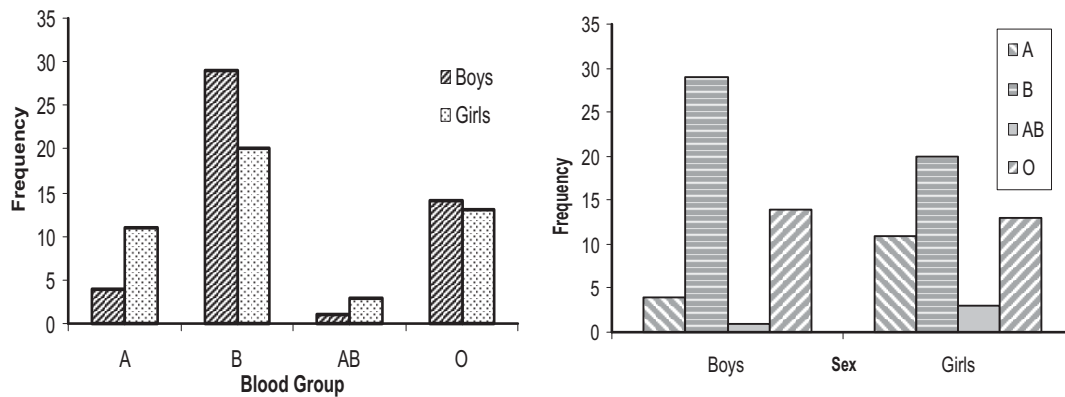


Figure 10.4 Clustered bar chart of blood group of 95 pharmacy students by sex

4. The Stacked Bar Chart

Figure 10.5 shows a stacked bar chart for the blood group and sex data shown in Table 10.13. Instead of appearing side by side, as in the clustered bar chart of Figure 10.4, the bars are now stacked on top of each other. Stacked bar charts are appropriate if we want to compare the total number of subjects in each group (total number of boys and girls for example), but not so good if we want to compare category sizes between groups.

Example 10.7 Blood groups of girls with blood groups of boys can be compared.

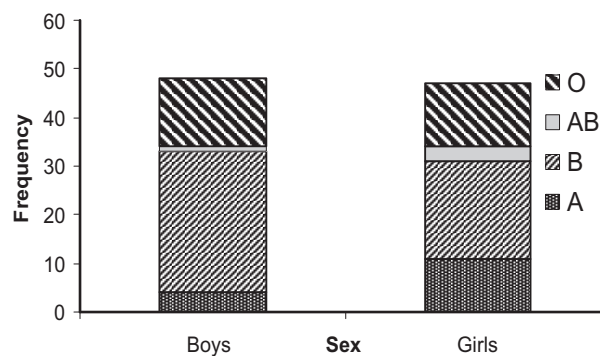


Figure 10.5 A stacked bar chart of blood group of 95 pharmacy students by sex

5. Pictograms

Pictograms are similar to bar charts. They present the same type of information, but the bars

are replaced with a proportional number of icons. This type of presentation for descriptive statistics dates back to the beginning of civilization when pictorial images were used to record numbers of people, animals or objects.

Example 10.8

Population of different districts of Western Maharashtra. Each diagram indicates one lakh population.

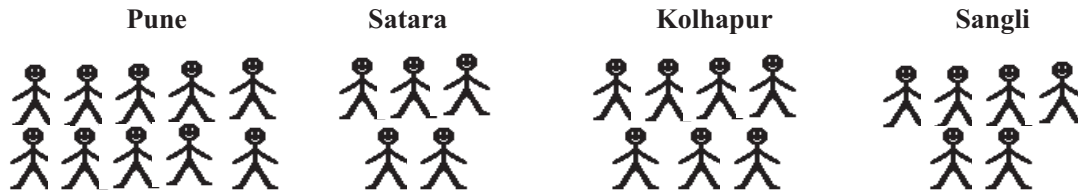


Figure 10.6 Population of different districts of Western Maharashtra

Graphical Presentation of Metric Discrete Data

1. Bar Chart

We can use bar charts to graph discrete metric data in the same way as with ordinal data.

Example 10.9:

The data given in table 10.10 are the number of times inhaler used in past 24 h by 53 children with asthma. Construct bar chart.

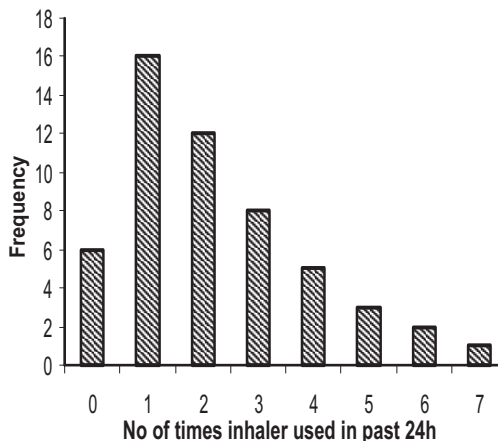


Figure 10.7 The frequency distribution table for metric discrete data

2. Line Plot

A line chart is similar to a bar chart except that thin lines, instead of thicker bars, are used to represent the frequency associated with each level of the discrete variable.

Example 10.10

Assay result of 30 amoxicillin capsules are as follows

Data: 251, 250, 253, 249, 250, 252, 247, 248, 254, 245, 250, 253, 251, 250, 249, 252, 249, 251, 246, 250, 250, 254, 248, 252, 251, 248, 250, 247, 251, 249.

Table 10. 14 Frequency distribution table for assay result of 30 amoxicillin tablets

Assay Result (mg)	Tally marks	Frequency n=30
245		1
246		1
247		2
248		3
249		4
250		7
251		5
252		3
253		2
254		2

Using the data presented in Table 10.14, a corresponding line chart is presented in Figure 10.8.

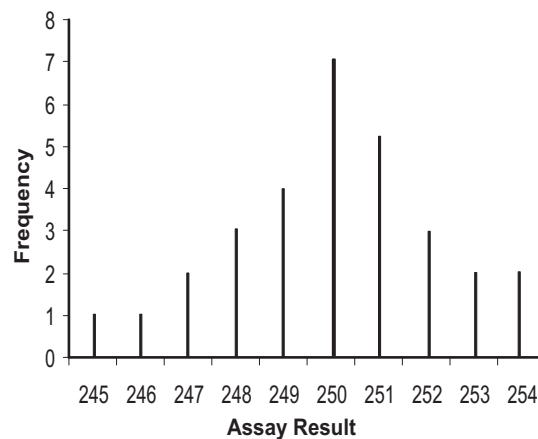


Figure 10.8 Line Plot of assay result of 30 amoxicillin tablets

3. Point Plot

Point plots are identical to line charts, however, instead of a line, a number of points or dots equivalent to the frequency are stacked vertically for each value of the horizontal axis. Also referred to as dot diagram, point plots are useful for small data sets.

Using the data presented in Table 10.14, a corresponding point diagram is presented in Figure 10.9.

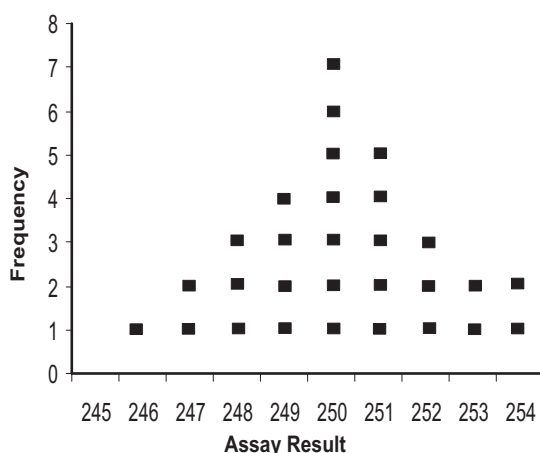


Figure 10.9 Point Plot of assay result of 30 amoxicillin tablets

Graphical Presentation of Metric Continuous Data

In all work with graphs, we can use two axes. The vertical axis is always labeled as Y axis. It is also called as “Ordinate”. The values taken along this axis are called ordinate values. The horizontal axis is always labeled as ‘X’ axis. It is also called “Abscissa”. The X axis and the Y axis meet at right angles at a point of origin (O).

In most of the graphs X axis is longer than Y axis. Usually a ratio of 3:2 or 4:3 will result in a good graph. If X axis is taken 18 cm long then Y axis should be 12 cm long. On X axis we will mark different groups, scores, class intervals, and on Y axis we will usually take frequencies.

1. The Histogram

A continuous metric variable can take a very large number of values, so it is usually impractical to plot them without first grouping the values. The grouped data is plotted using a frequency histogram, which has frequency plotted on the vertical axis and group size on the horizontal axis.

A histogram looks like a bar chart but without any gaps between adjacent bars. This emphasises the continuous nature of the underlying variable. If the groups in the frequency table are all of the same width, then the bars in the histogram will also be of the same width. One limitation of the histogram is that it can represent only one variable at a time (like the pie chart), and this can make comparisons between two histograms difficult.

Figure 10.10 shows a histogram of the grouped weight in kg of 60 final year pharmacy students as per the data in Table 10.8.

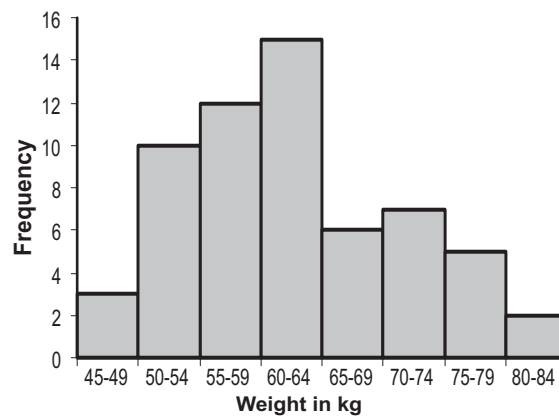


Figure 10.10 A histogram of the grouped weight in kg of 60 final year pharmacy students

2. Frequency Polygon

A frequency polygon can be constructed by placing a dot at the midpoint (class mark) for each class interval in the histogram and then these dots are connected by straight lines. This frequency polygon gives a better conception of the shape of the distribution. The class interval midpoint for a section in a histogram is calculated as follows:

$$\text{Midpoint} = (\text{highest} + \text{lowest point})/2 \quad \dots 3$$

The frequency polygon is then created by listing the midpoints (class marks) on the x axis, frequencies on the y-axis, and drawing lines to connect the midpoints for each interval.

Construction of frequency polygon

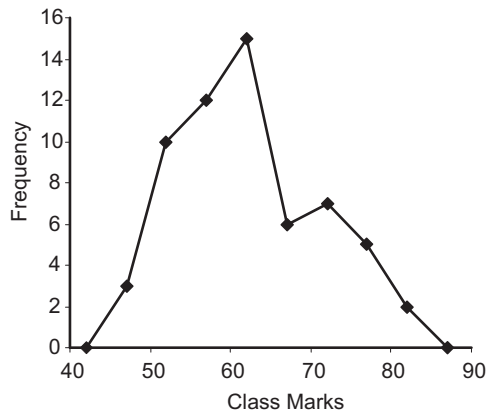
We will take the values in table 10.7 for drawing frequency polygon. Let us mark frequencies on y-axis and groups or class intervals (weights) on x-axis.

Steps in drawing frequency polygon are given below:

1. Let us take the first group with class interval 45-49. The class intervals in this example are not continuous and hence the class boundaries are defined first. Then the midpoint (class mark) is calculated (see table 15) using the formula given above. So, place a point corresponding to 47 on X axis and 3 on Y axis.
2. The same way, mid point of next group (50-54) is 52. This group is having frequency of 10. So, we have to place a point corresponding to 52.2 on X-axis and 10 on Y-axis.
3. The same way all frequencies are marked on the corresponding mid points of the groups.
4. Then with a ruler we should connect these points with straight line.
5. Rather than leaving the graph suspended in space, we assume that there is another interval above and below which is having frequency of zero. So, the mid points of group 40-44 and 85-89 are assumed having frequency of zero. The group is now allowed to meet X-axis on both ends.

Table 10.15

Class limits	Class Boundaries	Midpoint(Class Mark)	Frequency
45-49	44.5-49.5	47	03
50-54	49.5-54.5	52	10
55-59	54.5-59.5	57	12
60-64	59.5-64.5	62	15
65-69	64.5-69.5	67	06
70-74	69.5-74.5	72	07
75-79	74.5-79.5	77	05
80-84	79.5-84.5	82	02

**Figure 10.11 A** Frequency Polygon of the grouped weight in kg of 60 final year pharmacy students**3. Relative Cumulative Frequency Curve (Ogive)**

With continuous metric data, there is assumed to be a smooth continuum of values, so we can chart relative cumulative frequency with a correspondingly smooth curve, known as a relative cumulative frequency curve, or ogive.

Construction of Ogive

Ogive is a graph of the cumulative relative frequency distribution. So, to draw Ogive we should convert ordinary frequency distribution into relative cumulative frequency.

Example: Construct relative cumulative frequency diagram for the data given in Example 10.3

Solution

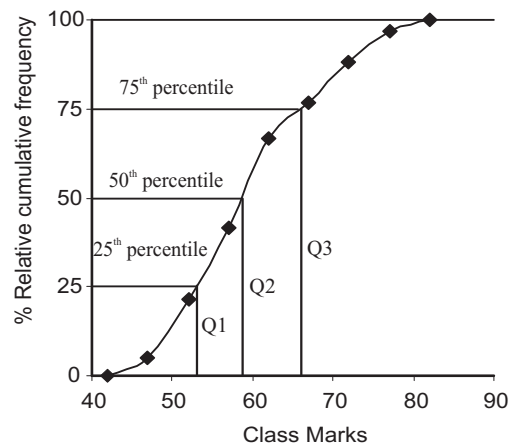
1. Construct the table of the cumulative frequency distribution as shown below. Here cumulative frequency means “Total number of students in each particular weight range from lowest value to that particular value”. It is obtained by cumulating the frequency of previous classes including the class in question.

Table 10.16

Class limits	Class Boundaries	Midpoint	Frequency	Cumulative frequency	Relative cumulative frequency
45-49	44.5-49.5	47	03	03	05.0
50-54	49.5-54.5	52	10	13	21.7
55-59	54.5-59.5	57	12	25	41.7
60-64	59.5-64.5	62	15	40	66.7
65-69	64.5-69.5	67	06	46	76.7
70-74	69.5-74.5	72	07	53	88.3
75-79	74.5-79.5	77	05	58	96.7
80-84	79.5-84.5	82	02	60	100

2. Now plot the relative cumulative frequencies on y-axis while midpoints of class intervals on x-axis. So, place a point corresponding to 47.5 on x-axis and 5 on y-axis. Likewise plot all relative cumulative frequencies on corresponding midpoints of the group.

3. Now, join the points with a free hand to give a smooth curve. This is the Ogive as shown below.

**Figure 10.12** Ogive of the grouped weight of 60 final year students

Applications of Ogive

1. By using Ogive we can locate any percentile that will divide the series into two parts.
2. Quartiles: There are three different points located on the entire range of variable (here it is weight in kg). These are Q1, Q2, Q3.

Q1 or lower quartile will have 25% observations falling in its left and 75% observations on its right side.

Q2 is the median, i.e., 50% values lies on either side.

Q3 is the upper quartile, will have 75% observations falling on its left side and 25%

observations on its right side.

By using these quartiles we can calculate semi inter quartile range and inter quartile range (Q1-Q3).

3. Quintiles: This divides the distribution into 5 equal parts. So, 20th percentile or 1st quintile will have 20% observations falling to its left and 80% to its right.

4. Deciles: This divides the distribution into 10 equal parts. First decile (10th percentile) will have 10% values to its left and 90% values to its right. 5th decile is the median and contains 50% values on either side.

4. Stem and Leaf plot

The stem-and-leaf plot is a visual representation for continuous data and contains features common to both the frequency distribution and dot diagrams. Digits, instead of bars are used to illustrate the spread and shape of the distribution. Each piece of data is divided into "leading" and "trailing" digits. For example, based on the range of data points, the observation 116 can be divided into either 11 and 6, or 1 and 16, as the leading and trailing digits. All the leading digits are sorted from lowest to highest and listed to the left of a vertical line. These digits become the stem. The trailing digits are then written in the appropriate location to the right of the vertical line. These become the leaves.

Example 10.11

Plot the Stem and Leaf graph for following data of age in years of 27 patients.

8, 13, 16, 25, 26, 29, 30, 32, 37, 38, 40, 41, 44, 47,
49, 51, 54, 55, 58, 61, 63, 67, 75, 78, 82, 86, 95

Solution

1. We can group 27 patients (8-95 yrs) according to age groups. Here we will take groups of 10 yrs each.

0-9 10-19 20-29 30-39 40-49 50-59 60-69 70-79 80-89 90-99

2. We can now write these groups on left side of the vertical line and place the last digit of the age on the right of the vertical line to the corresponding group (see figure 10.13).

3. Now, groups can be considered as stem and corresponding digits on the right as leaves. This can be rewritten in figure 10.14.

4. From stem and leaves, we can read the exact ages of the patients and the data is not lost.

5. Box and Whiskers plot

One simple plot that displays a great deal of information about a continuous variable is the box-and-whisker plot. The box plot illustrates the bulk of the data as a rectangular box in which the upper and lower lines represent the third quartile (75% of observations below Q3) and first quartile (25% of observations below Q1), respectively. The second quartile (50% of the observations below this point) is depicted as a horizontal line through the box. Vertical lines (whiskers) extend from the

top and bottom lines of the box to an upper and lower adjacent value. The details of drawing box and whiskers plot is discussed in next chapter (Example 12.5).

Stem	Leaves
0-9	8
10-19	3 6
20-29	5 6 9
30-39	0 2 7 8
40-49	0 1 4 7 9
50-59	1 4 5 8
60-69	1 3 7
70-79	5 8
80-89	2 6
90-99	5

Figure 10.13 Stem and Leaf plot of grouped data

Stem	Leaves
0	8
1	3 6
2	5 6 9
3	0 2 7 8
4	0 1 4 7 9
5	1 4 5 8
6	1 3 7
7	5 8
8	2 6
9	5

Figure 10.14 Stem and Leaf plot of ungrouped data

6. Scatter diagram or Scatter plot

A scatter diagram is an extremely useful presentation for showing the relationship between two continuous variables. The two dimensional plot has both horizontal and vertical axes which cover the ranges of the two variables. Plotted data points represent paired observations for both the x and y variable (Figure 10.14). These types of plots are valuable for correlation and regression inferential tests.

Example 10.12:

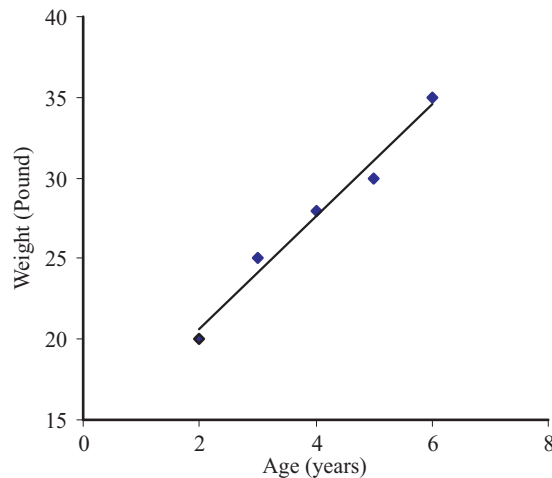
The table shows the age and the weight of a child. Plot the scatter graph.

Table 10.17 Age and weight of five children

Age in years	Weight in pounds
2	20
3	25
4	28
5	30
6	35

Solution

1. Take the weight of children on y-axis while age in years on x-axis.
2. Take appropriate scale on y-axis covering the weight from 20-35.
3. Take appropriate scale on x-axis covering the age from 1-6.
4. Now put the marks of weight to the corresponding ages of children.
5. Draw the line connecting the marks (points) to get the scatter plot.

**Figure 10.15** Scatter plot**7. Time Series chart**

If the data collected are from measurements made at regular intervals of time (minutes, weeks, years, etc.), we can present the data with a time series chart. Usually these charts are used with metric data, but may also be appropriate for ordinal data. Time is always plotted on the horizontal axis, and data values on the vertical axis.

Charting of Data Using MS - Excel

1. Construction of Pie Chart

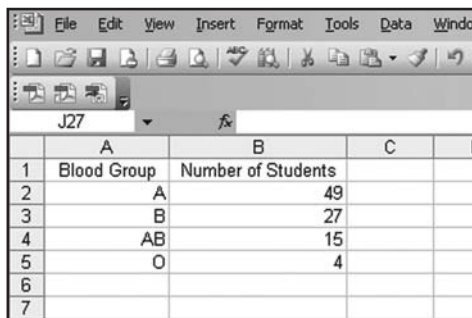
Example 10.13

Blood group of 95 students are as follows

Category of blood group	A	B	AB	O
Number of students	49	27	15	04

Excel Solution

- Step I**
1. Open New MS Excel File.
 2. Put labels in column A1 as “Category of blood group” and B1 as “Number of students”.
 3. Put values of blood group from cell A2 to A5 and number of students from Cell B2 to B5.
 4. Sheet will appear as shown in figure 10.16.



	A	B	C	D
1	Blood Group	Number of Students		
2	A	49		
3	B	27		
4	AB	15		
5	O	4		
6				
7				

Figure 10.16 Entering data into worksheet

- Step II**
1. Select cells from A1 to B5.
 2. Click on Insert from menu bar. Instantly pull down menu will appear.
 3. Select ‘Chart’ from pull down menu. Instantly “Chart Wizard- Step 1 of 4 - Chart Type” display box will appear as shown in figure 10.17.

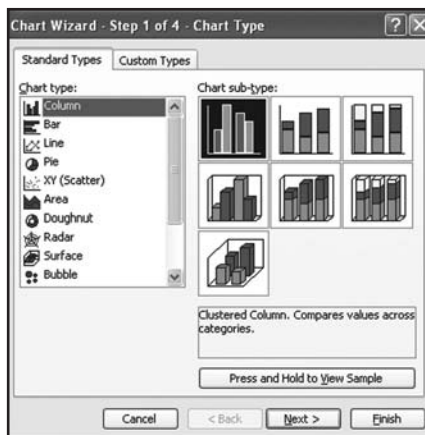


Figure 10.17 Window of Chart type

4. Select ‘Pie’ from Chart type as shown in figure 10.7 and click on Next button

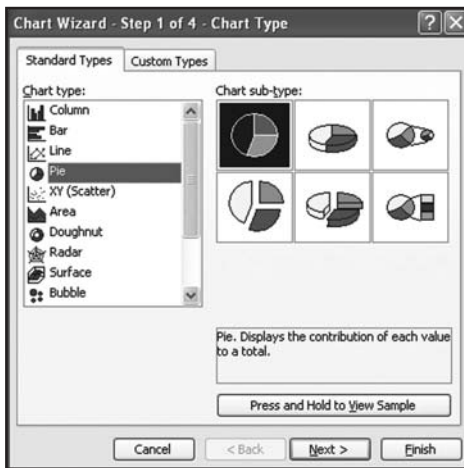


Figure 10.18 Window of selecting Pie Chart

5. Instantly Chart Wizard- Step 2 of 4- Chart Source Data will appear as shown in figure 10.19.
6. If the appeared graph is correct then click on next button.
7. If Graph is not correct click on series and make the corrections by changing Values and Category Labels, as shown in figure 10.20. Then click on Next button.

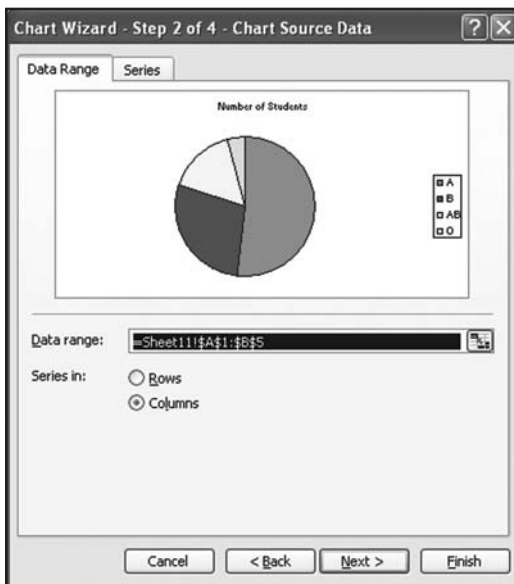


Figure 10.19 Window of Chart Source Data

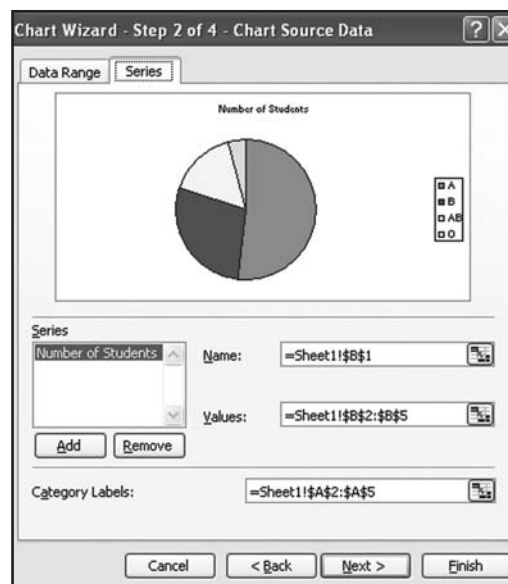


Figure 10.20 Window of selecting Series option

8. Instantly Chart Wizard - Step 3 of 4- Chart Options will appear.
9. Click to Value from 'Label Contains' as shown in figure 10.21 and then click to Next

button.

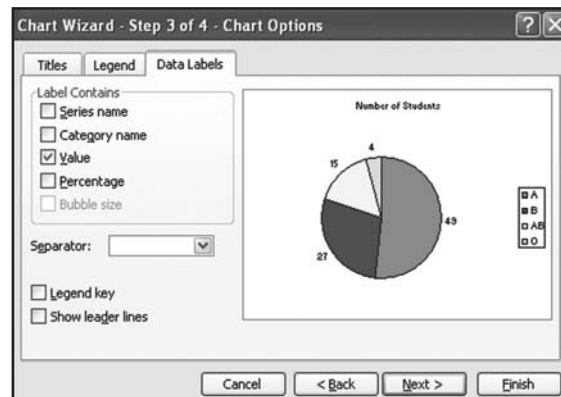


Figure 10.21 Window of Chart Options

10. Instantly Chart Wizard - Step 4 of 4 - Chart Location will appear as shown in figure 10.22

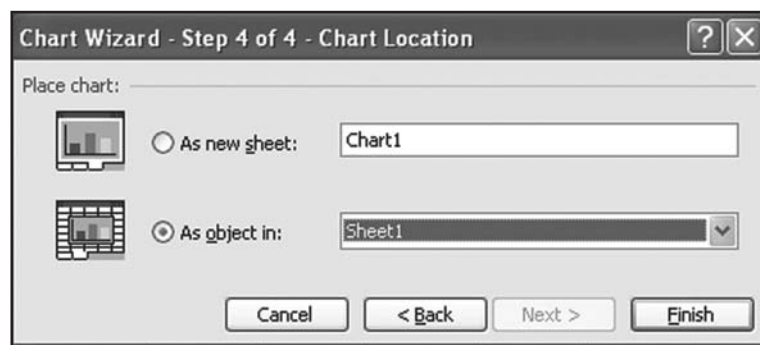


Figure 10.22 Window of Chart Location

11. Ignore the things and click on 'Finish' button. Following graph will be displayed on same sheet.

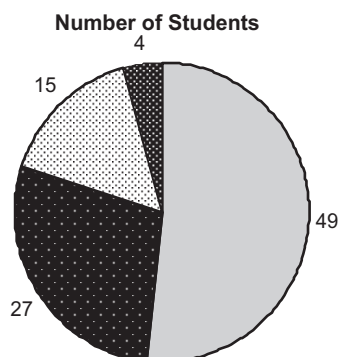


Figure 10.22 Pie Chart

13. We can make changes in the graph properties according to our needs.

2. Construction of Line or Scatter Graph in Excel

1. First put the values of X and Y in the excel sheet and label the columns. Make sure that your X values are in one column and that Y values are in a second column with each row containing the matching X, Y values.
2. Now click on the picture of a graph in the toolbar at the top. Alternatively, we can go to “Insert Chart” from the drop-down menus.
3. Step 1 of the Chart Wizard will appear as shown in previous example. Pick the picture and description that best matches the type of graph we are trying to draw. Choose the “XY (Scatter)” category. We can additionally pick “Scatter with data points connect by smoothed lines” or “Scatter with data points connected by lines” to connect the points with a smoothed line or with straight lines. Click “Next”.
4. Step 2 of the Chart Wizard, Chart Source Data, will appear. Go to the tab ‘Series’. It contains the following fields. Fill in the fields as listed below:
 - a. Name: Fill in the words that will appear in the legend.
 - b. X Values: Click on the picture of the spreadsheet at the right of the blank. This takes us to the spreadsheet where we can highlight data for X (just the numbers, not the column heading).
 - c. Y Values: Click on the spreadsheet, and highlight data for Y (just the numbers, not the column heading).Click Next.
5. Step 3 of the Chart Wizard has several different tabs, as shown below:
 - a. Titles tab –
Chart Title: This is the graph title that will appear above the graph by default.
Value X axis: Enter the X axis label here. Don't forget to include appropriate units in parentheses.
Value Y axis: Enter the Y axis label here. Don't forget to include appropriate units in parentheses.
 - b. Axes tab – We can alter the appearance of the X or Y-axis.
 - c. Gridlines tab – We can check the boxes on this tab to make gridlines appear or disappear.
 - d. Legend tab – We can check the boxes on this tab to make series legend appear or disappear and even we can indicate where on the page legend should appear.
 - e. Data Labels tab – We can check boxes to show specific data labels and values. click “Next” after making the choices on step 3.
6. Step 4 of the chart wizard simply asks where we want graph to appear. We can check for the graph to appear by itself on a separate sheet in the workbook, or to appear as an object embedded in the current sheet of the notebook, next to the data.

3. Construction of a Bar Graph in Excel

1. Put the X axis and Y axis label in the Excel sheet. Then type the category data (word or numbers) in column of X axis label and type corresponding frequency values in column of Y axis label.

2. Now click on the picture of a graph in the toolbar at the top. Alternatively, we can go to “Insert Chart” from the drop-down menus.
3. Step 1 of the Chart Wizard will appear. A bar graph is called a “Column Chart” in Excel; so select “Clustered Column Chart”. Select that and click “Next”
4. Step 2 of the Chart Wizard has 2 tabs, the “Data Range” & “Series” tabs.
 - a. The “Data Range” tab asks us to pick data range. If data is not highlighted previously we can use the selection tool to do so from this screen.
 - b. Click on the “Series” tab.
In the space labeled “Name” we can type the title of the graph.
In the space labeled “Category (x) axis labels”, we can highlight a data range that contains the words or numbers that we would like to appear under each bar. Click on the picture of the spreadsheet beside the blank to allow us to highlight the range on the spreadsheet. Hit “enter”, then click “Next”.
5. Step 3 of the Chart Wizard has several different tabs.
 - a. Titles tab –
Chart Title: This is the graph title that will appear above the graph by default. We have already entered title under step 2.
Category X axis : Enter the X axis label here.
Value Y axis: Enter the Y axis label here.
 - b. Axes tab – We can change the type of scaling on X axis & Y axis.
 - c. Gridlines tab – We can check the boxes on this tab to make gridlines appear or disappear.
 - d. Legend tab – We can check the boxes to make series legend appear or disappear.
 - e. Data Labels tab – We can check boxes to show specific data labels and values.
 - f. Data Table tab – We can use this tab showing the data table on the graph.
 - g. Hit “Next” after making choices on step 3.
6. Step 4 of the chart wizard simply asks where we want the graph to appear. We can check for the graph to appear by itself on a separate sheet in the workbook, or to appear as an object embedded in the current sheet of the notebook, next to the data.

Summary

1. Tabulation of data

Frequency distribution table

1. Frequency: Repetition of observations
2. Relative frequency: Frequency converted into Percentage
3. Cumulative frequency: Cumulative total of frequencies
4. Cumulative relative frequency: Cumulative frequency converted into percentage

2. Graphical presentation of nominal and ordinal data

- | | | |
|----------------------|---------------------|------------------------|
| 1. Pie Chart | 2. Simple Bar Chart | 3. Clustered Bar Chart |
| 4. Stacked Bar Chart | 5. Pictogram | |

3. Graphical presentation of metric discrete data

1. Bar Chart 2. Line Chart 3. Point Chart

4. Graphical presentation of metric continuous data

1. Histogram 2. Frequency polygon
 3. Ogive 4. Stem and Leaf plot
 5. Box and Whisker plot 6. Scatter plot
 7. Time series plot.

Choosing an appropriate chart

Chart Type	Nominal	Ordinal	Type of data	
			Metric discrete	Metric Continuous
Pie chart	yes	no	no	no
Bar chart	yes	yes	yes	no
Histogram	no	no	yes	yes
Frequency polygon	no	no	yes	yes
Ogive	no	no	yes (cumulative)	yes (cumulative)
Box & Whiskers	no	no	no	yes
Stem & Leaf	no	no	no	yes

Multiple choice questions

- Which of the following would be most suitable for displaying the proportions of a budget spent on different items by pharmaceutical company?
 - Pie chart
 - Bar chart
 - Line graph
 - Histogram
- Bar charts may be distinguished from histograms at a glance because:
 - bar charts are not used for time series data
 - histograms are used to display discrete data
 - bar charts are based on area under the curve
 - histograms do not have spaces between consecutive columns
- A graph that uses vertical bars to represent data is called a _____.
 - Line graph
 - Bar graph
 - Scatter plot
 - Vertical graph
- Ogive is _____.
 - Subset of population
 - Data collected over a period of time
 - A variable with qualitative data
 - A graph of cumulative distribution
- Which of the following is used for graphical presentation of metric discrete data.
 - Bar Chart
 - Line Chart
 - Point Chart
 - All of above

6. The difference between class boundaries is called _____.
 - a. class limit
 - b. class interval
 - c. class frequency
 - d. class mark
7. When the value is away from general mass of data, it is called as _____.
 - a. mean
 - b. open ended
 - c. outlier
 - d. close ended
8. _____ divides the distribution into 5 equal parts.
 - a. Quartile
 - b. Quintiles
 - c. Deciles
 - d. Median
9. _____ plot is useful for showing the relationship between two continuous variables.
 - a. Scatter
 - b. Point
 - c. Stem and leaf
 - d. Box and Whiskers
10. Cumulative frequency converted into percentage is called _____.
 - a. relative frequency
 - b. cumulative relative frequency
 - c. frequency
 - d. none of above

Exercise

1. Based on 1210 patients involved in clinical trials of an anticancer drug, it was observed that 841 experienced no adverse effects, while 256, 91 and 22 subjects suffered mild, moderate and severe adverse effects respectively. Prepare tabular and graphical presentation for this data with proper reasoning for choosing particular graphs.

2. The following assay results (percentage of label claim) were observed in 50 random samples during a production run.

102, 100, 96, 99, 101, 102, 100, 105, 97, 100, 92, 103, 101, 100, 99, 102, 96, 100, 101, 98, 107, 95, 98, 100, 100, 99, 97, 104, 101, 103, 98, 101, 100, 105, 99, 101, 102, 100, 87, 98, 101, 103, 93, 99, 101, 97, 100, 102, 99, 104.

Tabulate the data and report results as a stemplot.

3. Construct table of frequency, relative frequency, cumulative frequency and relative cumulative frequency for following data of weights in kg of 50 pharmacy students and present data graphically.

51, 53, 52, 39, 58, 48, 45, 56, 62, 64, 66, 67, 42, 48, 43, 44, 45, 47, 52, 54, 50, 49, 38, 39, 38, 32, 34, 36, 33, 34, 36, 38, 42, 43, 48, 49, 50, 51, 52, 30, 31, 30, 32, 45, 45, 47, 46, 48, 49, 58.

4. Prepare Ogive curve for the following data.

Height of groups (cm)	72-76	76-80	80-84	84-88	88-92	92-96	96-100
Frequency	5	7	10	15	11	9	7

5. Construct scatter plot of the responses given by a drug in 15 patients.

Concentration of drug (mg/l)	4	6	8	10	12	14	16	18	20	22	24	26
Response	1	2	3	4	5	6	7	8	9	10	11	12

6. Following data provides the number of deaths for several leading causes for the year 2010. Display

the results in a pie chart.

Cause of death	Number of deaths
Heart disease	10382
Cancer	8299
Cerebrovascular disease	2830
Accidents	1381
Other causes	11476

7. Draw a frequency polygon for the data given in Exercise 3.

8. The following is the distribution of number of painkiller tablet taken by 60 patients in last one week.

Number of tablets	0-1	2-3	4-5	6-7	8-9
Frequency	16	25	13	4	2

Find class marks, class boundaries and class intervals of the distribution.

9. List the data which correspond to the following stem and leaf plots:

a) 12		4	0	3	8	7	6	6	5
b) 34		42	05	19	61				

10. Among histograms, bar charts, and pie charts, which ones can be used to represent

- a) nominal data b) ordinal data c) metric data

Answers:

Multiple Choice Questions

1. a 2. d 3. b 4. d 5. d 6. b 7. c 8. b 9. a 10. b

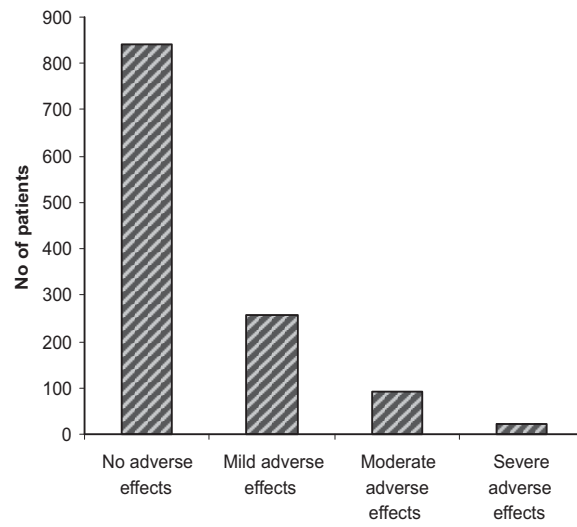
Exercise

1. Tabular presentation of data

Grade of adverse effect	Frequency n=1210	Relative frequency	Cumulative frequency	Relative Cumulative Frequency
0	841	69.50	841	69.50
1	256	21.15	1097	90.66
2	91	7.52	1188	98.18
3	22	1.81	1210	100

No adverse effect (0); Mild adverse effect (1); Moderate adverse effect (2); Severe adverse effect (3)

Graphical presentation of data: The given data is ordinal categorical type and hence can be represented as Bar chart.



2. Stem plot: The given data varies from 87 to 107 and therefore the stems should be 8, 9 and 10

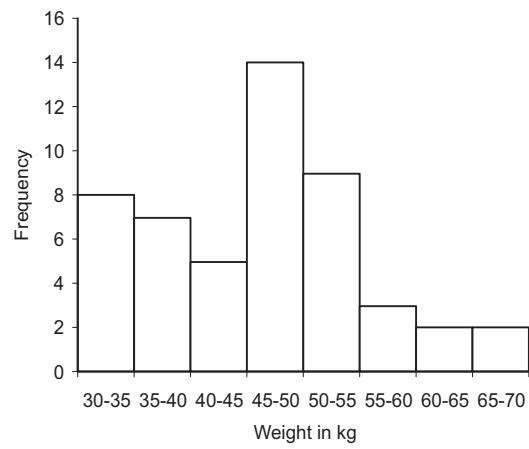
8	7
9	2 3 5 6 6 7 7 7 8 8 8 8 9 9 9 9 9
10	0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 2 2 2 2 2 3 3 3 4 4 5 5 7
Stems	Leaves

Histogram is to be given.

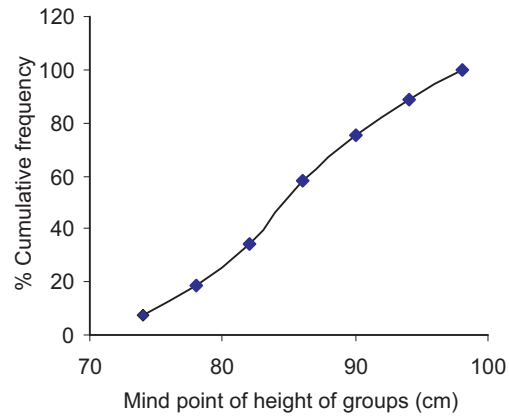
3. Tabular presentation of data

Class interval	Frequency n=50	Relative frequency	Cumulative frequency	Relative Cumulative Frequency
30-35	8	16	8	16
35-40	7	14	15	30
40-45	5	10	20	40
45-50	14	28	34	68
50-55	9	18	43	86
55-60	3	6	46	92
60-65	2	4	48	96
65-70	2	4	50	100

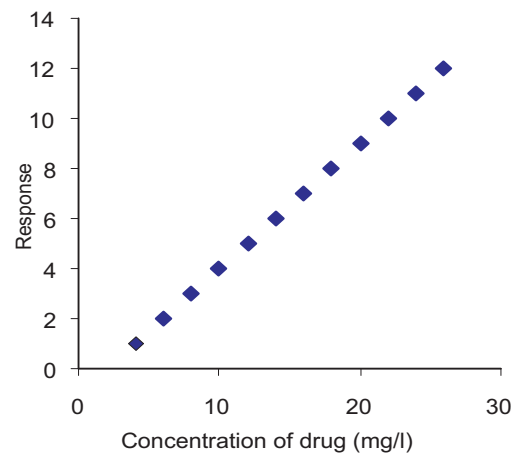
As the data is metric continuous, histogram is the best choice



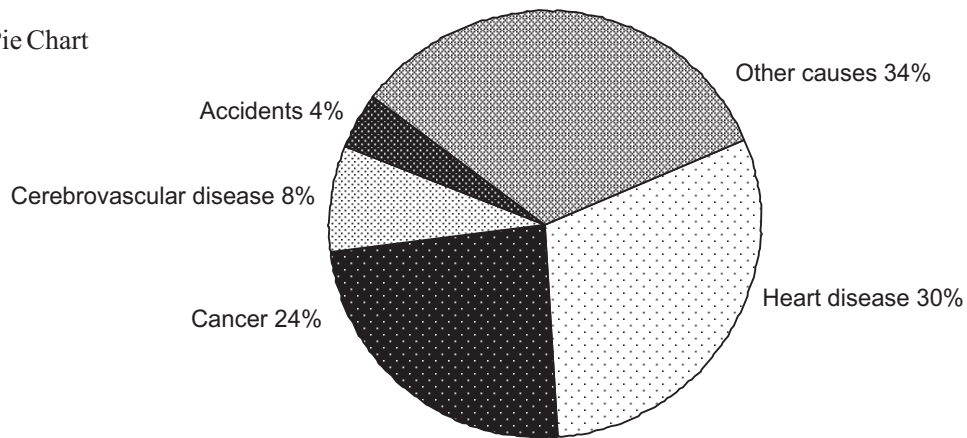
4. Ogive



5. Scatter plot

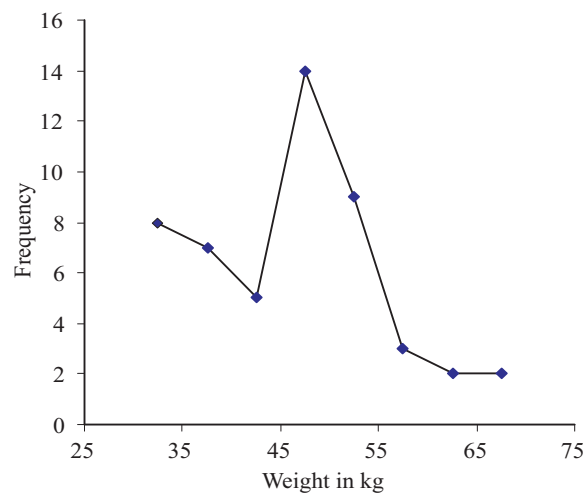


6. Pie Chart



Leading cause of Death

7. Frequency polygon



8. a) Class marks: 0.5 2.5 4.5 6.5 8.5
 b) Class boundaries: -0.5 1.5 3.5 5.5 7.5 9.5
 c) 2
9. a) 124 120 123 128 127 126 126 125
 b) 3442 3405 3419 3461
10. a) nominal data: Histogram
 b) ordinal data: Bar chart
 c) metric data: Nominal data

Chapter 11

SHAPE OF DISTRIBUTION OF DATA

Learning objectives

When we have finished this chapter, we should be able to:

1. Explain what is meant by the 'shape' of a frequency distribution.
2. Sketch and explain: negatively skewed, symmetric and positively skewed distributions.
3. Sketch and explain a bimodal distribution.
4. Describe the approximate shape of a frequency distribution from a frequency table or chart.
5. Sketch and describe a Normal distribution.

Shape of distribution of data

The choice of the most appropriate procedures for summarising and analysing data will not only depend on the type of variable but also on the shape of the distribution.

The shape of distribution of data may be of following types:

1. Uniform distribution: The values are fairly evenly spread throughout their possible range. This is a uniform distribution.

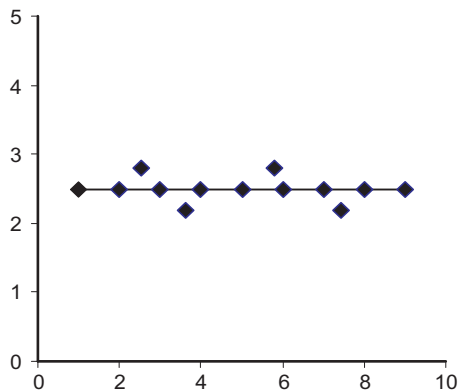


Figure 11.1 Chart showing uniform distribution of data

2. Positively skewed distribution: The values are concentrated towards the bottom of the range, with progressively fewer values towards the top of the range. This is a right or positively skewed distribution. See figure 11.2 for positively skewed distribution.

3. Negatively skewed distribution: The values are concentrated towards the top of the range, with progressively fewer values towards the bottom of the range. This is a left or negatively skewed distribution. See figure 11.3 for negatively skewed distribution.

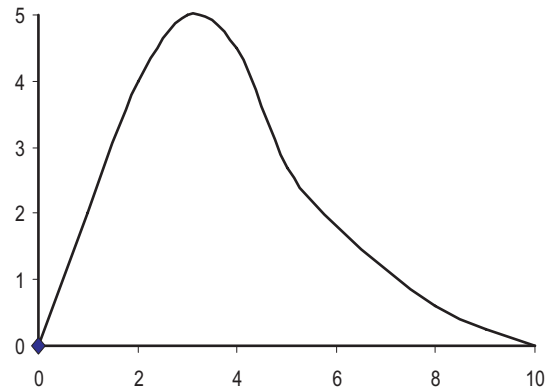


Figure 11.2 Chart showing positively skewed distribution of data

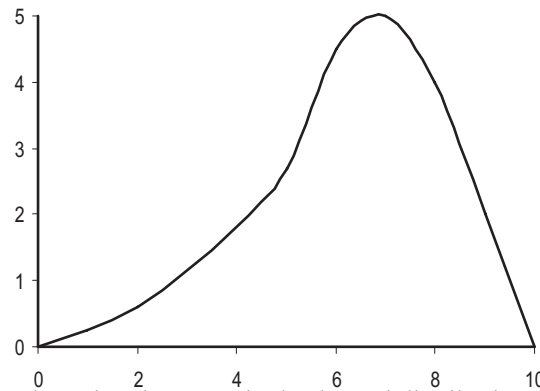


Figure 11.3 Chart showing negatively skewed distribution of data

4. Symmetric or Mound-shaped distribution: The values are clumped together around one particular value, with progressively fewer values both below and above this value. This is a symmetric or mound-shaped distribution

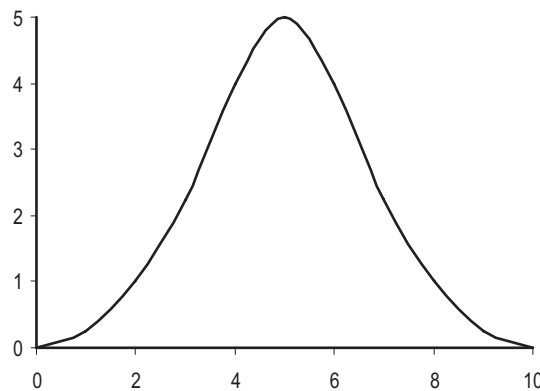


Figure 11.4 Chart showing symmetric or mound shaped distribution of data

5. Bimodal or Multimodal distribution: The values are clumped around two or more particular values. This is a bimodal or multimodal distribution.

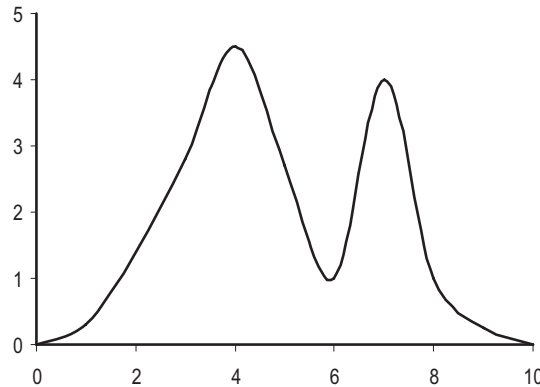


Figure 11.5 Chart showing bimodal distribution of data

One simple way to assess the shape of a frequency distribution is to plot a bar chart, or a histogram.

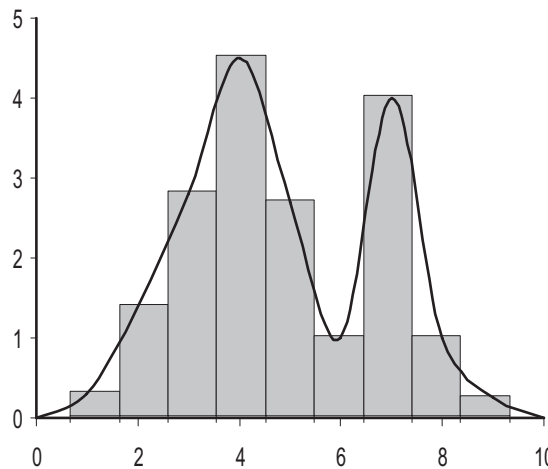


Figure 11.6 Frequency curve superimposed on a histogram

Bell Shaped distribution (Normal distribution)

There is one particular symmetric bell-shaped distribution, known as the Normal distribution. Many human clinical features are distributed normally, and the normal distribution has a very important role to play.

Characteristics of normal distribution

1. It is bell shaped curve
2. It is symmetrical in distribution; variables on either side of mean are equal in number.
3. Its maximum height is at the mean.
4. Mean= mode= median , in case of normal distribution coincide.

5. Skewness of the curve is zero.
6. It is asymptotic, in that tails never touch baseline.
7. Total area of curve is one and standard deviation is also one.
8. It has two curves. Central part is convex and when it comes down, it becomes concave on both sides.

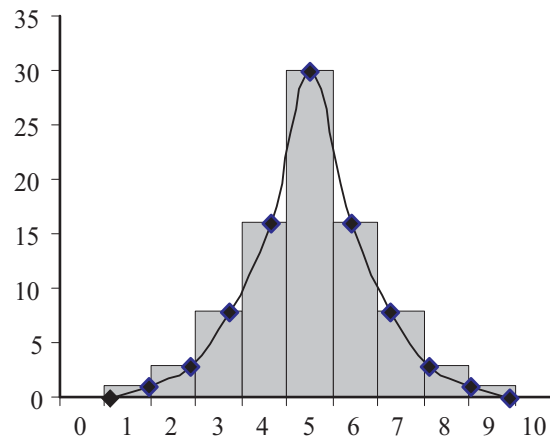


Figure 11.7 Frequency curve superimposed on a histogram

Skewness

Skewness measures asymmetry around the mean. The parameter is best interpreted as relative to the normal distribution (whose skewness equals to zero). The interpretation of the skewness is

- Skewness > 0 asymmetric tail with more values above the mean
- Skewness < 0 asymmetric tail with more values below the mean

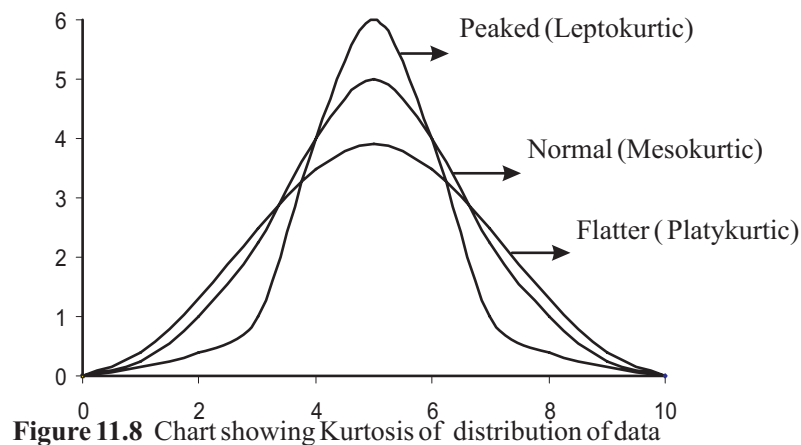
Skewed data is required to be treated using non parametric tests while normal curve data is treated using parametric tests.

Kurtosis

Kurtosis is a property associated with a frequency distribution and refers to the shape of the distribution of values regarding its relative flatness and peakedness. Compared with normal distribution, the interpretation of the kurtosis is:

- Kurtosis > 0 peaked relative to Normal distribution
- Kurtosis < 0 flat relative to Normal distribution

Here are some examples of the shapes described above.



Summary

Shapes of data:

1. Uniform distribution
2. Positively skewed distribution
3. Negatively skewed distribution
4. Symmetric or Mound-shaped distribution
5. Bimodal or Multimodal distribution

If the data is skewed it is required to be treated using non parametric tests while normal data is treated using parametric tests.

Normal distribution:

It is a bell shaped curve, symmetric and usually has Mean = Mode = Medium

Skewness

Skewness measures asymmetry around the mean.

Kurtosis

Kurtosis measures relative flatness and peakedness.

Multiple Choice Questions

1. If a distribution is skewed to the left, then it is _____.
 - a. Negatively skewed
 - b. Positively skewed
 - c. Symmetrically skewed
 - d. Symmetrical
2. If a test was generally very easy, except for a few students who had very low scores, then the distribution of scores would be _____.
 - a. positively skewed
 - b. negatively skewed
 - c. not skewed at all
 - d. normal

- ### Exercise:

- Answers:**

1. a 2. b 3. c 4. a 5. b



Chapter 12

MEASURES OF CENTRAL TENDENCY

Learning objectives

When we have finished this chapter, we should be able to:

1. Explain what a summary measure of location is, and calculate the mode, median and mean for a set of values.
2. Demonstrate the role of data type and distributional shape in choosing the most appropriate measure of central tendency.
3. Explain what a percentile is, and calculate any given percentile value.
4. Explain what a summary measure of spread is, and calculate, the range, the interquartile range and the standard deviation.
5. Estimate percentile values from an ogive.
8. Demonstrate the role of data type and distributional shape in choosing the most appropriate measure of spread.
9. Draw a boxplot and explain how it works.
10. Explain the use of Excel in estimating all measures of central tendency.

Introduction

As we have seen in the previous two chapters, we can ‘describe’ raw data by charting it, or arranging it in table form or we can examine its shape. These procedures help us to see patterns in the data. However, it is often more useful to summarise the data numerically. There are two principal features of a set of data that can be summarised with a single numeric value:

1. Measure of Location: A value around which the data has a tendency to congregate or cluster, is called a summary measure of location.

2. Measure of Dispersion: A value which measures the degree to which the data are spread out is called a summary measure of spread or dispersion.

With these two summary values we can compare different sets of data quantitatively.

Summary measures of location

A summary measure of location is a value around which most of the data values tend to congregate or centre. Let us discuss three measures of location: the mode; the median; and the mean.

1. The Mode

The mode is that category or value in the data that has the highest frequency (i.e. it occurs most often). The mode is not particularly useful with metric continuous data where no two values are

same. The other shortcoming of this measure is that there may be more than one mode in a set of data.

1. Mode for Ungrouped data

Example 12.1

Calculate mode for ungrouped data given below

6, 8, 9, 7, 12, 3, 2, 4, 8, 1, 8, 5

Solution

1. First arrange the data in ascending order

1, 2, 3, 4, 5, 6, 7, 8, 8, 8, 9, 12.

2. Mode is frequently occurring value. In this example 8 is frequently occurring and hence **mode** is **8**.

2. Mode for grouped data

The mode for grouped data can be calculated by using the formula

$$\text{Mode} = l_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times i \quad \dots 1$$

Where,

l_1 = Lower limit of modal class

f_1 = Frequency of modal class

f_0 = Frequency before the modal class

f_2 = Frequency after the modal class

i = Class interval of modal class

Example 12.2

Calculate mode for following grouped data

Table 12.1 Frequency distribution of Sales Per day

Sales volume (Class interval)	53-56	57-60	61-64	65-68	69-72	72 and above
Number of days (Frequency)	2	4	5	4	4	1

Solution

Since the largest frequency corresponds to the class interval 61-64, hence it is the modal class.

l_1 = Lower limit of modal class = 61; f_1 = Frequency of modal class = 5;

f_0 = Frequency before the modal class = 4; f_2 = Frequency after the modal class = 4

i = Class interval of modal class = 3

Formula

$$\text{Mode} = l_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times i$$

$$\text{Mode} = 61 + \left(\frac{5-4}{10-4-4} \right) \times 3 = 62.5$$

Hence, the **modal** sale is of **62.5** units.

Example 12.3

Calculate mode for following grouped data

Table 12.2 Frequency distribution for the heights of the Pharmacy students

Height (inches)	57-59	59-61	61-63	63-65	65-67	67-69
Frequency	47	23	65	51	20	08

Solution

Since the largest frequency corresponds to the class interval 61-63, hence it is the modal class.

l_1 = Lower limit of modal class = 61; f_1 = Frequency of modal class = 65;

f_0 = Frequency before the modal class = 23; f_2 = Frequency after the modal class = 51

i = Class interval of modal class = 2

Formula

$$\text{Mode} = l_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times i$$

$$\text{Mode} = 61 + \left(\frac{65 - 51}{130 - 23 - 51} \right) \times 2 = 61.5$$

Hence, the **modal** height is of **61.5** inches.

2. The Median

If we arrange the data in ascending order of size, the median is the middle value. Thus, half of the values will be equal to or less than the median value, and half will be equal to or above it. The median is thus a measure of central-ness. If we have an even number of values, the median is the average of the two values either side of the 'middle'.

An advantage of the median is that it is not much affected by skewness in the distribution, or by the presence of outliers. However, it discards a lot of information, because it ignores most of the values, apart from those in the centre of the distribution.

1. Median for ungrouped data

In this case the data is arranged in either ascending or descending order of magnitude. If the number of observations (n) is an odd number, then the median is represented by the numerical value corresponding to $(n+1)/2$ th ordered observation.

$$\text{Median} = \text{Size or value of } \left(\frac{n+1}{2} \right) \text{th observation} \quad \dots 2$$

If the number of observations (n) is an even number, then the median is defined as the arithmetic mean of the numerical values of $n/2$ th and $(n/2 + 1)$ th observations in the data array.

$$\text{Median} = \frac{\frac{n}{2} \text{th} + \left(\frac{n}{2} + 1 \right) \text{th observation}}{2} \quad \dots 3$$

Example 12.4

Calculate the median of the following data that relates to the number of patients per day in the outpatient ward in a Civil Hospital.

100, 200, 120, 170, 130, 150, 180.

Solution:

1. First arrange the data in an ascending order

100, 120, 130, 150, 170, 180, 200.

2. Since the number of observations in the data array are odd ($n=7$), the median for this data is

$$\text{Median} = \text{Size or value of } \left(\frac{n+1}{2} \right) \text{th observation}$$

$$\text{Median} = \left(\frac{7+1}{2} \right) = 4 \text{th observation}$$

4th observation in the data array = 150.

Thus the **median** number of patients examined per day in OPD in a Civil Hospital are **150**.

Example 12.5

Calculate the median of the following data that relates to the sale in lakh per month of a company in last one year. 12, 18, 15, 14, 13, 12, 20, 10, 11, 18, 19, 16

Solution:

1. First arrange the data in ascending order

10, 11, 12, 12, 13, 14, 15, 16, 18, 18, 19, 20.

2. Since the number of observations in the data array are even ($n=12$), the median for this data is

$$\text{Median} = \frac{\frac{n}{2} \text{th} + \left(\frac{n}{2} + 1 \right) \text{th}}{2}$$

$$\text{Median} = \frac{\frac{12}{2} \text{th} + \left(\frac{12}{2} + 1 \right) \text{th}}{2} = \frac{(6 \text{th value} + 7 \text{th value})}{2} = \frac{14 + 15}{2} = 14.5$$

Thus the **median** sale of company per month is **14.5** lakhs.

2. Median for grouped data (Metric continuous)

To find the median for grouped data, first we need to identify the class interval which contains the median value or $(n/2)$ th observation of the data set. To identify such class interval, we should find the cumulative frequency of each class until the class for which the cumulative frequency is equal to or greater than the value of $(n/2)$ th observation. The value of the median within that class is found by using interpolation. It is assumed that the observation values are evenly spaced over the entire class interval. The following formula is used to determine the median of grouped data:

$$\text{Median} = l_1 + \frac{(n/2) - \text{c.f.}}{f} \times i \quad \dots 4$$

Where

l_1 = Lower limit of median class.

c.f. = Cumulative frequency of the class prior to the median class.

f = Frequency of median class.

i = Class interval of median class.

Median class is the class in which $n/2^{\text{th}}$ observation lies. To use above formula, data should be continuous.

Example 12.6

A survey was conducted to determine the age (in years) of 130 pharmacists. The result of such a survey is as follows:

Table 12.3 Frequency distribution of age of Pharmacist

Age of pharmacist	20-25	25-30	30-35	35-40	40-45	45-50	50-55
Number of Pharmacist	6	13	29	48	22	8	4

Solution:

First, we should find the cumulative frequencies to locate the median class (see table 12.4).

Table 12.4 Calculations for median age of pharmacist

Age of pharmacist (in years)	No of pharmacist (f)	Cumulative frequency (cf)
20-25	06	06
25-30	13	19
30-35	29	48
35-40	48	96
40-45	22	118
45-50	08	126
50-55	04	130
n=130		

← Median class

Here the total number of observations are $n = 130$. Median is the size of $(n/2)$ th = $130/2 = 65$ th observation in the data set. This observation lies in the class interval 35-40.

l_1 = Lower limit of median class = 35

c.f. = Cumulative frequency of the class prior to the median class interval = 48

f = Frequency of median class = 48

i = Class interval of median class = 5

Formula

$$\text{Median} = l_1 + \frac{(n/2) - \text{c.f.}}{f} \times i$$

$$\text{Median} = 35 + \frac{(130/2) - 48}{48} \times 5 = 35 + \frac{17}{48} \times 5 = 35 + 1.77 = 36.77$$

Hence the **median** age of pharmacist is **36.77 years**.

Example 12.7

A survey was conducted to determine the height (in inches) of 50 pharmacists. The result of such a survey is as follows:

Table 12.5 Frequency distribution of age of Pharmacist

Height of pharmacist	44-48	48-52	52-58	58-62	62-66	66-70	70-74
Number of Pharmacist	1	2	7	8	18	12	2

Solution:

First, we should find the cumulative frequencies to locate the median class (see table 12.6).

Table 12.6 Calculations for median age of pharmacist

Height of pharmacist (in inches)	No of pharmacist (f)	Cumulative frequency (cf)	
44-48	01	01	
48-52	02	03	
52-58	07	10	
58-62	08	18	
62-66	18	36	← Median class
66-70	12	48	
70-74	02	50	
n=50			

Here the total number of observations are $n = 50$. Median is the size of $(n/2)$ th = $50/2 = 25$ th observation in the data set. This observation lies in the class interval 62-66.

l_1 = Lower limit of median class = 62

c.f.= Cumulative frequency of the class prior to the median class interval = 18

f= Frequency of median class = 36

i= Class interval of median class = 4

$$\text{Median} = l_1 + \frac{(n/2) - \text{c.f.}}{f} \times i$$

$$\text{Median} = 62 + \frac{(50/2) - 18}{36} \times 4 = 62.78$$

Hence the **median** age of pharmacist is **62.78 inches**.

3. Median for Metric Discrete data

Median for metric discrete data is calculated as follows

Median = value of (n/2)th observation

Example 12.8

The information on the number of defective components in 1000 boxes, is given below

Table 12.7 Data of number of defective components in 1000 boxes

No. of defective components	0	1	2	3	4	5	6
Number of boxes	25	306	402	200	51	10	6

Calculate the median of defective components for the whole of the production line.

Solution

For the calculation of median defective components for the whole production line the frequency table as shown in table 12.8 is required.

Table 12.8 Frequency table for calculation of median for Metric discrete data

No. of defective components	No. of Boxes (f)	Cumulative frequency (cf)
0	25	25
1	306	331
2	402	733
3	200	933
4	51	984
5	10	994
6	06	1000

n = 1000

Median = value of (n/2)th observation

= value of (1000/2) th observation

= value of 500th observation

In cumulative frequency column 500th observation comes after 331 and before 733.

Hence **median** is **2** defective components.

3. The Mean

The mean, or the arithmetic mean, is more commonly known as the average. One advantage of the mean over the median is that it uses all of the information in the data set. However, it is affected by skewness in the distribution, and by the presence of outliers in the data. This may, sometimes, produce a mean that is not very representative of the general mass of the data. Moreover, it cannot be used with ordinal data (as ordinal data are not real numbers, so they cannot be added or divided).

1. Mean for Ungrouped data

The mean of ungrouped data is calculated by adding the values of all observations and dividing the total by the number of observations.

$$\text{Mean } (\bar{X}) = \frac{\text{Sum of all observations } (\sum X)}{\text{Total number of observations } (N)} \quad \dots 5$$

Example 12.9

The following data gives weight of 20 paracetamol tablets in mg. Calculate average weight of a paracetamol tablet.

625, 617, 633, 630, 620, 631, 618, 620, 619, 632,

625, 628, 626, 624, 622, 625, 627, 631, 619, 624.

Solution:

$$\text{Mean } (\bar{X}) = \frac{\text{Sum of all observations } (\sum X)}{\text{Total number of observations } (N)} = \frac{12496}{20} = 624.8$$

Hence **mean** weight of paracetamol tablet is **624.8**.

Example 12.10

The blood serum cholesterol levels of 10 subjects are given as:

245 262 292 247 253 286 274 265 279 252

Calculate mean.

Solution:

$$\text{Mean } (\bar{X}) = \frac{\text{Sum of all observations } (\sum X)}{\text{Total number of observations } (N)} = \frac{2655}{10} = 265.5$$

Hence **mean** cholesterol levels of 10 subject is **265.5**.

2. Mean for metric continuous data (grouped data)

Mean for metric continuous grouped data can be calculated by using following formula

$$\text{Mean } (\bar{X}) = A + \frac{\sum fd}{N} \times i \quad \dots 6$$

Where

A = Assumed mean

f = frequency of ith class interval

N = Summation of all frequencies

d = (mi - A) / i, deviation from assumed mean

m = mid value of ith class interval

i = width of the class interval

Example 12.11

A company is planning to improve plant safety. For this, accident data for the last 50 weeks was compiled. These data are grouped into the frequency distribution as shown below. Calculate mean of the number of accidents per week.

Table 12.9 Number of accidents in last 50 weeks

No. of accidents	0-4	5-9	10-14	15-19	20-24
Number of weeks	5	22	13	8	2

Solution

1. Construct frequency distribution table as per given in table 12.10.

Table 12.10 Calculations for mean accidents per week

No. of Accidents class interval	No of weeks (f)	Mid value of class interval (m)	Deviation from assumed mean (d)	fd
0-4	5	02	-2	-10
5-9	22	07	-1	-22
10-14	13	12 ← A	0	0
15-19	8	17	1	8
20-24	2	22	2	4
N= 50				Σ fd = -20

2. Take the mid point of any class interval as assumed mean (A). Here we have taken the mid point of class interval of 10-14, which is 12, as assumed mean.

3. Set third column as d. That gives deviation from assumed mean, which is calculated by using formula:

(mid point of ith class interval - assumed mean) / width of class interval.

Here class interval is 5.

4. Finally set forth column fd and calculate $\sum fd$.

5. Now put the values in formula for calculation of mean for grouped data

Data:

A = Assumed mean = 12

N = Summation of all frequencies = 50

$\sum fd = -20$

i = width of the class interval = 5

Formula:

$$\text{Mean } (\bar{X}) = A + \frac{\sum fd}{N} \times i$$

$$\text{Mean } (\bar{X}) = 12 + \frac{-20}{50} \times 5 = 10$$

Hence **mean** of the number of accidents per week is **10**.

Example 12.12

The following are the weights in kg of 60 final year pharmacy students. Calculate mean weight.

Table 12.11 Weights of 60 final year students

Weight (kg)	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84
Frequency	3	10	12	15	6	7	5	2

Solution

1. Construct frequency distribution table as per given in table 12.12.

Table 12.12 Calculations for mean weight of final year students

Weight (kg)	Frequency n=60	Mid value of class interval (m)	Deviation from assumed mean (d)	fd
45-49	03	47	-3	-9
50-54	10	52	-2	-20
55-59	12	57	-1	-12
60-64	15	62 ← A	0	0
65-69	06	67	1	6
70-74	07	72	2	14
75-79	05	77	3	15
80-84	02	82	4	8
N= 60				$\sum fd = 2$

2. Take the mid point of any class interval as assumed mean (A). Here we have taken the mid point of class interval of 60-64, which is 62, as assumed mean.
3. Set third column as d. That gives deviation from assumed mean, which is calculated by using formula:

$$(\text{mid point of } i\text{th class interval} - \text{assumed mean}) / \text{width of class interval}.$$
 Here class interval is 5.
4. Finally set forth column fd and calculate $\sum fd$.
5. Now put the values in formula for calculation of mean for grouped data

Data:

A = Assumed mean = 62

N = Summation of all frequencies = 60

$\sum fd = 2$

i = width of the class interval = 5

Formula:

$$\text{Mean } (\bar{X}) = A + \frac{\sum fd}{N} \times i$$

$$\text{Mean } (\bar{X}) = 62 + \frac{2}{60} \times 5 = 62.16$$

Hence **mean** weight of final year pharmacy students is **62.16** kg.

3. Mean for Metric Discrete data (frequency data)

When observations are grouped as a frequency distribution, then mean is calculated by using following formula:

$$\text{Mean } (\bar{X}) = \frac{\sum f_i X_i}{N} \quad \dots 7$$

Where

$N = \sum f_i$

f_i = frequency with which variable X_i occurs

Example 12.13

Following data gives number of times that inhaler used in past 24h by 53 children with asthma. Find mean number of times inhaler used by children with asthma.

Table 12.13 Number of times inhaler used by children

No. of times inhaler used	0	1	2	3	4	5	6	7
Number of children	6	16	12	8	5	3	2	1

Solution:

1. Construct frequency distribution table as per given in table 12.14.
2. Set third column as $f_i X_i$.
3. Determine summation of f_i and $f_i X_i$.
4. Now put the values in formula for calculation of mean for grouped discrete data

Table 12.14 Calculation of mean for metric discrete data

Number of times inhaler used in past 24 h (X_i)	Number of children (f_i)	$f_i X_i$
0	06	0
1	16	16
2	12	24
3	08	24
4	05	20
5	03	15
6	02	12
7	01	7
$N = \sum f_i = 53$		$\sum f_i X_i = 118$

Applying formula 7, the mean is

$$\text{Mean } (\bar{X}) = \frac{\sum f_i X_i}{N} = \frac{118}{53} = 2.23$$

Hence mean number of times inhaler used by children with asthma in past 24 h was **2.23**.

Example 12.14

Calculate mean number of living children per woman from the following table.

Table 12.15 Number of living children per woman

No. of living children	0	1	2	3	4	5
Number of women	42	49	57	40	31	22

Solution:

1. Construct frequency distribution table as per given in table 12.16.
2. Set third column as $f_i X_i$.
3. Determine summation of f_i and $f_i X_i$.
4. Now put the values in formula for calculation of mean for grouped discrete data.

Table 12.16 Calculation of mean for metric discrete data

Number of living children (X_i)	Number of Women (f_i)	$f_i X_i$
0	42	0
1	49	49
2	57	114
3	40	120
4	31	124
5	22	110
$N = \sum f_i = 241$		$\sum f_i X_i = 517$

Applying formula 7, the mean is

$$\text{Mean } (\bar{X}) = \frac{\sum f_i X_i}{N} = \frac{517}{241} = 2.145$$

Hence **mean** number of living children per woman was **2.145**.

Choosing the most appropriate measure

The most appropriate measure of location for a given set of data is given below in the table 12.10. The main thing to remember is that the mean cannot be used with ordinal data (because they are not real numbers), and that the median can be used for both ordinal and metric data (particularly when the latter is skewed).

Table 12.17 A guide to choosing an appropriate measure of location

Type of variable	Summary measure of location		
	Mode	Median	Mean
Nominal	yes	no	no
Ordinal	yes	yes	no
Metric continuous	yes	yes, if skewed	yes
Metric discrete	yes	yes, if skewed	yes

4. Percentiles

The measures of central values discussed so far are averages. They locate the centre or mid point of a distribution. It may also be of interest to locate other points in the range like percentiles. They are values of a variable which divide the total observations by an imaginary line into two parts, expressed in percentages such as 10% and 90% or 25% and 75%, etc. In all, there are 99 percentiles. Percentiles are values in a series of observations arranged in ascending order of magnitude which divide the distribution into 100 equal parts. Thus, the median is 50th percentile. The 50th percentile

will have 50% observations on either side. Accordingly, 10th percentile will have 10% observations to the left and 90% to the right.

Quartiles are three different points located on the entire range of a variable- Q1, Q2 and Q3. Q1 or lower quartile will have 25% observations falling on its left and 75% on its right; Q2 or median will have 50% observations on either side and Q3 or upper quartile will have 75 % observations on its left and 25% on its right.

Quintiles are four in number and divide the distribution into 5 equal parts. So 20th percentile or first quintile will have 20% observations falling to its left and 80% to its right.

Deciles are nine in number and divide the distribution into 10 equal parts, first decile or 10th percentile will divide the distribution into 10% and 90% while 9th decile will divide into 90% and 10% .

Calculating a percentile value

Percentile value is calculated by using formula

$$P\text{th Percentile} = \frac{P}{100} (N + 1)\text{th value} \quad \dots 8$$

Where

P= percentile

N = number of values

Example 12.15

Following table records the percentage mortality in 26 ICUs. Calculate the 25th and 75th percentiles for the ICU percent mortality values.

Table 12.18 Percent mortality in 26 ICUs

ICU	1	2	3	4	5	6	7	8	9	10	11	12	13
% mortality	15.2	31.3	14.9	16.3	19.3	18.2	20.2	12.8	14.7	29.4	21.1	20.4	13.6
ICU	14	15	16	17	18	19	20	21	22	23	24	25	26
% mortality	22.4	14	14.3	22.8	26.7	18.9	13.7	17.7	27.2	19.3	16.1	13.5	11.2

Solution:

1. First, arrange values in ascending order for 26 ICUs as shown below

ICU	1	2	3	4	5	6	7	8	9	10	11	12	13
% mortality	11.2	12.8	13.5	13.6	13.7	14	14.3	14.7	14.9	15.2	16.3	16.1	17.7
ICU	14	15	16	17	18	19	20	21	22	23	24	25	26
% mortality	18.2	18.9	19.3	19.3	20.2	20.4	21.1	22.4	22.8	26.7	27.2	29.4	31.3

2. Now, the 25th percentile can be calculated by using formula

$$P^{\text{th}} \text{ Percentile} = \frac{P}{100} (N + 1)^{\text{th}} \text{ value}$$

$$P = 25^{\text{th}} \text{ percentile} = 25$$

$$N = \text{no. of ICU} = 26$$

$$25^{\text{th}} \text{ Percentile} = \frac{25}{100} (26 + 1) = 0.25 \times 27^{\text{th}} \text{ value} = 6.75^{\text{th}} \text{ value}$$

The 6th value is 14 % while the 7th value is 14.3%,

A difference is of 0.3,

The 25th percentile is 14% + 0.75 of 0.3,

$$14\% + 0.75 \times 0.3\% = 14\% + 0.225 = 14.225$$

So, the **25th percentile** for the ICU percent mortality is **14.225**.

3. Now, the 75th percentile can also be calculated using same formula, $P = 75$ and $N = 26$.

$$75^{\text{th}} \text{ Percentile} = \frac{75}{100} (26 + 1) = 0.75 \times 27^{\text{th}} \text{ value} = 20.25^{\text{th}} \text{ value}$$

The 20th value is 21.1 % while the 21st value is 22.4%, a differences of 1.3,

The 75th percentile is 21.1% + 0.25 of 1.3 = 21.1% + 0.25 x 1.3% = 21.1% + 0.325 = 21.425

So, the **75th percentile** for the ICU percent mortality is **21.425**.

Summary measures of spread

There are three main measures of spread or dispersion in common use. Here also, the type of data influences the choice of an appropriate measure.

1. The Range

The range is the distance from the smallest value to the largest. The range is not affected by skewness, but is sensitive to the addition or removal of an outlier value.

$$\text{Range} = \text{Lowest value to Highest value.} \quad \dots 9$$

Example 12.16

Weight of six paracetamol tablets in mg are given below. Find weight range for paracetamol tablet.

Data: 625, 612, 615, 632, 628, 618

Solution:

1. Arrange data in ascending order

612, 615, 618, 625, 628, 632

2. Lowest value = 612 mg

3. Highest value = 632 mg

Range is 612 to 632 mg.

2. The interquartile range (iqr)

One solution to the problem of the sensitivity of the range to extreme value (outliers) is to chop a quarter (25 per cent) of the values off both ends of the distribution, which removes any troublesome outliers, and then measure the range of the remaining values. This distance is called the interquartile range, or iqr. The interquartile range is not affected either by outliers or skewness, but it does not use all of the information in the data set since it ignores the bottom and top quarter of values.

Calculating interquartile range

To calculate the interquartile range, first we need to determine two values:

1. The value which cuts off the bottom 25th percentile of values; this is known as the first quartile and denoted as Q1.
2. The value which cuts off the top 75th percentile of values, known as the third quartile and denoted as Q3.

The interquartile range is then written as (Q1 to Q3).

Example 12.17

Calculate the iqr for the ICU percentage mortality values in table 12.19.

Table 12.19 Percent mortality in 26 ICUs

ICU	1	2	3	4	5	6	7	8	9	10	11	12	13
% mortality	15.2	31.3	14.9	16.3	19.3	18.2	20.2	12.8	14.7	29.4	21.1	20.4	13.6
ICU	14	15	16	17	18	19	20	21	22	23	24	25	26
% mortality	22.4	14	14.3	22.8	26.7	18.9	13.7	17.7	27.2	19.3	16.1	13.5	11.2

Solution

1. Arrange the data in ascending order
2. Calculate 25th percentile as a Q1
3. Calculate 75th percentile as a Q3
4. We have already calculated the 25th and 75th percentiles before in example 12.15

$$Q1 = 14.225\%$$

$$Q3 = 21.425\%$$

Therefore interquartile range = (14.225 to 21.425) %.

Estimating the median and interquartile range from Ogive

We can also estimate interquartile range from ogive. If we draw horizontal lines from the values 25 per cent, 50 per cent and 75 per cent on the y axis, to the ogive, and then down to the x axis, the points of intersection on the x axis approximately gives values for Q1, Q2 (the median), and Q3.

Example 12.18

Estimate iqr for the data of weight in kg of 60 final year pharmacy students given in Table 10.6.

Table 12.20 Weights of 60 final year students

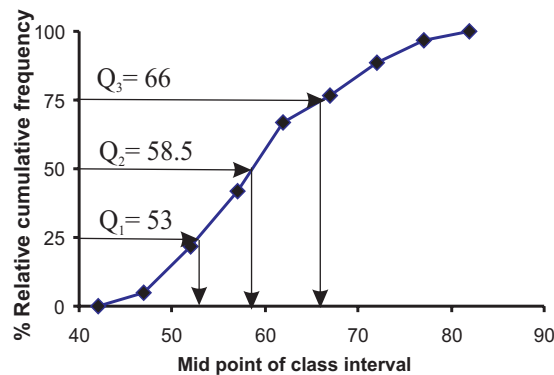
Weight (kg)	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84
Frequency	3	10	12	15	6	7	5	2

Solution

We have already drawn ogive of this example.

Ogive is drawn by plotting % cumulative frequency on Y axis against midpoint of class interval on X axis.

Now, draw perpendicular to ogive at 25th percentile (Q_1), 50th percentile (Q_2) and 75th percentile (Q_3) as shown in figure below.

**Figure 12.1** % cumulative frequency curve of weight to estimate median and iqr

The weight values corresponding to these perpendiculars are taken as Q_1 , Q_2 and Q_3 and here they are $Q_1=53$, $Q_2=58.5$ and $Q_3=66$

Interquartile range (IQR) = Q_1 to Q_3

Interquartile range (IQR) = 53 to 66

The interquartile range for the data of weights of final year students is 53 to 66 kg.

The Boxplot

Now that we have discussed the median and interquartile range, we can now better understand the boxplot which we had seen in chapter 11. Boxplots provide a graphical summary of the three quartile values, the minimum and maximum values, and any outliers. They are usually plotted with value on the vertical axis. Like the pie chart, the boxplot can only represent one variable at a time, but a number of boxplots can be set alongside each other.

Let's understand the boxplot,

1. The bottom end of the lower 'whisker' (the line sticking out of the bottom of the box), corresponds to the minimum value.
2. The bottom of the box is the 1st quartile value, Q1.
3. The line across the inside of the box, is the median, Q2. The median line will always not be at the centre. The more asymmetric (skewed) the distributional shape, the median line will be further away from the middle of the box. If it is closer to the top of the box, it indicates negative skew while if it is closer to the bottom of the box, it indicates positive skew.
4. The top of the box is the third quartile Q3.
5. The top end of the upper whisker is the 'maximum'. This is the maximum value that can be considered still to be part of the general mass of the data.
6. There is one outlier. This is, the actual maximum value in the data.

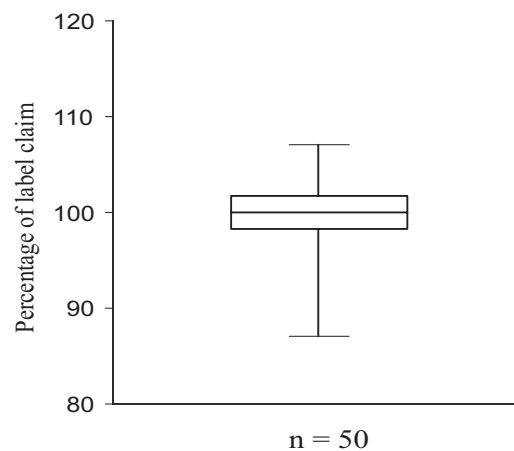


Figure 12.1 Boxplot

Example 12.21

Sketch the box plot for the percentage mortality in ICUs shown in table 12.10. We have already calculated Q1 and Q3 for same.

Solution:

Calculation of Q2, Median

The 50th percentile can be calculated using formula

$$\text{Pth Percentile} = \frac{P}{100} (N + 1)\text{th value}$$

$$P = 50^{\text{th}} \text{ percentile} = 50; N = 26$$

$$50\text{th Percentile} = \frac{50}{100} (26 + 1) = 0.5 \times 27\text{th value} = 13.5\text{th value}$$

The 13th value is 17.7% while the 14th value is 18.2%. A difference of 0.5.

The 50th percentile is 17.7% + 0.5 of 0.5 = 17.7% + 0.5 × 0.5% = 17.7% + 0.25 = 17.95

So, the 25th percentile for the ICU percent mortality is 14.225.

So, the 50th percentile for the ICU percent mortality is 17.95.

So, the 75th percentile for the ICU percent mortality is 21.425.

Now, the box plot can be sketched as under

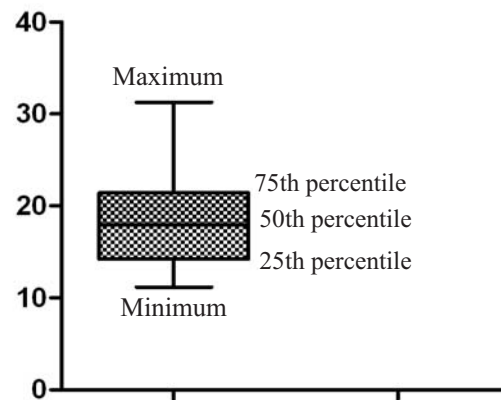


Figure 12.2. Box plot for percentage mortality in ICUs

3. Standard Deviation and Variance

The limitation of the interquartile range as a summary measure of spread is that, it is not using all the information in the data, since it omits the top and bottom quarter of values. An alternative approach we can use is to measure the mean (average) distance of all the data values from the overall mean of all of the values. The smaller this mean distance is, the narrower the spread of values will be, and vice versa. This approach is known as the standard deviation, or s.d.

Standard deviation can be calculated by using following steps:

1. Subtract the mean of the sample from each of the n sample values in the sample, to give the difference values.
2. Square each of these differences.
3. Add these squared values together (called the sum of squares).
4. Divide the sum of squares by $(N - 1)$; i.e. divide by 1 less than the sample size.
5. Take the square root. This is the standard deviation

Variance is the square of standard deviation and is used instead of standard deviation.

Example 12.23

A microbiologist found 8, 10, 19, 12, 6 and 5 *E. coli* in six culture. Calculate standard

deviation, s and variance, s^2 .

Solution

1. Let us calculate the mean

$$\text{Mean, } \bar{X} = \frac{8+10+19+12+6+5}{6} = 10$$

2. Now, construct the table as shown below:

x	$(x - \bar{x})$	$(x - \bar{x})^2$
08	-2	4
10	0	0
19	9	81
12	2	04
06	-4	16
05	-5	25
		$\Sigma(x - \bar{x})^2 = 130$

3. Let us use $\Sigma(x - \bar{x})^2$ value in the formula

Formula

$$\text{Standard deviation (s)} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}} \quad \dots 10$$

$$\text{Standard deviation (s)} = \sqrt{\frac{130}{6-1}} = \sqrt{26} = 5.1$$

4. Thus, standard deviation, s was found to be **5.1** while variance, s^2 was found to be **26**.

However, the calculations using this formula can be tedious when the numbers and mean are in decimals. So, the computing formula used is as follows

Formula

$$\text{Standard deviation (s)} = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N - 1}} \quad \dots 11$$

Let us calculate standard deviation using this formula

x	x^2
08	64
10	100
19	361
12	144
06	36
05	25
$\Sigma x = 60$	$\Sigma x^2 = 730$

Hence: $N = 6$, $\sum X = 60$, $\sum X^2 = 730$

$$\text{Standard deviation (s)} = \sqrt{\frac{730 - \frac{(60)^2}{6}}{6 - 1}} = 5.1$$

The, standard deviation, s by this method is also **5.1** and variance, s^2 is **26**.

1. Standard deviation (s) in ungrouped data

For calculation of standard deviation in ungrouped data following formula is used, where assumed mean method is used so that squaring of large values can be avoided.

$$\text{Standard deviation (s)} = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}} \quad \dots 12$$

Where,

N = number of observations

X = deviation from assumed mean = $x - A$.

A = Assumed mean

x = given observations

Example 12.23

Find standard deviation of incubation period of smallpox in 9 patients where it was found to be 14, 13, 11, 15, 10, 7, 9, 12, and 10.

Solution

1. Construct calculation table as per given in table no 12.21.
2. Set any assumed mean from set of observations (x).
3. Set second column as deviation from assumed mean i.e. $X = x - A$.
4. Set third column as square of deviation i.e. X^2
5. Determine summation of deviation from assumed mean (X) and square of deviation (X^2).
6. Finally put the values in equation 10 so as to get standard deviation for ungrouped data.

Table 12.21 Calculation of standard deviation for ungrouped data

Incubation period of smallpox (x)	Deviation from assumed mean (X= x - 11)	Square of deviation X ²
14	3	9
13	2	4
11 ← A	0	0
15	4	16
10	-1	1
7	-4	16
9	-2	4
12	1	1
10	-1	1
$\Sigma X = 2$		$\Sigma X^2 = 52$

Data

$$N = 9$$

Summation of deviation from assumed mean, $\Sigma X = 2$

Summation of square of deviation, $\Sigma X^2 = 52$

Formula

$$\text{Standard deviation (s)} = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N - 1}}$$

$$\text{Standard deviation (s)} = \sqrt{\frac{52 - \frac{2^2}{9}}{9 - 1}} = \sqrt{\frac{51.56}{8}} = \sqrt{6.443} = 2.54$$

Therefore, **standard deviation** of incubation period of smallpox in 9 patients was **2.54**.

Example 12.24

The following data gives weight of 10 paracetamol tablets in mg. Find standard deviation.

625 617 633 630 620 631 618 620 619 632

Solution

1. Construct calculation table as per given in table no 12.22.
2. Set any assumed mean from set of observations (x).
3. Set second column as deviation from assumed mean i.e. $X = x - A$.
4. Set third column as square of deviation i.e. X^2
5. Determine summation of deviation from assumed mean (ΣX) and square of deviation (ΣX^2).
6. Finally put the values in equation 10 so as to get standard deviation for ungrouped data.

Table 12.22 Calculation of standard deviation for ungrouped data

Weight of Paracetamol tablets (x)	Deviation from assumed mean (X= x - 625)	Square of deviation X ²
625 ← A	0	0
617	-8	64
633	8	64
630	5	25
620	-5	25
631	6	36
618	7	49
620	-5	25
619	-6	36
632	7	49
$\Sigma X = -5$		$\Sigma X^2 = 373$

Data

N = 10

Summation of deviation from assumed mean, $\Sigma X = -5$

Summation of square of deviation, $\Sigma X^2 = 373$

Formula

$$\text{Standard deviation (s)} = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N - 1}}$$

$$\text{Standard deviation (s)} = \sqrt{\frac{373 - \frac{(-5)^2}{10}}{10 - 1}} = 6.416$$

Therefore, **standard deviation** of weight of 10 paracetamol tablets was **6.416**.

2. Standard deviation (s) in grouped data

Standard deviation in grouped data is calculated by using formula:

$$\text{Standard deviation (s)} = \sqrt{\frac{\Sigma fd^2 - \frac{(\Sigma fd)^2}{N}}{N - 1}} \times i \quad \dots 13$$

Where,

f = frequency

N = summation of frequency, $f = \Sigma f$

$$d = (m.p. - A)/i$$

m.p. = mid point

A = Assumed mean

i = class interval

Example 12.25

Calculate SD of following data

Table 12.23 IQ of 50 students

IQ	0-20	20-40	40-60	60-80	80-100	100-120	120-140	140-160
Number of students	3	4	3	4	13	12	8	3

Answer:

1. Construct calculation table as per given in table no 12.24.
2. Construct third column as mid value of class interval.
3. Set any assumed mean (A) from set of mid points (m.p.).
4. Set forth column as deviation from assumed mean i.e. $d = (m.p. - A)/i$.
5. Determine fd and fd^2 .
6. Determine summation of f , fd and fd^2 .
7. Finally put the values in equation 11 so as to get standard deviation for grouped data.

Table 12.24 Calculation of SD in grouped data

IQ	No of Students (f)	Mid value of class interval (m.p.)	d (mp - A)/i	d ²	f d	f d ²
0-20	3	10	-4	16	-12	48
20-40	4	30	-3	9	-12	36
40-60	3	50	-2	4	-6	12
60-80	4	70	-1	1	-4	4
80-100	13	90 ← A	0	0	0	0
100-120	12	110	1	1	12	12
120-140	8	130	2	4	16	32
140-160	3	150	3	9	9	9

$$N = \Sigma f = 50$$

$$\Sigma fd = 3 \quad \Sigma fd^2 = 171$$

Data:

Class interval, $i = 20$;

Assumed mean, $A = 90$;

$$N = \Sigma f = 50;$$

$$\Sigma fd = 3;$$

$$\Sigma fd^2 = 171$$

Formula:

$$\text{Standard deviation (s)} = \sqrt{\frac{\Sigma fd^2 - \frac{(\Sigma fd)^2}{N}}{N-1}} \times i$$

Solution: Now, putting the values in above formula, we get

$$\text{Standard deviation (s)} = \sqrt{\frac{171 - \frac{(3)^2}{50}}{50-1}} \times 20 = \sqrt{3.48} \times 20 = 37.2$$

Therefore **standard deviation** of given IQ data is 37.2.

Example 12.26

Calculate SD of following data of intelligence quotient (IQ) of 50 students.

Table 12.25 Given data

Classes	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Frequency	5	10	15	20	25	12	8	4	1

Answer:

1. Construct calculation table as per given in table no 12.26.
2. Construct third column as mid value of class interval.
3. Set any assumed mean (A) from set of mid points (m.p.).
4. Set forth column as deviation from assumed mean i.e. $d = (\text{m.p.} - A)/i$.
5. Determine fd and fd^2 .
6. Determine summation of f , fd and fd^2 .
7. Finally put the values in equation 11 so as to get standard deviation for grouped data.

Table 12.26 Calculation of SD in grouped data

Classes	Frequency (f)	Mid value of class interval (m.p.)	d (mp - A)/i	d ²	f d	f d ²
0-10	5	5	-4	16	-20	80
10-20	10	15	-3	9	-30	90
20-30	15	25	-2	4	-30	60
30-40	20	35	-1	1	-20	20
40-50	25	45 ← A	0	0	0	0
50-60	12	55	1	1	12	12
60-70	8	65	2	4	16	32
70-80	4	75	3	9	12	36
80-90	1	85	4	16	4	16
N = $\Sigma f = 100$					$\Sigma fd = -56$	$\Sigma fd^2 = 346$

Data:

Class interval, $i = 10$;

Assumed mean, $A = 45$;

$N = \sum f = 100$;

$\sum fd = -56$;

$\sum fd^2 = 346$

Formula:

$$\text{Standard deviation (s)} = \sqrt{\frac{\sum fd^2 - \frac{(\sum fd)^2}{N}}{N - 1}} \times i$$

Solution:

Now, putting the values in above formula, we get

$$\text{Standard deviation (s)} = \sqrt{\frac{346 - \frac{(-56)^2}{100}}{100 - 1}} \times 10 = 17.8$$

Therefore **standard deviation** of given data is **17.8**.

3. Standard deviation in metric discrete data

Standard deviation for metric discrete data is given by the formula:

$$\text{Standard deviation (s)} = \sqrt{\frac{\sum f_i X_i^2 - \frac{(\sum f_i X_i)^2}{N}}{N - 1}} \quad \dots 14$$

Where,

f_i = frequency with which variable X_i occurs

$N = \sum f_i$

Example 12.27

Calculate standard deviation of Example 12.13

No. of times inhaler used	0	1	2	3	4	5	6	7
Number of children	6	16	12	8	5	3	2	1

Solution

1. Construct the frequency table as shown below:

Number of times inhaler used in past 24 h (X_i)	Number of children (f_i)	$f_i X_i$	X_i^2	$f_i X_i^2$
0	06	0	0	0
1	16	16	1	16
2	12	24	4	48
3	08	24	9	72
4	05	20	16	80
5	03	15	25	75
6	02	12	36	72
7	01	7	49	49
$N = \sum f_i = 53$		$\sum f_i X_i = 118$	$\sum f_i X_i^2 = 412$	

Data

$$N = \sum f_i = 53$$

$$\sum f_i X_i = 118$$

$$\sum f_i X_i^2 = 412$$

Formula

$$\text{Standard deviation (s)} = \sqrt{\frac{\sum f_i X_i^2 - \frac{(\sum f_i X_i)^2}{N}}{N - 1}}$$

2. Now put the values in the formula

$$\text{Standard deviation (s)} = \sqrt{\frac{412 - \frac{(118)^2}{53}}{53 - 1}} = \sqrt{\frac{412 - 262.71}{53 - 1}} = 1.72$$

So, the **standard deviation** was found to be **1.72**.

Measurement related to sample standard deviation

The variability of data may often be better described as a relative variation rather than as an absolute variation (i.e., the standard deviation). This can be accomplished by calculating the coefficient of variation (CV) that is the ratio of the standard deviation to the mean.

$$\text{Coefficient of variation (CV)} = \frac{SD}{\text{mean}} \quad \dots 15$$

The CV is usually expressed as a percentage (relative standard deviation or RSD) and can be useful in many instances because it places variability in perspective to the distribution center.

$$\text{Relative Standard Deviation} = CV \times 100 \quad \dots 16$$

Example 12.28

In two series of adults aged 21 years and children 3 months old, following values were obtained for the height. Find which series shows greater variation?

Data	Persons	Mean height	SD
	Adults	160 cm	10 cm
	Children	60 cm	5 cm

Formula:

$$CV = SD/\text{mean}$$

$$RSD = CV \times 100$$

Solution:

Put the values in equation

Adults

$$CV \text{ of adults} = 10/160 = 0.0625$$

$$RSD = 0.0625 \times 100 = 6.25 \%$$

Children

$$CV \text{ of children} = 5/60 = 0.0833$$

$$RSD = 0.0833 \times 100 = 8.33 \%$$

Thus, heights in children show greater variation than adult.

Example 12.29

Chest circumference in cm of 10 malnourished children and normal children aged one year, are given below. Find which series shows greater variation?

Data	Children	Mean height	SD
	Malnourished	48.5	4.53
	Normal	34.8	4.57

Formula:

$$CV = SD/\text{mean}$$

$$RSD = CV \times 100$$

Solution:

Put the values in equation

Malnourished

$$CV = 4.53/48.5 = 0.09$$

$$RSD = 0.093 \times 100 = 9.3 \%$$

Normal

$$CV = 4.57/34.8 = 0.131$$

$$RSD = 0.131 \times 100 = 13.1 \%$$

Thus, Normal children show greater variation than malnourished children.

Choosing an appropriate measure of spread**Table 12.27 Choosing an appropriate measure of spread**

Type of variable	Summary measure of spread		
	Range	Interquartile range	Standard deviation
Nominal	no	no	no
Ordinal	yes	yes	no
Metric	yes	yes, if skewed	yes

Use of Excel in Measures of Central Tendency

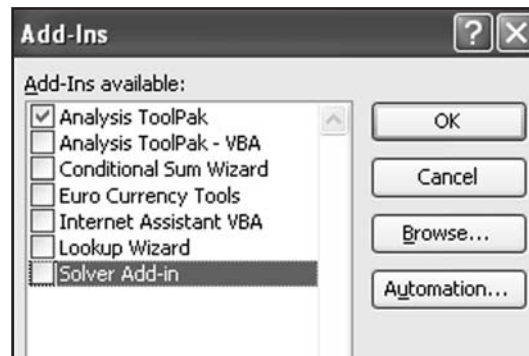
We can easily calculate both measures of central tendency i.e measures of location and measures of dispersion in excel where we need not have to worry about any formulas. Excel will readily calculate these values for us. So, lets see how we can use excel for calculating measures of central tendency.

Installing the Analysis ToolPak in Excel

Excel includes a number of add-in tools to assist with a number of data handling, reporting and analysis functions. So, first we will install this analysis toolpak as given below:

1. Open New workbook from MS-Excel and click to Tools menu from Menu bar.
2. Instantly, it will display pull-down menus.
3. Select on Add-Ins option.
4. In the Manage box, select Add-ins- Analysis ToolPak as shown in figure.
6. Then, click on OK button.

This unpacks the chosen Excel Add-in tools and makes them available for use. If the Add-in list is empty, there may have been a limited installation of MS Office. In this case, the MS Office CDs will be needed to install these Add-in tools.

**Figure 12.3 Window of Add-Ins**

Use of Excel in Descriptive Statistics

The Data Analysis ToolPak has a Descriptive Statistics tool that provides us with an easy way to calculate summary statistics for a set of sample data. Summary statistics includes Mean, Standard Error, Median, Mode, Standard Deviation, Variance, Kurtosis, Skewness, Range, Minimum, Maximum, Sum, and Count. This tool eliminates the need to type individual functions to find each of these results. Excel includes elaborate and customizable toolbars. We can follow the below steps:

Step 1. Open New Excel file. In sheet 1 put labels in first row (i.e A1, B1) and enter the data below these labels.

Step 2. Select the Tools menu from menu bar. Then, it will display pull-down menus.

Step 3. Click on the data analysis option. Instantly, dialog box will appear.

Step 4. Choose Descriptive Statistics from Analysis Tools list of Data Analysis dialog box. Lastly, click on OK button. This will display the Descriptive Statistics dialog box on the screen.

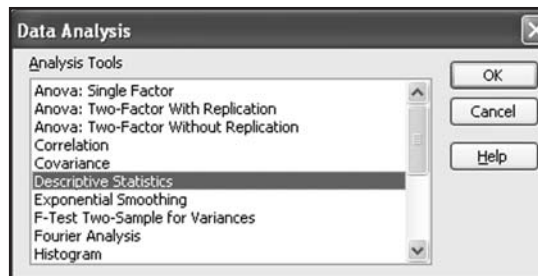


Figure 12.4 Window of Data Analysis

Step 4. When the dialog box appears; Enter range of data or select range (along with labels) in the Input Range box. Tick mark on Labels in first row and summary statistics of Output options, as shown below:

Note: Ensure that labels are given in first row.

Step 5. Ignore remaining settings and click on Ok button.

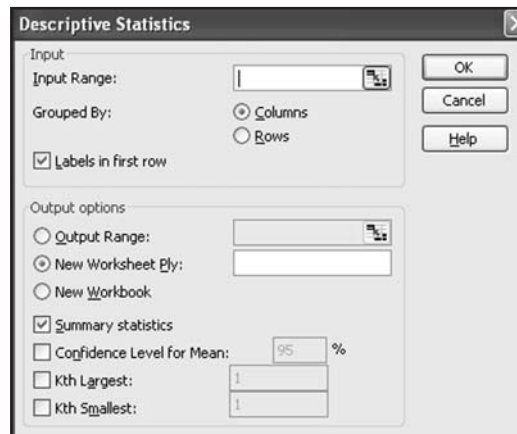


Figure 12.5 Window of Descriptive statistics

Step 6. We will get result of descriptive statistics as shown below.

Column1
Mean
Standard Error
Median
Mode
Standard Deviation
Sample Variance
Kurtosis
Skewness
Range
Minimum
Maximum
Sum
Count

Summary

The mean, median and mode, are common measures of location which represent either a typical or representative score and/or a value about which the data tend to center.

Mode

The mode (denoted by M) represents the most frequently occurring score. When more than two scores occur with the greatest frequency, the data set is said to be multimodal.

Mode for grouped data

$$\text{Mode} = l_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times i$$

Median

The median of a set of scores represents the middle value (50th percentile) when the scores are arranged as an array in order of increasing (or decreasing) magnitude.

Median for ungrouped data

Odd number

$$\text{Median} = \text{Size or value of } \left(\frac{n+1}{2} \right) \text{th observation}$$

Even number

$$\text{Median} = \text{value of } \frac{\frac{n}{2} \text{th} + \left(\frac{n}{2} + 1 \right) \text{th}}{2} \text{ observation}$$

Median for grouped data

$$\text{Median} = l_1 + \frac{(n/2) - \text{c.f.}}{f} \times i$$

Median for Metric Discrete data

$$\text{Median} = \text{value of } (n/2) \text{th observation}$$

Mean

Mean represents the most appropriate measure of central tendency for continuous-type data. It is obtained by adding all of the scores and dividing this sum by the number of scores.

Mean for ungrouped data

$$\text{Mean } (\bar{X}) = \frac{\text{Sum of all observations } (\sum X)}{\text{Total number of observations } (N)}$$

Mean for grouped data

$$\text{Mean } (\bar{X}) = A + \frac{\sum fd}{N} \times i$$

Mean for Metric Discrete data

$$\text{Mean } (\bar{X}) = \frac{\sum f_i X_i}{N}$$

Percentile

Percentiles are values in a series of observations arranged in ascending order of magnitude which divide the distribution into 100 equal parts.

$$\text{Pth Percentile} = \frac{P}{100} (N + 1) \text{th value}$$

The range, interquartile range, standard deviation are common measures of spread and represents the extent of spread of data

Range

The range is the distance from the smallest value to the largest.

$$\text{Range} = \text{Lowest value to Highest value.}$$

The interquartile range

The interquartile range is Q1 to Q3

Q1 is the value which cuts off the bottom 25th percent of values and is known as the first quartile (25th percentile), while Q3 cut off top 25th percent of values and is known as third quartile (75th percentile).

Standard Deviation (SD)

The standard deviation measures the variation of scores about the mean (average) score, and can be defined as the “root mean squared deviation.” Variance is square of standard deviation.

Standard deviation of ungrouped data

$$\text{Standard deviation } (s) = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}}$$

Standard deviation of grouped data

$$\text{Standard deviation } (s) = \sqrt{\frac{\sum fd^2 - \frac{(\sum fd)^2}{N}}{N - 1}} \times i$$

Standard deviation for metric discrete data

$$\text{Standard deviation (s)} = \sqrt{\frac{\sum f_i X_i^2 - \frac{(\sum f_i X_i)^2}{N}}{N - 1}}$$

Measurement related to sample standard deviation

$$\text{Coefficient of variation (CV)} = \frac{\text{SD}}{\text{mean}}$$

$$\text{Relative Standard Deviation} = \text{CV} \times 100$$

Choosing an appropriate measure of central tendency

Type of variable	Summary measures of location			Summary measures of spread		
	Mode	Median	Mean	Range	Iqr	SD
Nominal	yes	no	no	no	no	no
Ordinal	yes	yes	no	yes	yes	no
Metric	yes	yes, if skewed	yes	yes	yes, if skewed	yes

Multiple choice questions

- The range of a sample gives an indication of the _____.
 - way in which the values cluster about a particular point
 - number of observations bearing the same value
 - maximum variation in the sample
 - degree to which the mean value differs from its expected value.
- The observation which occurs most frequently in a sample is the _____.
 - median
 - mean deviation
 - standard deviation
 - mode
- What is the median of the sample 5, 5, 11, 9, 8, 5, 8?
 - 5
 - 6
 - 8
 - 9
- What is the median of the following set of scores? 18, 6, 12, 10, 14?
 - 10
 - 14
 - 18
 - 12
- All of the following are measures of central location except _____.
 - range
 - arithmetic mean
 - median
 - mode
- The measure of central location that has half of the observations below it and half of the observations above it is the _____.
 - arithmetic mean
 - geometric mean
 - median
 - mode
- The most commonly used measure of central location is the _____.
 - mean
 - range
 - median
 - mode

8. All of the following are measures of dispersion except _____.
a. interquartile range b. percentile c. variance d. standard deviation
9. The _____ is the value we calculate when we want the arithmetic average.
a. Mean b. Median c. Mode d. All of the above
10. The _____ is often the preferred measure of central tendency if the data are severely skewed.
a. Mean b. Median c. Mode d. Range
11. Which of the following is the formula for range?
a. $H + L$ b. $L \times H$ c. $L - H$ d. $H - L$
12. Why are variance and standard deviation the most popular measures of variability?
a. They are the most stable and are foundations for more advanced statistical analysis
b. They are the most simple to calculate with large data sets
c. They provide nominally scaled data
d. None of the above
13. Which of the following is NOT a measure of variability?
a. Median b. Variance c. Standard deviation d. Range
14. Which range characterizes the interquartile range?
a. From 5th percentile to 95th percentile
b. From 10th percentile to 90th percentile
c. From 25th percentile to 75th percentile
d. From 1 standard deviation below the mean to 1 standard deviation above the mean
15. The measure of dispersion most commonly used in conjunction with the arithmetic mean is the:
a. interquartile range b. range c. standard deviation d. variance

Exercise

1. Amount of 10 tablets of Atenolol were determined for its quality control. Calculate the mode of the observed amounts.

Data: 20, 22, 18, 24, 24, 24, 18, 19, 18, 18.

2. Calculate mode for following data

Data: 8, 5, 10, 10, 9, 6, 8, 8, 10, 8, 6, 8, 10, 10, 7.

3. A novel antidiabetic drug is developed by researcher. The reduction in glucose in 10 patients after 2 hours administration of drug is recorded. Calculate the mean reduction in blood glucose levels in the 10 patients.

Data: 20, 15, 22, 18, 25, 17, 23, 27, 19, 21.

4. The weights of 20 tablets, removed from a batch for quality control purpose are given below. Calculate the mean and median values of the tablet weights.

Data: 250, 252, 255, 245, 251, 260, 258, 256, 248, 275,
268, 240, 257, 262, 270, 280, 266, 279, 258, 265.

5. Serum bilirubin levels are measured for 10 cirrhosis patients as given below. Determine the range.

Data: 1, 8, 5, 7, 10, 2, 15, 12, 3, 9.

6. Drug content of injection were determined using ultraviolet spectroscopy. Results are reported in mg/ml. Calculate mean and standard deviation of the drug content of injection.

Data: 50.3, 48.1, 48.9, 51.5, 52.5, 50.6, 49.5, 53, 47.5, 50.

7. Time taken for dissolution of 50% of the original mass of drug from 10 tablets are summarized below. Determine the mean and standard deviation of the time required for the release of 50% of the original mass of drug.

$t_{50\%}$ (h): 24.7, 20.1, 25.3, 22, 22.5, 28.6, 21.6, 20.4, 23.5, 25.1.

8. Calculate mean and standard deviation of 15 readings in the preliminary study of urinary lead concentrations, micro mol/24h.

Data: 0.1, 0.4, 0.6, 0.8, 1.1, 1.2, 1.3, 1.5, 1.7, 1.9, 1.9, 2.0, 2.2, 2.6, 3.2

9. Results of new therapeutic agent with 50 mg/tablet administered to six healthy volunteers are listed below. Report median and mean.

Volunteer number	1	2	3	4	5	6
C_{\max} mg/l	60	71	111	46	81	96

10. Calculate median and mean of following assay data of 30 tetracycline capsules.

Assay result	245	246	247	248	249	250	251	252	253	254
Frequency	1	1	2	3	4	7	5	3	2	2

11. Calculate mode, median, mean and standard deviation for the following data.

Age in years	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No of persons	4	6	10	20	10	6	4

12. In addition to studying the lead concentration in the urine of 140 children, the paediatrician asked how often each of them had been examined by a doctor during the year 2010. Calculate mean, median and standard deviation.

No of Visits to doctor	0	1	2	3	4	5	6	7
No of Children	2	8	27	45	38	15	4	1

13. Age at Death (in days) of 78 cases of sudden infant death syndrome (SIDS) are given in following table.

Age in days	1-30	31-60	61-90	91-120	120-150	151- 180	181-210	211-240	241-270
No. of deaths	6	13	23	18	7	5	3	2	1

Determine Mean, median and standard deviation.

14. Ages of Patients Diagnosed with Multiple Sclerosis are given in following table.

Age in days	20-29	30-39	40-49	50-59	60-69	70-79	80-89
No. of deaths	4	44	124	124	48	25	4

Determine Mean and standard deviation.

15. The elimination half-lives of two synthetic steroids have been determined using two groups, each containing 15 volunteers. The results are shown below, with the values ranked from lowest to highest for each steroid. Find the median, range and quartiles.

Steroid 1

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Half life (h)	3.9	4.0	5.4	6.4	6.5	7.2	7.8	8.6	9.2	9.3	10	10.6	11.1	15.8

Steroid 2

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Half life (h)	4.4	4.5	5.5	5.8	5.9	6.1	6.6	7.2	7.2	7.3	7.8	8.5	8.6	9.1

16. From the 140 children whose urinary concentration of lead were investigated, 40 were chosen who were aged at least 1 year but under 5 years. The following concentrations of copper were found. Find the median, range and quartiles.

Data 0.70, 0.45, 0.72, 0.30, 1.16, 0.69, 0.83, 0.74, 1.24, 0.77,
 0.65, 0.76, 0.42, 0.94, 0.36, 0.98, 0.64, 0.90, 0.63, 0.55,
 0.78, 0.10, 0.52, 0.42, 0.58, 0.62, 1.12, 0.86, 0.74, 1.04,
 0.65, 0.66, 0.81, 0.48, 0.85, 0.75, 0.73, 0.50, 0.34, 0.88

17. Erythromycin contents of 10 tablets from an Alpha and a Bravo tableting machine was determined and given below. Determine coefficient of variation and relative standard deviation. Give your comment on it.

	Mean	SD
Alpha machine	248.7	8.72
Bravo machine	251.1	3.78

18. Fifteen patients were provided with their drugs in a child-proof container of a design that they had not previously experienced. A note was taken of the time it took each patient to get the container open for the first time. Determine median and quartile. The results are shown below

Times taken to open a child-proof container (s)

2.2 3 3.1 3.2 3.4 3.9 4 4.1 4.2 4.5 5.1 10.7 12.2 17.9 24.8

19. The percentage of ideal body weight was determined for 18 randomly selected insulin-dependent diabetics. The outcomes (%) are

107 119 99 114 120 104 124 88 114
116 101 121 152 125 100 114 95 117

Determine mean, CV and RSD.

20. Calculate all the following measures of central tendency for data given below:

9.4 7 7.6 6.3 6.7 8.6 6.8 10.6 8.9 9.4
a. Mean b. Median c. 25th percentile
d. 75th percentile e. 10th percentile f. Range
g. Iqr h. Standard deviation i. Variance
j. Coefficient of variance k. Relative standard deviation

Answers:

Multiple Choice Questions

1. c 2. d 3. c 4. d 5. a 6. c 7. a 8. b 9. a 10. b
11. c 12. a 13. a 14. c 15. c

Exercise

1. Mode 18.
2. Mode 10.
3. Mean 20.7.
4. Median 258, mean 259.75.
5. Range 1 to 15.
6. Mean 50.2, standard deviation 1.8.
7. Mean 23.4, standard deviation 2.6.
8. Mean 1.5, standard deviation 0.84.
9. Mean 3.1, median 3.
10. Mean 250, median 250.
11. Median 35, Mean 35, Mode 35, standard deviation 15.40.
12. Mean 3.25, median 3, standard deviation 1.25.
13. Median 86, Mean 56.68, standard deviation 51.85.

14. Mean 51.94, standard deviation 11.42.
15. Steroid 1. Median 8.2, Range: 3.9 to 15.8, Quartile Q_1 6.15, Q_3 10.15.
Steroid 2. Median 6.9, Range: 4.4 to 9.1, Quartile Q_1 5.73, Q_3 7.97.
16. Median 0.71, Range (0.1 to 1.24), Quartile Q_1 0.54, Q_3 0.83.
17. Alpha: CV 0.035, RSD 3.5; Bravo: CV 0.015, RSD 1.5.

The tablets prepared by Alpha machine show more variability than those prepared by Bravo machine.

18. Median 4.1, Quartile Q_1 3.3, Q_3 7.9.
19. Mean 112.78, SD 14.42, CV 0.128, RSD 12.8.
- 20.
- | | |
|---------------------------------------|----------------------------------|
| a. Mean 8.13 | b. Median 8.1 |
| c. 25th percentile 6.925 | d. 75th percentile 9.4 |
| e. 10th percentile 6.34 | f. Range 6.3 to 10.6 |
| g. Iqr 6.925 to 9.4 | h. Standard deviation 1.44 |
| i. Variance 2.09 | j. Coefficient of variance 0.177 |
| k. Relative standard deviation 17.7%. | |



Chapter 13

PROBABILITY AND PROBABILITY DISTRIBUTION

Learning objectives

When we have finished this chapter, we should be able to:

1. Define probability and calculate simple probabilities.
2. Explain the proportional frequency approach to calculate probability.
3. Explain probabilities of simple and composite outcomes, probability involving two variables and conditional probability.
4. Explain how probability can be used with the area properties of the normal distribution.
5. Explain probability distribution, binomial distribution and poisson distribution.

Classic Probability

Statistical concepts are essentially derived from probability theory. Thus, it would be only logical to review some of the fundamentals of probability.

Probability is a measure of the chance of getting some outcome of interest from some event. The event might be rolling a dice and the outcome of interest might be getting a six. Some basic ideas about probability are given below:

1. The probability of a particular outcome from an event will lie between zero and one.
2. The probability of an event that is certain to happen is equal to one. For example, the probability that everybody dies eventually.
3. The probability of an event that is impossible is zero. For example, throwing a seven with a normal dice.
4. If an event has as much chance of happening as of not happening (like tossing a coin and getting a head), then it has a probability of $1/2$ or 0.5 .
5. If the probability of an event happening is p , then the probability of the event not happening is $1 - p$.

Calculating Probability

We can calculate the probability of a particular outcome from an event by using the following formula:

$$\text{Probability } [P^E] = \frac{\text{Number of outcomes that favour the event (m)}}{\text{Total number of outcomes (N)}} \quad \dots 1$$

The probability of a particular outcome from an event is equal to the number of outcomes that favour that event, divided by the total number of possible outcomes.

Example 13.1

What is the probability of getting an odd number when we roll a dice?

Solution

Total number of possible outcomes = 6 (1 or 2 or 3 or 4 or 5 or 6)

Total number of outcomes favouring the event 'an odd number' = 3 (i.e. 1 or 3 or 5)

So probability of getting an odd number = $3/6 = 1/2 = 0.5$

The above method for determining probability works well with experiments where all of the outcomes have the same probability, e.g. rolling dice, tossing a coin, etc. In the real world, however, we will require to use the proportional frequency approach, which uses existing frequency data as the basis for probability calculations.

Example 13.2

Each box contains 10 strips of paracetamol. As a check on quality such 50 boxes were tested and distribution regarding number of defectives and the proportional frequency is calculated as category frequency divided by total frequency.

Table 13.1 Frequency table showing number of defective strips in 50 boxes

No. of defectives	Frequency (n=50)	Proportional frequency
0	30	$30/50 = 0.6$
1	10	$10/50 = 0.2$
2	5	$5/50 = 0.1$
3	3	$3/50 = 0.06$
4	1	$1/50 = 0.02$
5	1	$1/50 = 0.02$

Notice that the proportional frequencies sum to one, similar to probability.

Now if we ask the question, 'What is the probability that if we chose one of these 50 boxes at random and we will get one defective strip in it? The answer is the proportional frequency for the 'one' defective strip i.e. 0.2. In other words, we can interpret proportions as equivalent to probabilities.

Probability of simple outcomes

The outcomes that are mutually exclusive and exhaustive are called as simple outcomes. Let us see one example of determining probability of simple outcome.

Example 13.3

There are 52 cards in a deck, of which 26 are red; therefore, the probability of drawing a red card is

$$P(\text{Red}) = 26/52 = 0.50$$

Note that cards must be red or black, and cannot be both; thus representing mutually

exclusive and exhaustive simple outcomes.

Probability of composite outcomes

When the probability of group of simple outcomes is collectively taken it is referred to as probability of composite outcomes.

Addition theorem

According to addition theorem, the likelihood of two or more mutually exclusive outcomes equals the sum of their individual probabilities.

$$P(E_i \text{ or } E_j) = P(E_i) + P(E_j) \quad \dots 2$$

Example 13.4

The probability of a composite outcome of drawing a face card (jack, queen or king) would equal the sum of their probabilities

$$P(\text{Face card}) = P(\text{King}) + P(\text{Queen}) + P(\text{Jack}) = 1/13 + 1/13 + 1/13 = 3/13 = 0.231.$$

Probability of complementary event

For any outcome E , there is a complementary event (\bar{E}) which can be considered "not E ," Since either E or \bar{E} must occur, but cannot occur at the same time then $P(E) + P(\bar{E}) = 1$ or written for the complement.

$$p(\bar{E}) = 1 - p(E) \quad \dots 3$$

Probability involving two variables

In the case of two different variables, it is necessary to consider the likelihood of both variables occurring, $p(A)$ and $p(B)$, which are not mutually exclusive. A conjoint or union ($A \cup B$) is used when calculating the probability of either A or B occurring. An intersect ($A \cap B$) or joint probability is employed when calculating the probability of both A and B occurring at the same time.

1. Probability of Intersect ($A \cap B$)

The probability of an intersect is easily determined using the multiplication theorem, in which $p(A \cap B) = p(A) \times p(B)$ if A and B are independent of each other.

$$p(A \text{ and } B) = p(A) \times p(B)$$

Example 13.5

What is the probability of drawing a card which is both a queen and a heart (figure 13.1)

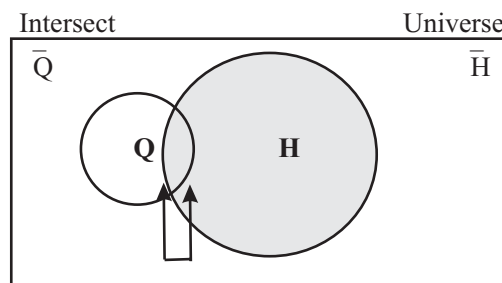


Figure 13.1 Probability of drawing a card which is both a queen and a heart

$$p(\text{queen and heart}) = p(Q \cap H) = 1/52$$

$$p(\text{queen and heart}) = p(\text{queen}) \times p(\text{heart}) = 1/13 \times 1/4 = 1/52$$

In this case there is obviously only one queen of hearts in a deck of cards.

2. Probability of conjoint

The probability of either A or B occurring i.e conjoint can be determined by using addition theorem and subtracting intersect. The $p(A \cup B)$ equals the sum of the two probabilities minus the probability associated with the intersect.

$$P(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Example 13.6

Take example of probability of drawing either queen or heart.

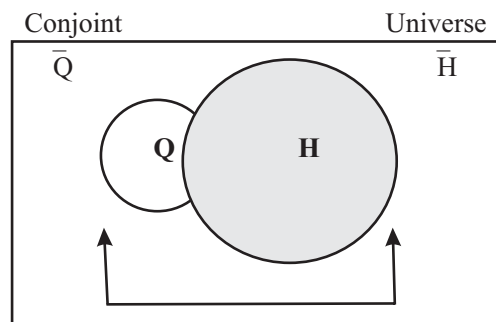


Figure 13.2 Schematics of Conjoint Probability Distribution

Solution:

We can compute the conjoint as given below

$$p(\text{queen or heart}) = P(Q \cup H) = p(Q) + p(H) - p(Q \cap H)$$

$$p(\text{queen or heart}) = P(Q \cup H) = 4/52 + 13/52 - 1/52 = 16/52$$

Here there are 13 heart cards and four queens for a total of 17, but one of the queens is also a heart, thus the 16 possible outcomes.

Let us take example illustrating conjoint and intersect.

Example 13.7

Consider that in a survey of 407 villages in Satara district, 219 villages had medical shops while 305 had doctors in their villages; and 192 villages had both doctor and medical shops. Assuming that this sample is representative of villages nationally,

1. What is the probability of selecting a village at random and finding that this village has a medical shop (MS)?
2. What is the probability of selecting a village at random and finding that this village has doctors?
3. What is the probability of selecting a village at random and finding that this village does not have doctor?

4. What is the probability of selecting a village at random that has both medical shop and doctor (intersect)?
5. What is the probability of selecting a village at random who has either medical shop or doctor but both (conjoint)?

Solution:

1. Probability of medical shop, $p(\text{MS})$

$$m(\text{MS}) = \text{No of outcomes favoring event (here medical shop)} = 219$$

$$N = \text{No of possible outcomes (here villages)} = 407$$

$$p(\text{MS}) = m(\text{MS})/N = 219/407 = 0.538$$

The probability of selecting a village having medical shops at random is **0.538**.

2. Probability of Doctor, $p(\text{D})$

$$m(\text{D}) = 305; N = 407$$

$$p(\text{D}) = m(\text{D})/N = 305/407 = 0.749$$

The probability of selecting a village having doctor at random is **0.749**.

3. Probability of no Doctor, $p(\text{nD})$

$$m(\text{nD}) = (407 - 305); N = 407; p(\text{D}) = 0.749$$

$$p(\text{nD}) = m(\text{nD})/N = (407 - 305)/407 = 0.251$$

or

$$p(\text{nD}) = 1 - p(\text{D}) = 1 - 0.749 = 0.251$$

The probability of selecting a village having no doctor at random is **0.251**.

4. Probability of medical shop and doctor

$$m(\text{MS} \cap \text{D}) = 192; N = 407$$

$$p(\text{medical shop and doctor}) = p(\text{MS} \cap \text{D}) = m(\text{MS} \cap \text{D})/N = 192/407 = 0.472$$

The probability of selecting a village having both medical shops and doctor at random is **0.472**.

5. Probability of medical shop or doctor $p(\text{MS or D})$

$$p(\text{MS}) = 0.538, p(\text{D}) = 0.749, p(\text{MS} \cap \text{D}) = 0.472$$

$$p(\text{medical shop or doctor}) = p(\text{MS} \cup \text{D}) = p(\text{MS}) + p(\text{D}) - p(\text{MS} \cap \text{D}) \\ = 0.538 + 0.749 - 0.472 = 0.915$$

The probability of selecting a village having medical shop or doctor at random is **0.915**.

Conditional Probability

Many times it is necessary to calculate the probability of an outcome, given that a certain value is already known for a second variable. For example, what is the probability of event A occurring given the fact that only a certain level (or outcome) of a second variable (B) is considered.

$$p(\text{A}) \text{ given B} = p(\text{A}|\text{B}) = p(\text{A} \cap \text{B})/p(\text{B}) \quad \dots 4$$

Example 13.8

What is the probability of drawing a queen from a stack of cards containing all the hearts from a single deck?

Solution

$$p(Q \cap H) = 1/52, \quad p(H) = 13/52$$

$$\text{Probability of (Queen| Heart), } p(Q|H) = p(Q \cap H)/p(H) = (1/52)/(13/52) = 1/13$$

In this example, if all the hearts are removed from a deck of cards, **1/13** is the probability of selecting a queen from the extracted hearts.

Example 13.9

1. From the previous example 13.7, if a selected village has a medical shop, what is the probability that this same village also has doctor?
2. If the selected village has doctor, what is the probability that this same village also has access to a medical shop?

Solution:

1. If selected village is having medical shop, probability that this same village will have doctor is:

$$p(D|MS) = p(MS \cap D) / p(MS) = 0.472/0.538 = \mathbf{0.877}$$

2. If selected village is having doctor, probability that this same village will have medical shop is:

$$p(MS|D) = p(MS \cap D) / p(D) = 0.472/0.749 = \mathbf{0.630}$$

Probability Distribution

Inferential statistics employs probability theory to make assumptions about the properties of populations on the basis of data recorded from smaller samples taken from population. An instrumental component of such estimations is the use of probability distributions, i.e. the relationships between particular variable and their probability of occurrence.

1. Discrete probability distribution

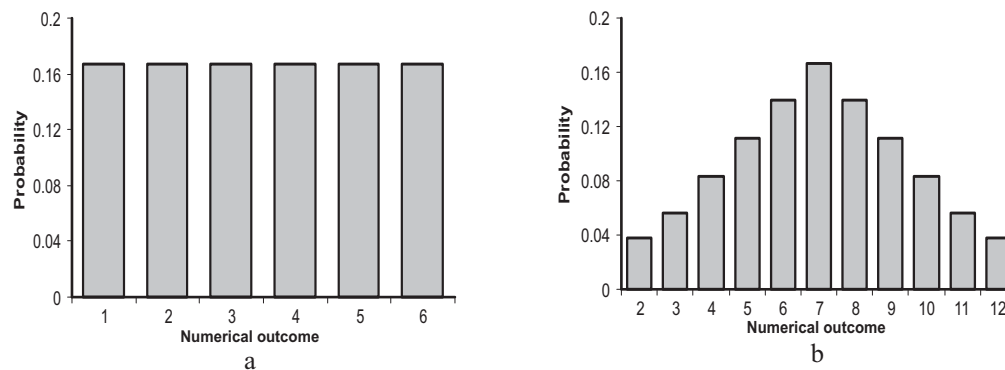
Discrete probability distributions are those in which the probability of the occurrence of discrete events is calculated and graphically portrayed. To illustrate these, consider firstly the numerical outcome following the rolling of one die and then two dice (Table 13.2)

The plot of theoretical probabilities against defined variables is referred to as a probability distribution. Frequency distributions represents the distribution of data derived from analysis of a sample taken from a population, whereas probability distributions reflect the distributions of a variable in a population.

In a probability distribution, the sum of all the individual probabilities is always 1 (the area under the plotted distribution is 1).

Table 13.2 Probabilities associated with defined numerical values obtained following the rolling of one or two dice

One dice		Two dice	
Variable (Numerical outcome)	Probability	Variable (Numerical outcome)	Probability
1	0.167	2	0.038
2	0.167	3	0.056
3	0.167	4	0.083
4	0.167	5	0.111
5	0.167	6	0.139
6	0.167	7	0.167
		8	0.139
		9	0.111
		10	0.083
		11	0.056
		12	0.038

**Figure 13.2** Probability distributions for the numerical values shown in table 13.2 after one dice (a) and two dice (b)

The distribution shown in figure 13.2 is a discrete probability distribution because the variable can adopt a countable number of values.

2. Binomial distribution

One of the distributions most commonly employed in the pharmaceutical and life science is the binomial distribution. This distribution is used whenever the outcome of an event consists of only two categories. An example of a binomial event has been described previously, namely tossing of a coin. The other examples of binomial data include:

- the outcome of a quality control assessment, which is either a pass or a fail.
- a new formulation may produce side effects: the outcomes is either positive or negative (no effects)
- the gender distribution in a population: the disease is either present or absent
- a new pharmaceutical agent is either clinically efficacious or non-efficacious.

In addition to the requirement for only two possible outcomes, each binomial trial must be independent, i.e. the occurrence of one events must not influence subsequent events. In the generation of the binomial distribution, it is assumed that the proportion of observations (or individuals) in one category is p and, consequently, the proportion of observations in the other category is $1-p$, i.e. q .

The probability of events using binomial data can be calculated by expansion of the binomial term, $(p+q)^n$, in which n denotes the sample size, p is probability of the occurrence of the first event, q is the probability of the occurrence of the second event.

The probability of X observations in a sample of size n that has been removed from a binomial distribution may be mathematically described as follows:

$$P(X) = \binom{n}{X} p^X q^{n-X} \quad \dots 6$$

Where p^X denotes the probability of a sample composed of X observations possessing a possibility p and q^{n-X} denotes the probability of a sample composed of $n-X$ observations possessing a probability q .

The above equation can be rewritten by substituting

$$\binom{n}{X} = \frac{n!}{X!(n-X)!} \quad \dots 7$$

$$P(X) = \frac{n!}{X!(n-X)!} p^X q^{n-X} \quad \dots 8$$

Example 13.3

In pharmaceutical manufacturing, tablets from a batch may be categorized into those that pass or those that fails quality control. As part of QC process for a batch of tablets, the probability associated with the pass category was 0.95 whereas the probability associated with fail category was 0.05. Using the binomial expansion calculate the probability of selecting three defective tablets in a sample of three.

Solution:

The probabilities were p (defective)= 0.05,

q (non defective)= 0.95,

The number of observations (X)= 3,

Sample size (n)= 3

$$\begin{aligned} P(X) &= \binom{n}{X} p^X q^{n-X} = \binom{3}{3} (0.05)^3 (0.95)^{3-3} \\ &= \frac{3!}{3!(0!)} \times 0.000125 = 0.000125 \end{aligned}$$

Thus, the probability of selecting three defective tablets in a sample of three is 0.000125.

3. Poisson distribution

The Poisson distribution is another discrete data distribution that is commonly employed to describe random occurrences when the probability of observing an event is small. the Poisson distribution approximates to the binomial distribution when the sample size is large and the probability of a specified event is small. Mathematically, the Poisson distribution is described by

$$P(X) = \frac{e^{-\mu} \mu^X}{X!} = \frac{\mu^X}{e^\mu X!} \quad \dots 9$$

Where

$p(X)$ = the probability of an event occurring in a single observation

μ = the mean number of occurrences (number of observations, x and probability, Np).

The above equation may be expanded to enable calculation of the probabilities of occurrence of an event or events (Table 13.3).

Table 13.3 Expansion of the equation that defines the Poisson distribution

Variable (Numerical outcome)	Probability
0	$P(0) = e^{-\mu}$
1	$P(1) = e^{-\mu} \mu$
2	$P(2) = e^{-\mu} \mu^2 / 2!$
3	$P(3) = e^{-\mu} \mu^3 / 3!$
4	$P(4) = e^{-\mu} \mu^4 / 4!$

Probability and Normal Distribution

If data is Normally distributed then about 95 per cent of the values will lie no further than two standard deviations from the mean (see Figure 13.3). In probability terms, we can say that there is a probability of 0.95 that a single value chosen at random will lie no further than two standard deviations from the mean.

1. About **68 %** of the observations lie within one standard deviation either side of mean.
2. About **95 %** of the observations lie within two standard deviation either side of mean.
3. About **99 %** of the observations lie within three standard deviation either side of mean.

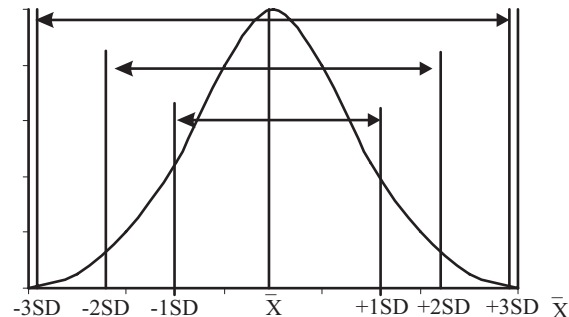


Figure 13.3 The area of properties of the normal distribution

Summary

Probability is a measure of the chance of getting some outcome of interest from some event.

$$\text{Probability } [P^E] = \frac{\text{Number of outcomes that favour the event (m)}}{\text{Total number of outcomes (N)}}$$

Probability of composite outcomes

$$P(E_i \text{ or } E_j) = P(E_i) + P(E_j)$$

Probability of complementary event

$$p(\bar{E}) = 1 - p(E)$$

Probability of intersect

$$p(A \text{ and } B) = p(A \cap B) = p(A) \times p(B)$$

Probability of conjoint

$$p(A \text{ or } B) = p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Conditional probability

$$p(A) \text{ given } B = p(A|B) = p(A \cap B) / p(B)$$

Probability of binomial distribution

$$P(X) = \binom{n}{X} p^X q^{n-X} \quad \binom{n}{X} = \frac{n!}{X!(n-X)!}$$

$$P(X) = \frac{n!}{X!(n-X)!} p^X q^{n-X}$$

Poisson distribution

$$P(X) = \frac{e^{-\mu} \mu^X}{X!} = \frac{\mu^X}{e^\mu X!}$$

Multiple Choice Questions

1. Approximately what percentage of scores fall within one standard deviation of the mean in a normal distribution?

- | | |
|--------|--------|
| a. 34% | b. 95% |
| c. 99% | d. 68% |

2. The sum of all probabilities is _____.

- | | |
|--------|-------|
| a. 1 | b. 2 |
| c. 100 | d. 10 |

- ## Exercise

1. A newly designed shipping containers for ampules was compared to the existing one to determine if the number of broken units could be reduced. One hundred shipping containers of each design (old and new) were subjected to identical rigorous abuse. The containers were evaluated and failures were defined as containers with more than 1% of the ampules broken. A total of 15 failures were observed and 12 of those failures were with the old container. If one container was selected at random:
- What is the probability that the container will be of the new design?
 - What is the probability that the container will be a "failure"?
 - What is the probability that the container will be a "success"?
 - What is the probability that the container will be both an old container design and a "failure"?

- e. What is the probability that the container will be either of the old design or a "failure"?
 - f. If one container is selected at random from only the new containers, what is the probability that the container will be a "failure"?
 - g. If one container is selected at random from only the old container design, what is the probability that the container will be a "success"?
2. Total of 150 healthy females volunteered to take part in a multi-center study of a new urine testing kit to determine pregnancy. One-half of the volunteers were pregnant, in their first trimester. Urinary pHs were recorded and 62 of the volunteers were found to have a urine pH less than 7.0 (acidic) at the time of the study. Thirty-six of these women with acidic urine were also pregnant.
- If one volunteer is selected at random:
- a. What is the probability that the volunteer is pregnant?
 - b. What is the probability that the volunteer has urine that is acidic, or less than a pH 7?
 - c. What is the probability that the volunteer has a urine 'which is basic or a pH equal to or greater than 7?
 - d. What is the probability that the volunteer is both pregnant and has urine which is acidic or less than pH 7?
 - e. What is the probability that the volunteer is either pregnant or has urine which is acidic or less than pH 7?
 - f. If one volunteer is selected at random from only those women with acidic urinary pHs, what is the probability that the volunteer is also pregnant?
 - g. If one volunteer is selected at random from only the pregnant women, what is the probability that the volunteer has a urine pH of 7.0 or greater?
3. If a medicine cures 80% of the people who take it, what is the probability that out of the eight people who take the medicine, 5 will be cured?
4. If a microchip manufacturer claims that only 4% of his chips are defective, what is the probability that among the 60 chips chosen, exactly three are defective?
5. If a telemarketing executive has determined that 15% of the people contacted will purchase the product, what is the probability that among the 12 people who are contacted, 2 will buy the product?
6. A department store buys 50% of its appliances from Manufacturer A, 30% from Manufacturer B, and 20% from Manufacturer C. It is estimated that 6% of Manufacturer A's appliances, 5% of Manufacturer B's appliances, and 4% of Manufacturer C's appliances need repair before the warranty expires. An appliance is chosen at random. If the appliance chosen needed repair before the warranty expired, what is the probability that the appliance was manufactured by Manufacturer A? Manufacturer B? Manufacturer C?

Answers:**Multiple Choice Questions**

1. d 2. a 3. c 4. b 5. d 6. b 7. d 8. a 9. b 10. c

Exercise

1.

- | | |
|-----------|-----------|
| a. 0.5, | b. 0.075, |
| c. 0.925, | d. 0.06, |
| e. 0.515, | f. 0.03, |
| g. 0.88 | |

2.

- | | |
|-----------|-----------|
| a. 0.5, | b. 0.413, |
| c. 0.587, | d. 0.24, |
| e. 0.673, | f. 0.581, |
| g. 0.52 | |

3. 0.1464

4. 0.2137

5. 0.292.

6. A-0.03, B-0.015, C-0.008



Chapter 14

SAMPLING TECHNIQUES

Learning objectives

When we have finished this chapter, we should be able to:

1. Explain various methods of sampling techniques
2. Explain characteristics of good sample

Samples are considered to be the true representatives of that population. It is not possible to include all the members of the population in an experimental study because of constraints of cost, time and labour involved. Hence, appropriate sampling techniques are utilised to ensure that the samples are randomly selected and that every observation is independently measured.

Various sampling techniques used are given below:

1. Random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling
5. Multistage sampling
6. Multiphase sampling
7. Sequential sampling

1. Random Sampling

In this method, every unit of the population has an equal chance of being selected. This method is applicable when population is small, homogeneous and readily available. This method is also called as 'Unrestricted Random Sampling'.

Randomization may be accomplished for a smaller number of units by using a random numbers table or by numbers generated at random using a calculator or computer.

Lottery Method

This is the simplest and most popular method of obtaining a random sample. Under this method, various units of population we are going to study are numbered on small and identical slips of paper, which are folded and put them into a box or a bag. Then they are thoroughly mixed and required number of slips for the sample are picked one after the other without replacement. While doing this, it has to be ensured that in successive drawings each of the remaining slips of population has equal chance of being chosen.

Use of Random Number

Common method of drawing sample is by making use of published tables of random

numbers. First give serial numbers to all the people of the study population. Then select at random, any page of the random number table and pick up the number in any row and column at random. The population units corresponding to the numbers are selected.

Advantages of random sampling

1. Scientific method
2. More representative
3. More economical

Disadvantages of random sampling

1. It needs complete list of study population which is often difficult to get.
2. If the sample size is small, this sample will not be a true representative of the universe.
3. Cases selected by random sampling tends to be widely dispersed geographically and cost of collecting data becomes too large.

2. Systematic Sampling

This method is popularly used in those cases when a complete list of population from which sample is to be drawn, is available. It is more often applied to field studies when the population is large, scattered and homogeneous.

The most practical way of sampling is to select every *i*th item on a list. Sampling of this type is known as systematic sampling. An element of randomness is introduced into this kind of sampling by using random numbers to pick up the unit with which to start. Sample interval for systematic sampling is calculated using following formula:

$$\text{Sample interval} = \text{Total population} / \text{Sample size desired}$$

For instance, if a 4 per cent sample is desired, the first item would be selected randomly from the first twenty-five and thereafter every 25th item would automatically be included in the sample. Thus, in systematic sampling only the first unit is selected randomly and the remaining units of the sample are selected at fixed intervals. Although a systematic sample is not a random sample in the strict sense of the term, but it is often considered reasonable to treat systematic sample as if it were a random sample.

Advantages

It can be taken as an improvement over a simple random sample in as much as the systematic sample is spread more evenly over the entire population. It is an easier, accurate and less costlier method of sampling and can be conveniently used even in case of large populations.

Disadvantages

If there is a hidden periodicity in the population, systematic sampling will prove to be an inefficient method of sampling. In practice, systematic sampling is used when lists of population are available and they are of considerable length.

3. Stratified Sampling

If a population from which a sample is to be drawn does not constitute a homogeneous group,

stratified sampling technique is generally applied in order to obtain a representative sample. Under stratified sampling the population is divided into several sub-populations that are individually more homogeneous than the total population (the different sub-populations are called 'strata') and then we select items from each stratum to constitute a sample. Since each stratum is more homogeneous than the total population, we are able to get more precise estimates for each stratum and by estimating more accurately each of the component parts, we get a better estimate of the whole. In brief, stratified sampling results in more reliable and detailed information.

Advantages

1. More representative
2. Greater accurate
3. Administrative convenience
4. More advantageous when distribution of population is skewed.

Disadvantages

1. It is very difficult task to divide the population into homogenous strata. This may require considerable time and money and statistical expertise.
2. The supplementary information to set up strata is not available some times.
3. Sometimes the different strata may overlap and the sampling would not be representative.

4. Multistage Sampling

As the name implies this method refers to the sampling procedures carried out in several stages using random sampling techniques. Ordinarily multi-stage sampling is applied in big inquiry extending to a considerable large geographical area, say, the entire country. In the first stage, random numbers of districts are chosen in all the states, followed by random numbers of talukas, villages and units, respectively.

Advantages

1. It is easier to administer than most single stage designs mainly because of the fact that sampling frame under multi-stage sampling is developed in partial units.
2. A large number of units can be sampled for a given cost under multistage sampling.
3. It introduces flexibility in sampling.
4. It enables the use of existing divisions and subdivisions which saves extra labour.

5. Cluster Sampling

If the total area of interest happens to be a big one, a convenient way in which a sample can be taken is to divide the area into a number of smaller non-overlapping areas and then to randomly select a number of these smaller areas called clusters, with the ultimate sample consisting of all units in these small areas or clusters.

Thus in cluster sampling the total population is divided into a number of relatively small subdivisions which are themselves clusters of still smaller units and then some of these clusters are randomly selected for inclusion in the overall sample.

Advantages

1. Data collection method is simple and economic.
2. Cluster sampling reduces cost by concentrating surveys in selected clusters.
3. Involves less time and cost.
4. Estimates based on cluster samples are usually more reliable per unit cost.

Disadvantages

1. Gives higher standard error.
2. It is less precise than random sampling.
3. There is not as much information in observations within a cluster.

6. Multiphase Sampling

In this method part of information is collected from the whole sample and part of information is from sub sample.

Advantages

1. Less cost
2. Less laborious
3. More purposeful.

7. Sequential Sampling

This sampling design is some what complex sample design. The ultimate size of the sample under this technique is not fixed in advance, but is determined according to mathematical decision rules on the basis of information yielded as survey progresses. This is usually adopted in case of acceptance sampling plan in context of statistical quality control. When a particular lot is to be accepted or rejected on the basis of a single sample, it is known as single sampling; when the decision is to be taken on the basis of two samples, it is known as double sampling and in case the decision rests on the basis of more than two samples but the number of samples is certain and decided in advance, the sampling is known as multiple sampling. But when the number of samples is more than two but it is neither certain nor decided in advance, this type of system is often referred to as sequential sampling. Thus, in brief, we can say that in sequential sampling, one can go on taking samples one after another as long as one desires to do so.

Characteristics of Sample

The main characteristics of a representative sample are as follows:

1. Precision and Accuracy
2. Unbiased character

1. Precision and Accuracy

Precision refers to how closely data are grouped together or the compactness of the sample data. Precision measures the variability of a group of measurements. A precise set of

measurements is compact and is reflected by a small relative standard deviation.

Assume that the smaller box within samples A, B and C represent the true value for the population from which the samples were taken. In this example, even though samples A and C have good precision, only sample C is accurate as a predictor of the population.

Precision is measured by the formula,

$$\text{Precision} = \frac{\sqrt{N}}{s}$$

Where,

N is number of sample.

s is standard deviation.

Accuracy is concerned with "correctness" of the results and refers to how closely the sample data represents the true value of the population. It is desirable to have data that is both accurate and precise.

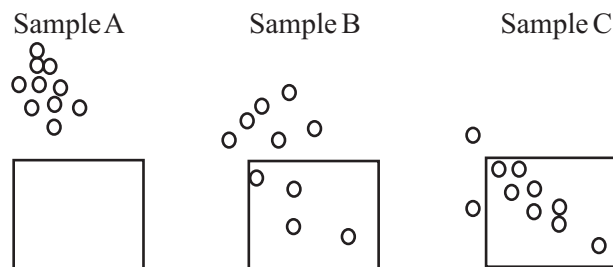


Figure 15.1 Sample comparing precision and accuracy

2. Bias

Bias can be thought of as systematic error which causes some type of constant error in the measurement with the measurement system.

Selection bias occurs when certain characteristics make potential observations more (or less) likely to be included in the study. For example, always sampling from the top of storage drums may bias the results based on particle size, assuming smaller particles settle to the lower regions of the drums. Bias can result from an incorrect sampling, inappropriate experimental design, inadequate blinding, or mistakes (blunders) in observing or recording the data.

The following are the characteristics of a good sample

1. The sample is randomly selected.
2. Every sample is independent of other.
3. The sample is precise and accurate.
4. The sample selection is unbiased.
5. The sample is reliable and valid.

Summary**Sample**

Samples are relatively small group of observations that are taken from a defined population and considered to be the true representative of that population.

Sampling techniques

1. Random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling
5. Multistage sampling
6. Multiphase sampling
7. Sequential sampling

Characteristics of Sample

1. Precision and accuracy
2. Unbiased character

Multiple choice questions

1. Which of the following techniques yields a simple random sample?
 - a. Choosing volunteers from an introductory psychology class to participate
 - b. Listing the individuals by ethnic group and choosing a proportion from within each ethnic group at random.
 - c. Numbering all the elements of a sampling frame and then using a random number table to pick cases from the table.
 - d. Randomly selecting schools, and then sampling everyone within the school.
2. Which of the following sampling techniques is an equal probability selection method in which every individual in the population has an equal chance of being selected?
 - a. Simple random sampling
 - b. Systematic sampling
 - c. Proportional stratified sampling
 - d. All of the above
3. Which of the following will give a more "accurate" representation of the population from which a sample has been taken?
 - a. A large sample based on the convenience sampling technique
 - b. A small sample based on simple random sampling
 - c. A large sample based on simple random sampling
 - d. A small cluster sample
4. Which of the following would generally require the largest sample size?
 - a. Cluster sampling
 - b. Simple random sampling

- c. Systematic sampling d. Proportional stratified sampling
5. Which of the following would usually require the smallest sample size because of its efficiency?
- a. One stage cluster sampling b. Simple random sampling
- c. Two stage cluster sampling d. Quota sampling
6. A technique used when selecting clusters of different sizes is called ____.
- a. Cluster sampling b. One-stage sampling
- c. Two-stage sampling d. Probability proportional to size or PPS
7. The process of drawing a sample from a population is known as ____.
- a. Sampling b. Census c. Survey research d. None of the above
8. _____ is a set of elements taken from a larger population according to certain rules.
- a. Sample b. Population c. Statistic d. Element
9. Determining the sample interval (represented by k), randomly selecting a number between 1 and k, and including each kth element in your sample are the steps for which form of sampling?
- a. Simple Random Sampling b. Stratified Random Sampling
- c. Systematic Sampling d. Cluster sampling
10. A _____ is a subset of a ____.
- a. Sample, population b. Population, sample
- c. Statistic, parameter d. Parameter, statistic

Exercise

1. Explain various methods of sampling techniques.
2. What are the characteristics of good sample?
3. From the batch of 50 tablets, sampling of 10 tablets is to be done. Explain how the systematic sampling and random sampling method can be used?
4. Write a note on stratified sampling.
5. Explain the terms: a) Precision b) Accuracy c) Bias

Answers:

Multiple Choice Questions

1. c 2. d 3. c 4. a 5. a 6. d 7. a 8. a 9. c 10. a



Chapter 15

ESTIMATION OF CONFIDENCE INTERVAL

Learning objectives

When we have finished this chapter, we should be able to:

1. Explain what the standard error of the sample mean is and calculate its value.
2. Explain how we can use the probability properties of the Normal distribution to measure the preciseness of the sample mean as an estimator of the population mean.
3. Estimate the confidence interval of the population mean.
4. Calculate and interpret a 95 per cent confidence interval for a population mean.

Concept of Confidence Interval

The mean and standard deviation of sample data are employed to estimate the true mean and true standard deviation. However, it is reasonable to know how reliable is the sample data at representing the population data. After calculating the mean and standard deviation of a sample, as is the normal approach in the pharmaceutical sciences, we need to provide an indication of the reliability of the data.

In light of the small sample size, it is unreasonable to predict that the population mean and standard deviation will be identical to these observed sample values, as each sample will produce different mean and standard deviation values. Therefore, when reporting the mean of sample data it is good practice to present some indication of the reliability of the data, i.e. the quality of the estimation of the true mean from the sample mean. This is performed using confidence intervals. The confidence intervals are quoted as a mean and range, the latter representing the probability of observing the true mean.

Standard Error of the Mean

The standard deviation (s or SD) describes the variability within a sample; whereas, the standard error of the mean (SEM) represents the possible variability of the mean itself. The SEM is sometimes referred to as the standard error (SE) and describes the variation of all possible sample means and equals the standard deviation (SD) of the sample data divided by the square root of the sample size. The distribution of sample means (the standard error of the mean) will always be smaller than the dispersion of the sample (the standard deviation). The SEM is calculated by the following formula:

$$SEM = \frac{SD}{\sqrt{N}} \quad \dots 1$$

Where

SEM = Standard error of mean

SD = Standard deviation

N = number of observations

Standard error of the mean can be considered as a measure of precision. Obviously, the smaller the SEM, the more confident we can be that our sample mean is closer to the true population mean. A general rule of thumb is that with samples of 30 or more observations, it is safe to use the sample standard deviation as an estimate of population standard deviation.

Example 15.1

Diastolic blood pressure of 322 males was taken. Mean BP was found to be 95 and SD 12 mm. Determine SEM.

Answer:

Data: N=322,

s = 12 mm

Formula:

$$SEM = \frac{s}{\sqrt{N}}$$

Solution:

$$SEM = \frac{12}{\sqrt{322}} = 0.67$$

SEM was found to be 0.67.

Example 15.2

Hardness of tablets from same batch was determined using Pfizer type hardness tester. The data is as follows:

5.9 6.4 5.7 5.4 4.9 6.2 5.8 5.5 5.3 5.2

Calculate SEM.

Solution

1. First, calculate standard deviation of the given data

Formula:

$$\text{Standard deviation (s)} = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}}$$

Solution:

Where

$$\sum X = 56.3$$

$$\sum X^2 = 318.89$$

$$\text{Standard deviation (s)} = \sqrt{\frac{318.89 - \frac{(56.3)^2}{10}}{10 - 1}}$$

$$\text{Standard deviation (s)} = \sqrt{0.21}$$

$$\text{Standard deviation (s)} = 0.458$$

X	X ²
5.9	34.81
6.4	40.96
5.7	32.49
5.4	29.16
4.9	24.01
6.2	38.44
5.8	33.64
5.5	30.25
5.3	28.09
5.2	27.04
56.3	318.89

2. Now determine SEM

$$N = 10; \quad s = 0.458$$

Formula:

$$\text{SEM} = \frac{s}{\sqrt{N}}$$

Solution:

$$\text{SEM} = \frac{0.458}{\sqrt{10}} = 0.144$$

SEM was found to be 0.144

Estimation of Confidence Interval

We have already seen in chapter 13 that we can be 95 percent confident that any sample mean is going to be within plus or minus two standard errors of the population mean.

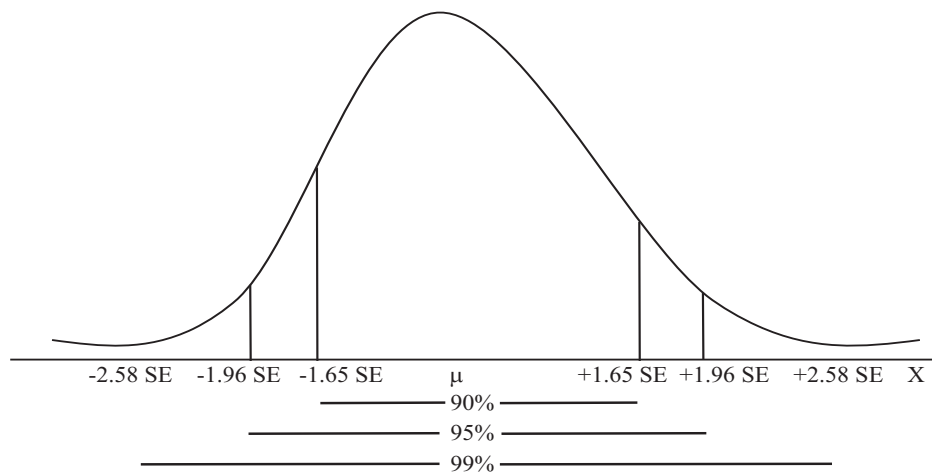


Figure 15.1 Confidence limits of sample mean (\bar{X}) from the population mean (μ) are ($\mu \pm 1.96 \text{ SE}$ and $\mu \pm 2.58 \text{ SE}$)

From this we can say that:

At 95% confidence interval level,

$$\text{Population mean} = \text{sample mean} \pm 1.96 \times \text{standard error} \quad \dots 2$$

That is:

1. We can be 95 per cent confident that the interval, from the sample mean $- 1.96 \times$ standard error, to the sample mean $+ 1.96 \times$ standard error, will include the population mean.
2. Or in probability terms, there is a probability of 0.95 that the interval from the sample mean $- 1.96 \times$ standard error, to the sample mean $+ 1.96 \times$ standard error, will contain the population mean.

In other words, if we pick one out of all the possible sample means at random, there is a probability of 0.95 that it will lie within two standard errors of the population mean. We call the distance from the sample mean $- 1.96 \times \text{SEM}(\bar{X})$, to the sample mean $+ 1.96 \times \text{SEM}(\bar{X})$, the confidence interval.

Confidence interval for mean can be calculated using following formula:

$$p \% = \bar{X} \pm \frac{Z_{1-\alpha/2} \sigma}{\sqrt{N}} \quad \dots 3$$

Where,

$p \%$ = selected confidence interval (90%, 95% or 99%)

\bar{X} = observed mean

$Z_{(1-\alpha/2)}$ = z value corresponding to the percentage confidence interval (1.65 for 90%), (1.96 for 95%) and (2.58 for 99%)

σ = standard deviation of population

N = number of observations

Example 15.3

Clinical study of new anticoagulant drug has been performed in 30 patients, in which the volume of distribution has been calculated as 10.2 ± 1.9 l. Calculate the 95% confidence interval of the mean value.

Answer:

Data:

\bar{X} = observed mean = 10.2

$Z_{(1-\alpha/2)}$ = z value corresponding to the 95% percentage confidence interval = 1.96

σ = standard deviation = 1.91

N = number of observations = 30

Formula:

$$p_{95\%} = \bar{X} \pm \frac{Z_{1-\alpha/2} \sigma}{\sqrt{N}}$$

Solution

$$p_{95\%} = 10.2 \pm \frac{1.96 \times 1.91}{\sqrt{30}} = 10.2 \pm 0.68$$

The 95% confidence interval was found to be 10.2 ± 0.68 L, i.e. 9.52-10.88 L.

Example 15.4

The following are the results of assay

Tablet No	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
Assay (mg)	75	74	72	78	78	74	75	77	76	78	73	77	75	74	72
Tablet No	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Assay (mg)	75	76	75	73	76	79	73	76	75	80	77	74	76	77	74

Assume that three assays are selected at random to give sample of tablets: 4, 18 and 26.

Estimate 90%, 95% and 99% confidence intervals assuming that population standard deviation is 2.04.

Answer:

Taking the assay results of sampled tablets 4, 18 and 26, let us calculate mean

Data:

78 75 77

Formula

$$\text{Mean } (\bar{X}) = \frac{\text{Sum of all observations } (\sum X)}{\text{Total number of observations } (N)}$$

Solution

$$\text{Mean } (\bar{X}) = \frac{78 + 75 + 77}{3} = \frac{230}{3} = 76.66$$

Mean = 76.66 mg

1. 90 % confidence interval**Data:**

\bar{X} = observed mean = 76.66

$Z_{(1-\alpha/2)}$ = z value corresponding to the 90% percentage confidence interval = 1.65

σ = standard deviation = 2.04

N = number of observations = 3

Formula

$$p_{90\%} = \bar{X} \pm \frac{Z_{1-\alpha/2} \sigma}{\sqrt{N}}$$

Solution

$$p_{90\%} = 76.66 \pm \frac{1.65 \times 2.04}{\sqrt{3}}$$

$$p_{90\%} = 76.66 \pm 1.95$$

$$74.71 < \mu < 78.61$$

90% confidence interval for population mean lies between 76.66 ± 1.95 mg.

2. 95 % confidence interval

Data:

$$\bar{X} = \text{observed mean} = 76.66$$

$$Z_{(1-\alpha/2)} = z \text{ value corresponding to the 95\% percentage confidence interval} = 1.96$$

$$\sigma = \text{standard deviation} = 2.04$$

$$N = \text{number of observations} = 3$$

Formula

$$p_{95\%} = \bar{X} \pm \frac{Z_{1-\alpha/2} \sigma}{\sqrt{N}}$$

Solution

$$p_{95\%} = 76.66 \pm \frac{1.96 \times 2.04}{\sqrt{3}}$$

$$p_{95\%} = 76.66 \pm 2.3$$

$$74.36 < \mu < 78.96$$

95% confidence interval for population mean lies between 76.66 ± 2.30 mg.

3. 99 % confidence interval

Data:

$$\bar{X} = \text{observed mean} = 76.66$$

$$Z_{(1-\alpha/2)} = z \text{ value corresponding to the 99\% percentage confidence interval} = 2.58$$

$$\sigma = \text{standard deviation} = 2.04$$

$$N = \text{number of observations} = 3$$

Formula

$$p_{99\%} = \bar{X} \pm \frac{Z_{1-\alpha/2} \sigma}{\sqrt{N}}$$

Solution

$$p_{99\%} = 76.66 \pm 3.01$$

$$p_{99\%} = 76.66 \pm \frac{2.58 \times 2.04}{\sqrt{3}}$$

$$73.65 < \mu < 79.67$$

99% confidence interval for population mean lies between 76.66 ± 3.01 mg.

Summary**Standard error of Mean**

$$SEM = \frac{SD}{\sqrt{N}} = \frac{S}{\sqrt{N}}$$

At 95% confidence, Population mean = Sample mean \pm 1.96 x SE

Confidence interval

$$p\% = \bar{X} \pm \frac{Z_{1-\alpha/2} \sigma}{\sqrt{N}}$$

Multiple choice questions

- Standard error of mean is a measure of _____.
 a. Accuracy b. Precision c. Bias d. All of above
- What does it mean when we calculate a 95% confidence interval?
 a. The process we used will capture the true parameter 95% of the time in the long run
 b. We can be "95% confident" that the interval will include the population parameter
 c. We can be "5% confident" that the interval will not include the population parameter
 d. All of the above statements are true
- What would happen (other things equal) to a confidence interval if we calculate a 99 percent confidence interval rather than a 95 percent confidence interval?
 a. It will be narrower b. It will not change
 c. The sample size will increase d. It will become wider
- What is the standard deviation of a sampling distribution called?
 a. Sampling error b. Sample error
 c. Standard error d. Simple error
- As a general rule, researchers tend to use ____ percent confidence intervals.
 a. 99% b. 95%
 c. 50% d. none of the above
- z value corresponding to the 95 percent confidence interval is _____.
 a. 1.96 b. 1.65
 c. 2.58 d. 1.80
- Standard error of mean (SEM) represents the variability of _____.
 a. sample b. mean itself
 c. SD d. sample size
- With samples less than 30, _____ is an estimate of population SD.
 a. SEM b. sample SD
 c. sample mean d. population mean

- ### Exercise:

- 52 57 49 51 53 52

- ## Answers

1. b 2. d 3. d 4. c 5. b 6. a 7. b 8. a 9. b 10. c

Exercise:

1. 95 % confidence interval = 0.09 ± 0.00196 .
2. Mean = 52.33, SD = 2.683, 95 % confidence interval = 52.33 ± 2.147 .
3. 95 % confidence interval = 524.3 ± 1.533 .
4. 99 % confidence interval = 18.85 ± 1.60 .
5. 90 % confidence interval = 5.62 ± 0.177 .



Chapter 16

HYPOTHESIS TESTING

Learning objectives

When we have finished this chapter, we should be able to:

1. Explain how a research question can be expressed in the form of a testable hypothesis.
2. Explain what a null hypothesis is.
3. Summarise the hypothesis test procedure.
4. Explain what a p-value is.
5. Use the p-value to appropriately reject or not reject a null hypothesis.
6. Describe type I and type II errors, and their probabilities.

Introduction

Hypothesis testing is the process of inferring from a sample whether or not to accept a certain statement about a population or populations. The sample is assumed to be a small representative proportion of the total population. Two errors can occur, rejection of a true hypothesis or failing to reject a false hypothesis.

Using an inferential statistics there are two possible outcomes:

H_0 : Hypothesis Under Test (Null Hypothesis)

H_a : Alternative Hypothesis (Research Hypothesis)

By convention, the Null hypothesis is stated as no real differences in the outcomes.

For example, if we are comparing three levels of a discrete independent variable (μ_1, μ_2, μ_3), the null hypothesis would be stated as $\mu_1 = \mu_2 = \mu_3$. The evaluation then attempts to nullify the hypothesis of no significant difference in favor of an alternative research hypothesis.

Null hypotheses reflect the conservative position of no difference, no risk, no effect, etc., hence the name, 'null' hypothesis. To test this null hypothesis, researchers will have to take samples and measure outcomes, and decide whether the data from the sample provides strong enough evidence to be able to refute or reject the null hypothesis or not. If evidence against the null hypothesis is strong enough for us to be able to reject it, then we are implicitly accepting that some specified alternative hypothesis, usually labeled H_a , is probably true.

Hypothesis Testing Procedure

The hypothesis testing process can be summarised as given below:

1. Select a suitable outcome variable.
2. Use the research question to define an appropriate and testable null hypothesis involving this

outcome variable.

3. Collect the appropriate sample data and determine the relevant sample statistic, e.g. sample mean, sample proportion, sample median, (or their difference or ratio), etc.
4. Use a decision rule that will enable us to judge whether the sample evidence supports or does not support the null hypothesis.
5. Thus, on the strength of this evidence, either reject or do not reject the null hypothesis.

Let's take a simple example. Suppose we want to test whether a coin is fair, i.e. not weighted to produce more heads or more tails than it should. The null hypothesis is that the coin is fair, i.e. it will produce as many heads as tails, so that the population proportion π , equals 0.5. The outcome variable is the sample proportion of heads, p . If we toss the coin 100 times, and get 41 heads, then $p = 0.41$. Is this outcome compatible with our hypothesised value of 0.5? Is the difference between 0.5 and 0.41 statistically significant or could it be due to chance? We decide what proportion of heads we might expect to get if the coin is fair, by using two approaches.

1. creation of a confidence interval or
2. a comparison with a critical value.

The estimation of confidence interval has already been studied in the previous chapter, This can be estimated by using the formula

$$p\% = \bar{X} \pm z\% \times \frac{\sigma}{\sqrt{N}} \quad \dots 1$$

Population Mean = Estimate Sample Mean \pm Reliability Coefficient \times Standard Error.

In the second method we would calculate a "test statistic", a value based on the manipulation of sample data. This value is compared to a preset "critical" value (usually given in a special table) based on a specific acceptable error rate (i.e., 5%). If the test statistic is extremely rare it will be to the extreme of our critical value and we will reject the null hypothesis under test in favor of the research hypothesis.

Decision rule for hypothesis testing

1. Determine the p-value for the output we have obtained. A p-value is the probability of getting the outcome observed (or one more extreme), assuming the null hypothesis to be true.
2. Compare it with the critical value, usually 0.05.
3. If the p-value is less than the critical value, reject the null hypothesis; otherwise do not reject it.

It's important to stress that the p-value is not the probability that the null hypothesis is true or not true. It's a measure of the strength of the evidence against the null hypothesis. The smaller the p-value, the stronger the evidence. This means it is less likely that the outcome we have got is occurred by chance. Note that the critical value, usually 0.05 or 0.01, is called the significance level of the hypothesis test and is denoted as α (alpha).

Types of Error

Whenever we decide either to reject or not to reject a null hypothesis, we may be making a mistake. After all, we are basing the decision on sample evidence. Even if we have done everything

right, the sample could still, by chance, not be very representative of the population. Moreover, the test might not be powerful enough to detect an effect if there is one.

There are two possible errors associated with hypothesis testing. Type I error is the probability of rejecting a true null hypothesis (H_0) and Type II error is the probability of accepting a false H_0 . Type I error is also called the level of significance and is denoted by the symbol α . Alternatively, level of confidence is given as $(1-\alpha)$. Type II error is symbolized using the Greek letter β . The probability of rejecting a false H_0 is called power $(1-\beta)$. Type I error is called as false positive while Type II error is called as false negative.

Table 16.1 Summary of the relationships between statistical outcome and statistical errors

Statistical outcome (decision)	Real results	
	Null hypothesis is true	Null hypothesis is false
Non rejection of the null hypothesis	Correct decision	Type II (β) error
Rejection of the null hypothesis/ acceptance of alternative hypothesis	Type I (α) error	Correct decision

One-tailed and two-tailed tests

If there are two possible statistical outcomes in the study, then a two tailed test must be used. Conversely, if there is only one outcome of interest to the investigator, the test statistic must be interpreted using a one tailed outcome.

The decision as to whether the calculated test statistic should be evaluated as a one tailed test or a two tailed test is crucial. An incorrect choice may result in the incorrect interpretation of the statistical analysis because a significant difference between treatments may be declared insignificant and vice versa.

Defining the critical region for statistical test

Assuming $\alpha = 0.05$, the critical region of the standardised normal distribution may be given in three ways. In all three cases, the chosen level of significance is 0.05, as denoted by shaded region. Calculated values of the z statistic (the test statistic that relates to the standardised normal distribution) that reside within these regions allow the analyst to reject the null hypothesis.

In a two tailed test, the rejection region is equally divided into two sections, each relating to a single tail that resides at either extreme of the probability distribution. As $\alpha = 0.05$, each tail occupies 2.5% of the distribution. In the case of the standardised normal distribution, the critical regions are denoted by z values of either $\geq +1.96$ or ≤ -1.96 . Conversely, in the one tailed test, the rejection region is distributed at either tail of the distribution.

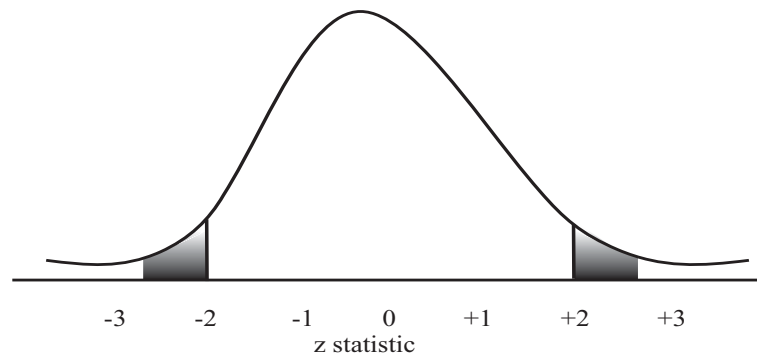


Figure 16.1 Standardised normal distribution showing the (shaded) region of the z statistic for a two-tailed test.

If the alternative hypothesis states, that ‘the mean of a treatment group is greater than the mean of another group’, then the critical value of z is $\geq +1.65$ and test is referred to as a positive one tailed test.

When the alternative hypothesis states, that the mean of a treatment group is lower than the mean of another group, the rejection region resides at the left-hand (negative) side of standardised normal distribution and is defined by z values that are ≤ -1.65 .

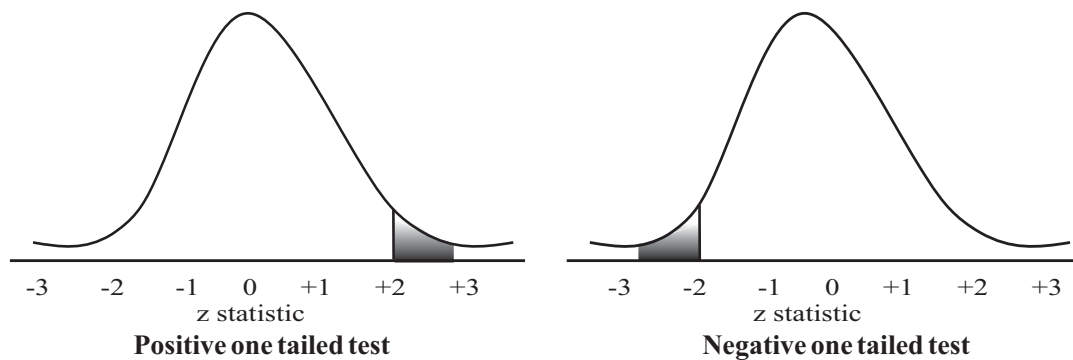


Figure 16.2 Standardised normal distribution showing the (shaded) region of the z statistic for a one-tailed test.

Key stages in hypothesis testing

1. State the null hypothesis.

e.g. $\mu_1 = \mu_2$; no difference in means of two groups.

2. State the alternative hypothesis.

e.g. $\mu_1 \neq \mu_2$; there is difference in means of two groups.

3. Select the level of significance

e.g. $\alpha = 0.05$ or 0.01 or 0.001 . Usually 0.05 is taken as level of significance. We have to set the level of significance before start of the study.

4. Select the number of tails

- e.g. one tailed $\mu_1 < \mu_2$; the mean of group 2 is more than mean of group 1.
 $\mu_1 > \mu_2$; the mean of group 1 is more than mean of group 2.
- two tailed $\mu_1 \neq \mu_2$; the mean of group 1 and group 2 are not equal but we don't know which is greater.

5. Test the statistics

Determine p value by suitable statistical test.

6. Compare table and observed value

p value calculated is compared with table value (critical value).

7. Decide whether to accept or reject null hypothesis

Based on the comparison, either reject or accept the null hypothesis. If the obtained p value is less than critical value, reject null hypothesis or otherwise accept it.

Summary**Hypothesis testing**

It is the process of inferring from a sample whether or not to accept a certain statement about a population (s).

Key stages in hypothesis testing

1. State the null hypothesis
2. State the alternative hypothesis
3. Select the level of significance
4. Select number of tails
5. Test the statistics
6. Compare table and observed value
7. Decide whether to accept or reject null hypothesis

Decision rule

1. Determine p value.
2. Compare it with the critical value, usually 0.05.
3. If the obtained p value is less than critical value, reject null hypothesis; otherwise accept it.

Types of error

Type I error is the probability of rejecting a true null hypothesis

Type II error is the probability of accepting a false H_0 .

Multiple choice questions

1. Hypothesis testing is the process of inferring from a _____ whether or not to accept a certain statement about a _____.

a. sample, population

b. population, sample

- ### Exercise

1. Write the alternative hypothesis for each of the following null hypotheses:
- $\mu_A = \mu_B$
 - $\mu_H \geq \mu_I$

- c. $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$
 - d. $\mu_A \leq \mu_B$
 - e. $\mu = 115$
 - f. populations C, D, E, F and G are the same
 - g. both samples come from the same population
2. What is hypothesis testing?
 3. Give key stages in hypothesis testing.
 4. Give various types of errors in hypothesis testing.
 5. What is one tailed and two tailed test?

Answers:**Multiple Choice Questions**

1. a 2. c 3. a 4. b 5. b 6. c 7. a 8. b 9. a 10. b



Chapter 17

CHOICE OF STATISTICAL TESTS

Learning objectives

When we have finished this chapter, we should be able to:

1. Choose appropriate statistical test for given type of data.
2. Understand parametric and non parametric statistical tests.

One of the major steps in the process of statistical hypothesis testing involves the choice of statistical test. The most appropriate statistical test is chosen according to the desired power of the study, the nature of the population from which the observations are taken and the nature of measurement of the variable.

Parametric and non parametric analysis

Statistical tests may be differentiated into two categories, known as parametric and non-parametric tests. The following assumptions should hold before a parametric statistical method is selected and used:

1. The samples should be removed from a normally distributed population.
2. The samples should be independent.
3. The variances of the populations under examination should be similar. This is termed as homoscedasticity.
4. The variable under examination must be metric, discrete or continuous.

Therefore, parametric tests are utilised for the metric data while nominal and ordinal data is analysed by non-parametric tests.

Let us see statistical tests used according to the type of data.

1. Nominal data

The statistical analysis of nominal data may be performed using a χ^2 (Chi square) analysis or a binomial-based test. Characterization of the association within nominal data is performed using the contingency coefficient.

2. Ordinal data

Many non-parametric tests are referred to as ranking tests and may be employed for the analysis of ordinal data. Characterization of the association within ordinal data is commonly performed using the Spearman correlation coefficient.

3. Metric discrete data

Metric discrete data may be analysed using parametric statistical tests. Indeed, if all the assumptions of parametric statistics are valid, statistical comparisons of groups of interval data

should be performed using parametric tests, e.g. t test, analysis of variance, etc.

4. Metric continuous data

Metric continuous data scales may be manipulated using conventional arithmetic and may therefore be conveniently analysed using either parametric or non-parametric methods. Characterisation of the association within normally distributed interval or ratio data is commonly performed using the correlation coefficient.

More specifically, non-parametric methods are exclusively employed to analyse nominal and ordinal data, whereas parametric and non-parametric methods may be used to examine continuous metric data. If metric data is skewed then non parametric tests are used or otherwise parametric tests are choice for metric data.

Chart Showing Choice of Statistical tests

Type of data	No. of samples to be compared	Matched/ Unmatched	Statistical test	
			Parametric test	Non-parametric test
Categorical data	1	N/A		One sample binomial test
	2	Unmatched		> 20, Chi-square test < 20, Fisher's Exact test
	2	Matched		McNemar's test
	>2	Unmatched		Chi-square test
Metric data	1	N/A	One sample t-test for means; One sample Chi-square test for variances	One sample Wilcoxon signed rank test
	2	Unmatched	Students t-test	Man Whitney U test
	2	Matched	Paired t test	Wilcoxon signed rank test
	>2	Unmatched	One-way ANOVA for means; Bartlett's test of homogeneity for variances	Kruskal Wallis one way ANOVA test
	>2	Matched	Repeated Measures ANOVA	Friedman rank sum test

Some Commonly Used Hypothesis tests**Parametric tests****1. Two-sample t test**

Used to test whether or not the difference between two independent population means is zero (i.e. the two means are equal). Both variables must be metric and Normally distributed. In addition the two population standard deviations should be similar, but for larger sample sizes this becomes less important.

2. Matched-pairs t test

Used to test whether or not the difference between two paired population means is zero. Both variables must be metric, and the differences between the two must be Normally distributed.

3. One way ANOVA

Used to test whether or not the differences between three or more independent groups means is zero. The variables should be metric and normally distributed.

Non parametric tests**1. Mann-Whitney U test**

Used to test whether or not the difference between two independent population medians is zero. Variables can be either metric or ordinal. No requirement as to shape of the distributions, but they need to be similar. This is the non-parametric equivalent of the two-sample t test.

3. Kruskal-Wallis test

Used to test whether the medians of three or more independent groups are the same. Variables can be either ordinal or metric. Distributions may be of any shape, but all need to be similar. This non-parametric test is an extension of the Mann-Whitney test.

4. Wilcoxon signed rank test

Used to test whether or not the difference between two paired population medians is zero. Variables can be either metric or ordinal. Distributions may be of any shape, but the differences should be distributed symmetrically. This is the non-parametric equivalent of the matched-pairs t test.

5. Chi-squared test (χ^2)

Used to test whether the proportions across a number of categories of two or more independent groups is the same. Variables must be categorical. The chi-squared test is also a test of the independence of the two variables.

6. Fisher's Exact test

Used to test whether the proportions in two categories of two independent groups is the same. Variables must be categorical. This test is an alternative to the 2×2 chi-squared test, when cell sizes are too small.

7. McNemar's test

Used to test whether the proportions in two categories of two matched groups is the same. Variables must be categorical.

Summary**Choice of statistical test**

Data Type	Parametric Ratio, Interval	Non-Parametric Ordinal	Frequency Nominal
Single sample	z-test, t-test	Sign test, K-S test	χ^2 Goodness of fit
Two independent samples	z-test, t-test	Mann-Whitney U	> 20, Chi-squared < 20, Fisher's Exact
Two independent paired samples	Paired t-test	Wilcoxon Signed ranks	McNemar's test
More than two independent samples	One-way ANOVA	Kruskall-Wallis	One-way ANOVA
Two factors	Two-way ANOVA		χ^2 Test of Independence
Correlation	Pearson	Spearman	Phi

Multiple Choice Questions

- If metric data is skewed, the choice of statistical test should be _____.
 - parametric tests
 - non parametric tests
 - one sample t test
 - z test
- If data is ordinal or nominal, which of the following test is not used.
 - Non parametric test
 - Kruskal Wallis test
 - Parametric test
 - Mann Whitney U test
- If the data is paired but skewed, the following test is used
 - t test
 - Paired t test
 - Mann Whitney U test
 - Wilcoxon signed rank test
- If the data is paired and metric, the following test is used
 - t test
 - Paired t test
 - Mann Whitney U test
 - One sample t test
- The following test is used to test whether or not the difference between two population means is zero.
 - t test
 - Paired t test
 - Mann Whitney U test
 - One sample t test
- The test for knowing association between two metric variable is _____.
 - Pearson correlation
 - Spearman correlation
 - Phi correlation
 - None of above

- ### Exercise

- Answers:**

1. b 2. c 3. d 4. a 5. a 6. a 7. b 8. b 9. c 10. b



Chapter 18

HYPOTHESIS TESTING FOR ONE SAMPLE MEAN

Learning objectives

When we have finished this chapter, we should be able to:

1. Understand and estimate z statistic for one sample z test.
2. Understand and estimate t statistic for one sample t test.

One sample testing involves the estimation of whether sample data, generated from an experimental procedure, are derived from a defined population or not. More specifically, one sample tests evaluate whether the mean of sample and the mean of the population are different.

Two parametric one sample tests are described here, one sample z test and one sample t test. The non parametric tests, the chi-square one sample test and the Kolmogorov-Smirnov one sample test are used but are out of scope of this book.

1. One Sample z test

One sample z test are employed for the analysis of one sample hypothesis when the sample size is large and, in addition, there is sufficient knowledge about the properties of the population with which the experimentally derived sample is being compared. Typically, z tests (and indeed t tests) compare whether the mean of the sample differs from that of a defined population.

The z statistics is calculated using the following formula

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{N}} \quad \dots 1$$

Where

\bar{X} = observed sample mean

μ_0 = hypothesised mean

σ = standard deviation of the population

N = sample size

Example 18.1

A pharmaceutical company has purchased a new liquid filling machine for use in its liquid oral division. The liquid filling properties of existing machine linked to the current manufacturing process are 5.04 ± 0.27 per bottle. A pilot study has been engaged to evaluate whether (or not) the filling performance of the new machine is similar to that of previous machine using identical process

conditions. Therefore, 100 vials of the product were removed from a pilot and the mean fill volume was determined gravimetrically to be 5.13 ml. Is there a difference between the performance of new and existing filling machines?

Solution: As the mean filling volume of one machine is given and the number of samples are more than 30, one sample z test is employed here. The following steps gives the systematic approach of hypothesis testing.

1. State the null hypothesis

The null hypothesis (H_0) states that there is no significant difference between the sample mean and the population mean.

$$H_0: \mu = 5.04 \text{ ml}$$

2. State the alternative hypothesis

The alternative hypothesis states that there is a difference between the sample mean and the population mean.

$$H_a: \mu \neq 5.04 \text{ ml}$$

3. State the level of significance

The choice of the level of significance (α) is an extremely important consideration in the establishment of the experimental design. Traditionally, α , the probability of rejecting the null hypothesis when it is in fact true, is chosen to be 0.05.

4. State the number of tails

The number of tails associated with the alternative hypothesis is determined before data collection. As alternative hypothesis states $\mu \neq 5.04 \text{ ml}$, it may be less or more than 5.04 ml and the outcome may be two tailed.

5. Select the most appropriate statistical test

A two tailed parametric, one sample test may be applied to the statistical question. As the sample size is large, the z test is the most suitable statistical method.

6. Perform the statistical analysis

Step 1 Calculate the test statistic

Formula:

The z statistic is calculated by using the following formula

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{N}}$$

Data:

\bar{X} = observed sample mean = 5.13

μ_0 = hypothesised mean = 5.04

σ = standard deviation of the population = 0.27

N = sample size = 100

Solution:

$$z = \frac{5.13 - 5.04}{0.27 / \sqrt{100}} = \frac{0.09}{0.027} = 3.33$$

Step 2 Define the critical z statistic

Identification of critical statistic requires knowledge of the chosen level of significance (α) and the number of tails associated with the experimental design. Therefore, the critical z statistics associated with the specified level of significance (0.05) and a two tailed design may be derived from the standardized normal distribution as 1.96. The regions of the z distribution associated with the null and alternative hypothesis may now be defined in terms of the z statistic as:

$$H_0: -1.96 < z < +1.96$$

$$H_a: z \geq +1.96 \text{ or } z \leq -1.96$$

7. Decision

As the calculated z value (+3.33) is greater than +1.96 (i.e. calculated z value does not lie in the region $-1.96 < z < +1.96$), the null hypothesis is rejected and the alternative hypothesis is accepted. It is therefore concluded that there is a significant difference in the performance of the new and existing filling machine, in terms of fill volumes.

Example 18.2

A pharmaceutical company has purchased a new tablet punching machine for use in its compression division. Existing tableting machine produces mean of 1000 tablets/min with a standard deviation of 150. A pilot study has been engaged to evaluate whether (or not) the tableting performance of the new machine is similar to that of previous machine using identical process conditions. The mean tableting rate for 50 samples were determined and found to be 1050 tablets/min. Is there a difference between the performance of new and existing filling machines?

Solution:

As the mean tableting rate of one machine is given and the number of samples are more than 30, one sample z test is employed here. The following steps gives the systematic approach of hypothesis testing.

1. State the null hypothesis

The null hypothesis (H_0) states that there is no significant difference between the sample mean and the population mean.

$$H_0: \mu = 1000 \text{ tablet/min}$$

2. State the alternative hypothesis

The alternative hypothesis states that there is a difference between the sample mean and the population mean.

$$H_a: \mu \neq 1000 \text{ tablet/min}$$

3. State the level of significance

The choice of the level of significance (α) is an extremely important consideration in the

establishment of the experimental design. Traditionally, α , the probability of rejecting the null hypothesis when it is in fact true, is chosen to be 0.05.

4. State the number of tails

As alternative hypothesis states $\mu \neq 1000$ tablets/min, it may be less or more than 1000 tablets/min and the outcome may be two tailed.

5. Select the most appropriate statistical test

A two tailed parametric, one sample test may be applied to the statistical question. As the sample size is large, the z test is the most suitable statistical method.

6. Perform the statistical analysis

Step 1. Calculate the test statistic

Formula:

The z statistic is calculated by using the following formula

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{N}}$$

Data:

\bar{X} = observed sample mean = 1050

μ_0 = hypothesised mean = 1000

σ = standard deviation of the population = 150

N = sample size = 50

Solution:

$$z = \frac{1050 - 1000}{150 / \sqrt{50}} = \frac{50}{21.2} = 2.36$$

Step 2. Define the critical z statistic

Identification of critical statistic requires knowledge of the chosen level of significance (α) and the number of tails associated with the experimental design. Therefore, the critical z statistics associated with the specified level of significance (0.05) and a two tailed design may be derived from the standardized normal distribution as 1.96. The regions of the z distribution associated with the null and alternative hypothesis may now be defined in terms of the z statistic as:

$$H_0: -1.96 < z < +1.96$$

$$H_a: z \geq +1.96 \text{ or } z \leq -1.96$$

7. Decision

As the calculated z value (+2.36) is greater than +1.96, the null hypothesis is rejected and the alternative hypothesis is accepted. It is therefore concluded that there is a significant difference in tableting rate of the new and existing tablet machine.

2. One Sample t test

In the use of the z statistic for the interpretation of one sample statistical hypothesis testing, it

has been assumed that the variance of the population is known. However, in many experiments it is impossible to obtain knowledge of the relevant population statistics and, accordingly, the only available estimates of the central tendency and variability of the population are the mean and variance that are associated with the sample data. Under these circumstances, one sample t test is employed to calculate the t statistic. The process for calculation of the one sample t test is similar to that for the one sample z test.

The t statistics is calculated using the following mathematical formula:

$$t_{(N-1)df} = \frac{\bar{X} - \mu_0}{s / \sqrt{N}} \quad \dots 2$$

\bar{X} = observed sample mean

μ_0 = hypothesised population mean

s = sample standard deviation

N = sample size

$t_{(N-1)df}$ = t value associated with N-1 degrees of freedom.

Example 18.3

To test a manufacturer claim that his fruit juice contains 60 mg of vitamin C per 100ml, a quality controller analysts takes six randomly selected samples with the following results: 65, 58, 62, 57, 62, 65. Is the manufacturer's claim justified (sample mean = 61.5; sample standard deviation (s)=3.39)?

Answer:

Hypothesis testing of this problem is done by using following steps:

1. State the null hypothesis

The null hypothesis states that there is no difference between the sample mean and hypothetical mean.

$$H_0: \mu = 60 \text{ mg}$$

2. State the alternative hypothesis

The alternative hypothesis states that there is a difference between the manufacturer's claim and observed mean.

$$H_a: \mu \neq 60 \text{ mg}$$

3. State the level of significance

It is assumed that the level of significance is 0.05.

4. State the number of tails

The number of tails associated with the alternative hypothesis is determined before data collection. As the sample mean may be either lower or higher than the manufacturers claim, we have to use two tailed test .

5. Select the appropriate statistical test

Whenever the population variance is unknown, one sample t test and not one sample z test is used. A two tailed parametric, one sample t test may be applied to the statistical question.

6. Perform the statistical analysis**Step 1. Calculate the t statistics**

The t statistic is calculated by using the following formula

Formula:

$$t_{(N-1)df} = \frac{\bar{X} - \mu_0}{s / \sqrt{N}}$$

Data:

\bar{X} = observed sample mean = 61.5

μ_0 = hypothesised population mean = 60

s = sample standard deviation = 3.391

N = sample size = 6

$t_{(N-1)df}$ = t value associated with N-1 degrees of freedom = 5

Solution:

$$t_5 = \frac{61.5 - 60}{3.391/\sqrt{6}} = \frac{1.5}{1.38} = 1.09$$

Step 2. Define the critical 't' statistics.

The critical value of 't' statistics for 5 degrees of freedom, two tailed and $\alpha = 0.05$ is given as +2.57 (Appendix 2). Therefore, the calculated t value should be in the range of -2.57 < t < +2.57.

7. Decision

As the calculated t value lies within the region -2.57 < t < +2.57, the null hypothesis is accepted. So the manufacturer's claim that his fruit juice contains 60 mg/100ml of vitamin C is justified.

Example 18.4

To test a manufacturer claim that his tablet contains 60 mg of drug, 10 tablets were selected randomly and analysed spectrophotometrically. Is the manufacturer's claim justified? Analysis results are given below: (sample mean = 62.7; sample standard deviation (s)=6.255)

53, 56, 57, 60, 61, 64, 67, 68, 70, 71

Solution:

Hypothesis testing of this problem is done by using following steps:

1. State the null hypothesis

The null hypothesis states that there is no difference between the sample mean and hypothetical mean.

$$H_0: \mu = 60 \text{ mg}$$

2. State the alternative hypothesis

The alternative hypothesis states that there is a difference between the manufacturer's claim and observed mean.

$$H_a: \mu \neq 60 \text{ mg}$$

3. State the level of significance

It is assumed that the level of significance is 0.05.

4. State the number of tails

The number of tails associated with the alternative hypothesis is determined before data collection. As the sample mean may be either lower or higher than the manufacturer's claim, we have to use two tailed test.

5. Select the appropriate statistical test

Whenever the population variance is unknown, the one sample t test and not the one sample z test is used. A two tailed parametric, one sample t test may be applied to the statistical question.

6. Perform the statistical analysis

Step 1. Calculate the t statistics

The t statistic is calculated by using the following formula

Formula:

$$t_{(N-1)df} = \frac{\bar{X} - \mu_0}{s / \sqrt{N}}$$

Data:

\bar{X} = observed sample mean = 62.7

μ_0 = hypothesised population mean = 60

s = sample standard deviation = 6.255

N = sample size = 10

$t_{(N-1)df}$ = t value associated with N-1 degrees of freedom = 9

Solution:

$$t_s = \frac{62.7 - 60}{6.255 / \sqrt{10}} = \frac{2.7}{1.978} = 1.36$$

Step 2. Define the critical 't' statistics.

The critical value of 't' statistics for 9 degrees of freedom, two tailed and $\alpha = 0.05$ is given as ± 2.26 (Appendix II). Therefore, the calculated t value should be in the range of $-2.26 < t < +2.26$.

7. Decision

As the calculated t value lies within the region $-2.26 < t < +2.26$, the null hypothesis is accepted. So the manufacturer's claim that his tablet contains 60mg drug is justified.

Summary**One sample z test**

Used to test a single sample mean (\bar{X}) when the population mean (μ) and variance (σ) is known.

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{N}}$$

One sample t test

Used to test a single sample mean (\bar{X}) when the population variance (σ) is unknown or $n < 30$.

$$t_{(N-1)df} = \frac{\bar{X} - \mu_0}{s / \sqrt{N}}$$

Multiple choice questions

- One sample z test is used to test a single sample when the _____ is known.
 - population variance
 - sample variance
 - a and b
 - none of above
- One sample t test is used to test a single sample when the _____ is unknown.
 - population variance
 - sample variance
 - a and b
 - none of above
- If we are testing one sample and sample size is more than 30, which of the following test can be used?
 - two sample z test
 - one sample t test
 - one sample z test
 - two sample t test
- df is given by $N-1$. df is _____.
 - degree of freedom
 - degree of fame
 - dance of freedom
 - degree of frame
- The non parametric test for one sample is _____.
 - one sample z test
 - one sample t test
 - chi square one sample test
 - none of above
- One sample tests evaluate whether the mean of sample and the mean of _____ are different.
 - other sample
 - population
 - subjects
 - None of above
- One sample t test is for _____.
 - Parametric data
 - non parametric data
 - nominal data
 - ordinal data
- If the calculated value of z lies in the critical region then null hypothesis is _____.
 - rejected
 - accepted

- During the manufacture of a 1 L infusion, 50 bottles were drawn at random and the volume of each container measured to check whether the full volume met the 1 liter claim. The mean volume was found to be 1008 ml and standard deviation was 65 ml. Is there sufficient evidence to suggest that the average volume of the infusion was not 1 liter at 5% level of significance?
- A random sample of 400 items gives the mean 4.45 with a standard deviation of 2. Can it be regarded as drawn from a normal population with mean 4 at 5% level of significance?
- A randomly 200 medical shops were taken from Satara district and the average sale per shop was found to be 520. Population mean sale was 510 per shop with a standard deviation of 40. Is the difference between sample mean and population mean statistically significant at 5% level of significance?
- A random sample of 45 items gives the mean 73.2 with a standard deviation of 8.6. Can it be regarded as drawn from a normal population with mean 76.7 at 1% level of significance?
- A random sample of 900 children was found to have a mean fatfold thickness at triceps of 3.4 mm with SD of 2.3mm. Can it be reasonably regarded as a representative sample of population having a mean thickness of 3.2mm at 5% level of significance?
- A batch of 40 L of paracetamol suspension (30 mg/ml) was filled into containers. 25 containers of product have been removed for analysis of their drug content. Does the mean concentration of drug in the batch conform to the 30 mg/ml at the 0.05 level of significance.
 Concentration of paracetamol (mg/ml) in 25 containers:

31.5	30.5	30.5	29.8	30.1	30.2	30.2	30.6	29.8	31.1
30.1	30.1	30.1	30.2	29.7	28.9	30	30.4	30.3	29.9
32.1	30.1	28.9	30.8	29.5					
- A pharmaceutical manufacturer does a chemical analysis to check the potency of products. The

standard release potency for cephalothin crystals is 910 and the manufacturer believes this claim may be too high. An assay of 16 lots gives the following potency data:

Data: 897 918 914 906 913 895 906 893 916 908
 918 906 905 907 921 901

Test the manufacturer's claim at the 0.01 level of significance.

8. Ten individuals are chosen at random from a normal population and their weights in kg are found to be

68 63 66 69 63 67 70 70 71 71

Does this sample adequately represent a population in which the mean weight was found to be 66 kgs?

9. Mean Hb % of the population is 14.3. Can a sample of 15 individuals with a mean of 13.5 and SD 1.5 be from same population?

10. Ten tablets are chosen at random from batch and their content in mg are found to be

47 51 48 49 48 52 47 49 46 47

Manufacturer claims that tablet contains 50 mg of drug. Test the manufacturers claim at the 0.05 level of significance.

Answers:

Multiple Choice Questions

1. a 2. a 3. c 4. a 5. c 6. b 7. a 8. b 9. a 10. d

Exercise

1. $z=0.87$; Accept null hypothesis because observed z value is between -1.96 and +1.96 for two tailed test at 0.05% level of significance. It is therefore concluded that the average volume of container was 1 liter.

2. $z= 4.5$; Accept alternative hypothesis because observed z value is more than +1.96 for two tailed test at 0.05% level of significance. It is therefore concluded that the sample has not been drawn from a population with mean 4.

3. $z= 3.53$; Accept alternative hypothesis because observed z value is more than +1.96 for two tailed test at 0.05% level of significance. It is therefore concluded that the sample mean and population mean differ significantly.

4. $z= - 2.73$; Accept alternative hypothesis because observed z value is more than -2.58 for two tailed test at 0.01% level of significance. It is therefore concluded that the sample has not been drawn from a population with mean 76.7.

5. $z= 2.61$; Accept alternative hypothesis because observed z value is more than +1.96 for two tailed

test at 0.05% level of significance. It is therefore concluded that the sample is not representative of population having mean thickness of 3.2 mm at 5% level of significance.

6. Initially determine sample mean (30.216) and standard deviation (0.691). As the population variance is unknown and also sample size is less than 30 use one sample t test. The critical value of 't' statistics for 24 degrees of freedom at 5% level of significance is given as 2.064. Observed t value is 1.52 which is less than t critical. So it can be concluded that mean concentration of drug in the batch conform to the 30 mg/ml at the 0.05 level of significance.

7. Initially determine sample mean (907.75) and standard deviation (8.48). As the population variance is unknown and also sample size is less than 30 use one sample t test. The critical value of 't' statistics for 15 degrees of freedom and two tail at 1% level of significance is given as 2.95. Observed t value is -1.06 which is in between - 2.95 and +2.95. It can be concluded that sample mean for potency of cephalothin crystals does not differ significantly with manufacturers claim. So, the manufacturer claim is false about high potency of cephalothin crystals.

8. Initially determine sample mean (67.8) and standard deviation (3.01). As the population variance is unknown and also sample size is less than 30 use one sample t test. The critical value of 't' statistics for 9 degrees of freedom and two tail at 0.05 level of significance is given as 2.26. Observed t value is 1.89 which is in between - 2.26 and +2.26. It can be concluded that sample adequately represent a population in which the mean weight was found to be 66 kgs.

9. As the population variance is unknown and also sample size is less than 30 use one sample t test. The critical value of 't' statistics for 14 degrees of freedom and two tail at 0.05 level of significance is given as 2.14. Observed t value is -2.58 which is not between - 2.26 and +2.26. It can be concluded that sample of 15 individuals with a mean of 13.5 and SD 1.5 is not from same population.

10. Initially determine sample mean (48.4) and standard deviation (1.9). As the population variance is unknown and also sample size is less than 30, use one sample t test. The critical value of 't' statistics for 9 degrees of freedom and two tail at 0.05 level of significance is given as 2.26. Observed t value is -2.67 which is not in the range of -2.26 and +2.26. Reject null hypothesis and accept alternative hypothesis.



Chapter 19

HYPOTHESIS TESTING FOR TWO SAMPLE MEANS

Learning objectives

When we have finished this chapter, we should be able to;

1. Understand and estimate the z statistics for two independent data sets.
2. Understand and perform the t statistics for two independent data sets.
3. Understand and perform paired t test for matched data sets.

One of the most common statistical testing used in the pharmaceuticals is the examination of differences between two sets of sample data, i.e. whether or not the two populations whose properties are estimated by the sample statistics differ from one another.

There are two types of two sample statistical tests; independent and paired, as shown in following table:

Type	Parametric test	Non-parametric test
Independent samples	z test (Sample > 30) t test (sample < 30)	Chi-square test (sample > 20) Fisher exact test (sample < 20)
Paired samples	Paired t test	Mann Whitney U test (Metric) McNemar's test (Categorical) Wilcoxon signed rank test (Metric)

In the two sample test, the null hypothesis specifically states that there is no difference between the mean values of each set of data.

1. The z statistics for two independent samples

The z statistic can be calculated by using following formula

$$z = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)} \quad \dots 1$$

Where

\bar{X}_1 & \bar{X}_2 = sample means of two independent data sets

$SE(\bar{X}_1 - \bar{X}_2)$ is the standard error of difference between two means

The standard error of difference of two means can be calculated as

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \quad \dots 2$$

Where

σ_1, σ_2 = population standard deviations of samples N_1 and N_2 respectively

The following steps are used for calculating z statistic for two sample means:

1. Calculate the mean \bar{X}_1 and \bar{X}_2 and standard deviation (σ_1) and (σ_2) of each population.
2. Calculate the standard error of difference between two means $SE(\bar{X}_1 - \bar{X}_2)$.
3. Calculate z statistics using formula given above.
4. If the observed difference between the two means is greater than 1.96 times the standard error of difference, it is significant at 5% level of significance.
5. If the observed difference is greater than 3 times the SE, it is real variability in more than 99% cases, and biological or due to chance in less than 1% cases

Example 19.1

To determine whether two drugs affected human mental concentration equally, 50 students were given one drug and 50 others the second drug. All the students were then given an examination to measure their mental concentration index. The mean scores for the two groups were 65 and 70, and the respective standard deviations were 15 and 18. Is there sufficient evidence to suggest that the drugs affected mental concentration differently?

Solution: Let us test the difference between two groups systematically.

1. State the null hypothesis

The null hypothesis states that there is no difference between the effect of two drugs on mental concentration.

$$H_0: \mu_1 = \mu_2$$

2. State the alternative hypothesis

The alternative hypothesis states that there is a difference between the effect of two drugs on mental concentration.

$$H_a: \mu_1 \neq \mu_2$$

3. State the level of significance

It is assumed that the level of significance is 0.05.

4. State the number of tails

The number of tails associated with the alternative hypothesis is determined before data collection. As the effect produced by two drugs on mental concentration may differ to any side, we have to use two tailed test.

5. Select the appropriate statistical test

The most relevant statistical method to assess the validity of the null hypothesis, is a two-tailed z test for two independent samples. As the sample size is more than 30, the z test is the most suitable statistical method.

6. Perform the statistical analysis

Step 1. Calculate z statistic

The following steps are used for calculating z statistic

1. Calculate the mean \bar{X}_1 and \bar{X}_2 and standard deviation (σ_1) and (σ_2) for each population.

Here they have given.

$$\bar{X}_1 = 65 \quad \sigma_1 = 15 \quad \bar{X}_2 = 70 \quad \sigma_2 = 18$$

2. Calculate the standard error of difference between two means $SE(\bar{X}_1 - \bar{X}_2)$.

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{15^2}{50} + \frac{18^2}{50}} = \sqrt{10.98} = 3.31$$

3. Calculate z statistics using formula given above.

$$z = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)} = \frac{65 - 70}{3.31} = \frac{-5}{3.31} = -1.51$$

Step 2. Find critical value of z statistics

The critical value of z statistics for two tailed test and at 0.05 level of significance, is given as ± 1.96 . Therefore it should lie in critical region $-1.96 < z < +1.96$ (Appendix I).

7. Decision

The observed z value (-1.51) is less than z critical (-1.96). It lies in critical region. Therefore, there is no reason to reject H_0 . The evidence would suggest that both drugs have the same effect on mental concentration.

Example 19.2

In a group of 196 adults the mean serum cholesterol was 180 mg/dl with a standard deviation of 42 mg/dl. In a comparable group of 144 adults, the mean serum cholesterol was 150 mg/dl with a standard deviation of 48 mg/dl. Is the difference in cholesterol level of the two classes statistically significant? **Solution:** Let us test the difference between two groups systematically.

1. State the null hypothesis

The null hypothesis states that there is no difference between the cholesterol level of the two groups.

$$H_0: \mu_1 = \mu_2$$

2. State the alternative hypothesis

The alternative hypothesis states that there is a difference between the cholesterol levels of

two classes.

$$H_a: \mu_1 \neq \mu_2$$

3. State the level of significance

It is assumed that the level of significance is 0.05.

4. State the number of tails

The number of tails associated with the alternative hypothesis is determined before data collection. As the cholesterol levels may differ to any side, we have to use two tailed test.

5. Select the appropriate statistical test

The most relevant statistical method to assess the validity of the null hypothesis, is a two-tailed z test for two independent samples. As the sample size is more than 30, the z test is the most suitable statistical method.

6. Perform the statistical analysis

Step 1. Calculate z statistic

The following steps are used for calculating z statistic

1. Calculate the mean \bar{X}_1 and \bar{X}_2 and standard deviation (σ_1) and (σ_2) for each population.

The data is already provided.

$$\bar{X}_1 = 180 \quad \sigma_1 = 42 \quad \bar{X}_2 = 150 \quad \sigma_2 = 48$$

2. Calculate the standard error of difference between two means $SE(\bar{X}_1 - \bar{X}_2)$.

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{42^2}{196} + \frac{48^2}{144}} = 5$$

3. Calculate z statistics using formula given above.

$$z = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)} = \frac{180 - 150}{5} = \frac{30}{5} = 6$$

Step 2. Find critical value of z statistics

The critical value of z statistics for two tailed test and at 0.05 level of significance, is given as ± 1.96 (Appendix I). Therefore it should lie in the critical region $-1.96 < z < +1.96$.

7. Decision

As observed z value (6) is more than z critical (1.96), reject null hypothesis and accept alternative hypothesis. The evidence would suggest that difference in cholesterol level of the two groups is statistically significant.

2. The t test for two independent samples

In the equation for the calculation of the z statistic the population variance is used, indicating that there is prior knowledge of this parameter. Conversely, in the calculation of the t statistic the variances of the populations from which the samples were drawn are estimated from the variances of

the sample data.

The t statistic can be calculated by using following formula,

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{SE(\bar{X}_1 - \bar{X}_2)} \quad \dots 3$$

Where,

\bar{X}_1 & \bar{X}_2 = mean values of the two sets of sample data.

$SE(\bar{X}_1 - \bar{X}_2)$ = the standard error of difference between two means.

The standard error of difference of two means can be calculated as,

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_p^2}{N_1} + \frac{S_p^2}{N_2}} \quad \dots 4$$

Where,

S_p^2 is calculated as pooled variance.

$$S_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \quad \dots 5$$

The following steps are used for calculating t statistics:

1. Calculate mean and standard deviation of each group.
2. Calculate the pooled variance from sample size and sample variances using following formula.

$$S_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

Where,

S_1 and S_2 are standard deviations of two sample groups.

3. Calculate the standard error of difference between two means by using formula.

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_p^2}{N_1} + \frac{S_p^2}{N_2}}$$

4. Calculate t statistic using formula.

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{SE(\bar{X}_1 - \bar{X}_2)} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_p^2}{N_1} + \frac{S_p^2}{N_2}}}$$

5. Find the degrees of freedom, which is $N_1 + N_2 - 2$.

6. For given degrees of freedom, find critical value from t table. If the calculated value is less than critical value, then the null hypothesis is accepted while if calculated t statistic exceeds then the null hypothesis is rejected.

Example 19.3

Two formulations of same drug was placed for stability testing under controlled storage conditions at 37 °C. After 3 months, samples of each formulation were removed and assayed individually. The analytical results are given in following table. Determine whether there is a difference in the mean assay of drug in each formulation following the period of storage.

Data:

Formulation 1 104.1, 108.2, 108.6, 100.8, 106.5, 101.0, 102.6, 99.2, 95.2, 100.8

Formulation 2 102.9, 99.6, 98.1, 104.2, 90.2, 101.0, 99.9, 89.5, 95.5, 98.6

Solution: This problem can be solved in following steps:

1. State the null hypothesis

The null hypothesis states that there is no difference between the mean of assay of two formulations.

$$H_0: \mu_1 = \mu_2$$

2. State the alternative hypothesis

The alternative hypothesis states that there is a difference between the mean of assay of two formulation

$$H_a: \mu_1 \neq \mu_2$$

3. State the level of significance

It is assumed that the level of significance is 0.05.

4. State the number of tails

The number of tails associated with the alternative hypothesis is determined before data collection. As there are two possible outcomes to the study, this is a two tailed experimental design.

5. Select the appropriate statistical test

As the population variances are not known and the sample size is less than 30, two sample t test is the appropriate choice.

6. Perform the statistical analysis

1. Calculate mean and standard deviation of each group.

Calculation of mean

$$\bar{X} = \frac{\sum(X)}{N}$$

$$\text{Mean of formulation 1} = \bar{X}_1 = 102.7$$

$$\text{Mean of formulation 2} = \bar{X}_2 = 97.95$$

Calculation of standard deviation

$$SD = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}}$$

$$\text{Standard deviation of formulation 1} = s_1 = 4.21$$

Standard deviation of formulation 2 = $s_2 = 4.91$

2. Calculate the pooled standard deviation from sample size and sample SD of two groups using following formula

$$S_p = \frac{(N_1 - 1)s_1 + (N_2 - 1)s_2}{N_1 + N_2 - 2} = \frac{(9 \times 4.21) + (9 \times 4.91)}{20 - 2} = 4.56$$

3. Calculate the standard error of difference between two means by using formula

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_p^2}{N_1} + \frac{S_p^2}{N_2}} = S_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}} = 4.56 \sqrt{\frac{1}{10} + \frac{1}{10}} = 2.04$$

4. Calculate t statistic using formula

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{SE(\bar{X}_1 - \bar{X}_2)} = \frac{102.7 - 97.95}{2.04} = 2.32$$

5. Find the critical t values

The critical t values for 18 degrees of freedom, two tailed design and at 0.05 level of significance is given as ± 2.1 (Appendix II). Therefore the critical region of acceptance is $-2.1 < t < +2.1$.

5. Decision

The calculated t value more than critical value suggests rejection of null hypothesis. Therefore it can be concluded that there was significant difference in the assay of drug in two formulations after storage for three months at 37°C.

Example 19.4

Weight in kg of 10 boys and 10 girls aged between 15 to 20 are given below. Determine whether there is a significant difference in the mean weight of two groups.

Data:

Boys (Wt in kg)	42	46	50	48	50	52	41	49	51	56
Girls (Wt in kg)	38	41	36	35	30	42	31	29	31	35

Solution: This problem can be solved in following steps:

1. State the null hypothesis

The null hypothesis states that there is no difference between the mean weight of two groups.

$$H_0: \mu_1 = \mu_2$$

2. State the alternative hypothesis

The alternative hypothesis states that there is a difference between the mean weight of two

groups.

$$H_a: \mu_1 \neq \mu_2$$

3. State the level of significance

It is assumed that the level of significance is 0.05.

4. State the number of tails

The number of tails associated with the alternative hypothesis is determined before data collection. As there are two possible outcomes to the study, this is a two tailed experimental design.

5. Select the appropriate statistical test

As the population variances are not known and the sample size is less than 30, two sample t test is the appropriate choice.

6. Perform the statistical analysis

1. Calculate mean and standard deviation of each group.

Calculation of mean

$$\bar{X} = \frac{\sum(X)}{N}$$

Mean of weight of Boys = $\bar{X}_1 = 48.5$

Mean of weight of Girls = $\bar{X}_2 = 34.8$

Calculation of standard deviation

$$SD = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}}$$

Standard deviation of weight of Boys = $s_1 = 4.53$

Standard deviation of weight of Girls = $s_2 = 4.56$

2. Calculate the pooled standard deviation from sample size and sample SD of two groups using following formula

$$S_p = \frac{(N_1 - 1)s_1 + (N_2 - 1)s_2}{N_1 + N_2 - 2} = \frac{(9 \times 4.53) + (9 \times 4.56)}{20 - 2} = 4.54$$

3. Calculate the standard error of difference between two means by using formula

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_p^2}{N_1} + \frac{S_p^2}{N_2}} = S_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}} = 4.54 \sqrt{\frac{1}{10} + \frac{1}{10}} = 2.03$$

4. Calculate t statistic using formula

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{SE(\bar{X}_1 - \bar{X}_2)} = \frac{48.5 - 34.8}{2.03} = 6.7$$

5. Find the critical t values

The critical t values for 18 degrees of freedom, two tailed design and at 0.05 level of significance is given as ± 2.1 (Appendix II). Therefore calculated value should lie in acceptance region of $-2.1 < t < +2.1$.

5. Decision

The calculated t value more than critical value suggests rejection of null hypothesis. Therefore it may be concluded that there was significant difference between the mean weight of two groups.

3. Paired t test for two matched samples

The paired t test is a parametric statistical method that examines the significance of the mean of the differences between the pairs of data and a fixed value, determined from the null hypothesis.

It is applied to paired data of independent observations from one sample only when each individual gives a pair of observations. Testing by this method eliminates individual sampling variations because the sample is one and the observations on each person in the sample are taken before and after the experiment.

The following steps are used for calculating paired t statistics.

1. Find the difference in each set of paired observations before and after ($X_1 - X_2 = x$) and its squares.
2. Calculate the mean of the difference (\bar{x}).
3. Calculate SD of differences by using formula

$$SD = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}} \quad \dots 6$$

4. Determine 't' value by using following formula:

$$t = \frac{\bar{x} - O}{SE} = \frac{\bar{x}}{SD/\sqrt{N}} \quad \dots 7$$

As per null hypothesis, there should be no real difference in means of two sets of observations, i.e. theoretically it should be 0.

5. Find the degrees of freedom, which is $N-1$.
6. For given degrees of freedom, find critical value from t table. If the calculated value is less than critical value, then the null hypothesis is accepted while if calculated t statistic exceeds then the null hypothesis is rejected.

Example 19.5

Systolic blood pressure of 9 normal individuals was taken. Then a known hypotensive drug was given and blood pressure was again recorded. Did the hypotensive drug lowers the systolic blood pressure?

Data: Blood pressure of nine healthy volunteers before and after injection of hypotensive drug.

BP Before (X_1)	122	121	120	115	126	130	120	125	128
BPAfter (X_2)	120	118	115	110	122	130	116	124	125

Solution:

Let us test the hypothesis using following steps:

1. State the null hypothesis

The null hypothesis states that there is no difference between the mean of systolic BP of healthy volunteers before and after injection of hypotensive drug treatment.

$$H_0: \mu_{1 \text{ before}} = \mu_{2 \text{ after}}$$

2. State the alternative hypothesis

The alternative hypothesis states that the mean of systolic BP of healthy volunteers before injection of hypotensive drug treatment is greater than mean of systolic BP after injection of drug.

$$H_a: \mu_{1 \text{ before}} > \mu_{2 \text{ after}}$$

3. State the level of significance

It is assumed that the level of significance is 0.05.

4. State the number of tails

The number of tails associated with the alternative hypothesis is determined before data collection. As there is only one possible outcome to the study i.e. injection of drug lowers the blood pressure, this is a one tailed experimental design.

5. Select the appropriate statistical test

As the given experiment consists of data of independent observations from same sample giving pair of observations, paired t test is applied.

6. Perform the statistical analysis

1. Find the difference in each set of paired observations before and after ($X_1 - X_2 = x$) and its squares

BP Before (X_1)	122	121	120	115	126	130	120	125	128	
BP After (X_2)	120	118	115	110	122	130	116	124	125	
Difference (x)	2	3	5	5	4	0	4	1	3	$\Sigma x = 27$
Squares (x^2)	4	9	25	25	16	0	16	1	9	$\Sigma x^2 = 105$

2. Calculate the mean of the difference (\bar{x}).

Formula:

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

Data: $\Sigma x = 27$; $N = 9$

Mean of the difference (\bar{x}) = $27/9 = 3$

3. Calculate SD of differences by using formula

Formula:

$$SD = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}}$$

Data: $\sum X^2 = 105$; $\sum X = 27$; $N = 9$

$$SD = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}} = \sqrt{\frac{105 - \frac{(27)^2}{9}}{9-1}} = \sqrt{\frac{24}{8}} = 1.73$$

SD of differences = 1.73

4. Determine 't' value by using following formula:

$$t = \frac{\bar{x} - \bar{y}}{SE} = \frac{\bar{x} - \bar{y}}{SD/\sqrt{N}} = \frac{3 - 1}{1.73/\sqrt{9}} = \frac{2}{0.58} = 3.45$$

5. Find critical t value

For 8 degrees of freedom and at level of significance of 0.05 one tailed experiment value from appendix II gives ± 1.86 . Therefore, calculated t value should be in critical region $-1.86 < t < +1.86$.

7. Decision

The calculated t value is more than critical value and therefore it may be concluded that there was significant difference in blood pressure before and after administration of hypotensive drug. Reject the null hypothesis and accept the alternative hypothesis i.e. the mean of systolic BP of healthy volunteers before injection of hypotensive drug treatment is greater than mean of systolic BP after injection of drug.

Example 19.6

Serum digoxin levels were determined for nine healthy males following rapid intravenous injection of the drug. The measurements were made 4h after the injection and again at the end of an 8h period. Is the difference in the serum digoxin concentration at the end of 4h and at the end of 8h statistically significant?

Data: Serum digoxin concentration $\mu\text{g/ml}$ after 4h and 8h.

After 4h (X_1)	1.0	1.3	0.9	1.0	1.0	0.9	1.3	1.1	1.0
After 8h (X_2)	1.0	1.3	0.7	1.0	0.9	0.8	1.2	1.0	1.0

Answer: Let us test the hypothesis using following steps:

1. State the null hypothesis

The null hypothesis states that there is no difference between the mean serum concentration of digoxin after 4h and 8h.

$$H_0: \mu_{14h} = \mu_{28h}$$

2. State the alternative hypothesis

The alternative hypothesis states that the mean serum concentration of digoxin after 4h is greater than mean serum concentration of digoxin after 8h .

$$H_a: \mu_{14h} > \mu_{28h}$$

3. State the level of significance

It is assumed that the level of significance is 0.05.

4. State the number of tails

The number of tails associated with the alternative hypothesis is determined before data collection. As there is only one possible outcome to the study, this is one tailed experiment.

5. Select the appropriate statistical test

As the given experiment consists of data of independent observations from same sample giving pair of observations, paired t test is applied.

6. Perform the statistical analysis

1. Find the difference in each set of paired observations before and after ($X_1 - X_2 = x$) and its squares

BP Before (X_1)	1.0	1.3	0.9	1.0	1.0	0.9	1.3	1.1	1.0
BPAfter (X_2)	1.0	1.3	0.7	1.0	0.9	0.8	1.2	1.0	1.0
Difference (x)	0	0	0.2	0	0.1	0.1	0.1	0	$\Sigma x = 0.6$
Squares (x^2)	0	0	0.04	0	0.01	0.01	0.01	0	$\Sigma x^2 = 0.08$

2. Calculate the mean of the difference (\bar{x}).

Formula:

$$\bar{x} = \frac{\Sigma x}{N}$$

Data: $\Sigma x = 0.6$; $N = 9$

Mean of the difference (\bar{x}) = $0.6/9 = 0.067$

3. Calculate SD of differences by using formula

Formula:

$$SD = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N-1}}$$

Data: $\Sigma x^2 = 0.08$; $\Sigma x = 0.6$; $N = 9$

$$SD = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N-1}} = \sqrt{\frac{0.08 - \frac{(0.6)^2}{9}}{9-1}} = \sqrt{\frac{0.04}{8}} = 0.07$$

SD of differences = 0.07

4. Determine 't' value by using following formula:

$$t = \frac{\bar{x}}{SE} = \frac{\bar{x}}{SD/\sqrt{N}} = \frac{0.067}{0.07/\sqrt{9}} = \frac{0.067}{0.023} = 2.91$$

5. Find critical t value

For 8 degrees of freedom and at level of significance of 0.05, one tailed experiment value of t is given in appendix II as ± 1.86 . Therefore, calculated t values should be within $-1.86 < t < +1.86$.

7. Decision

The calculated t value is more than critical value and therefore it may be concluded that there was significant difference in mean serum digoxin concentration at the end of 4h and 8h. Reject the null hypothesis and accept the alternative hypothesis i.e. mean serum concentration of digoxin after 4h is greater than mean serum concentration of digoxin after 8h

Use of Microsoft Excel in Hypothesis Testing of Two Sample Means

1. Performing z test for two independent samples using Excel

Excel Solution: The systematic steps for calculating z test are given below:

Step I

1. Open new MS-Excel file from MS-office.
2. Enter data into Sheet1.
3. In first row put the labels for variables (i.e. formulation 1 in A1 and formulation 2 in B1 cell).
4. Enter data for formulation 1 in column 'A' (From A2 cell).
5. Enter data for formulation 2 into column 'B' (From B2 cell).

Step II

1. After entering data into the Sheet1, select Tools/Data Analysis/z-Test: Two Sample for means.
- a. Select Tools menu from the Menu bar of MS-Excel, Instantly, it will display pull down menus.
- b. Then, click on the Data Analysis option from pull down menus.
- c. When the Data Analysis dialog box appears: Choose z-Test: Two Sample for means and click on OK button.
- d. 'z-Test: Two Sample for means' dialog box will appear as shown below.

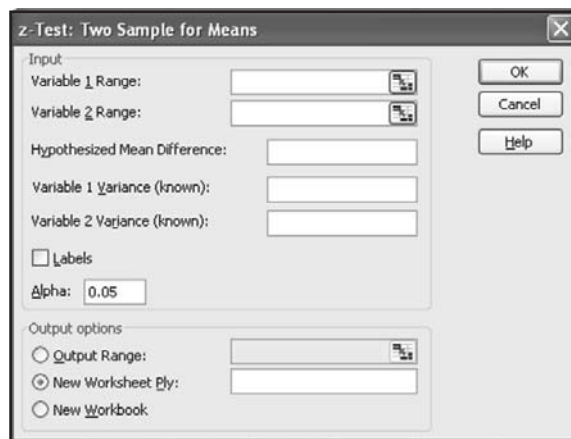


Figure 19.1 Window of z-Test: Two samples for means

2. Variable 1 Range - Enter the cell reference for the first range of data we want to analyze. The range must consist of a single column or row of data.
 3. Variable 2 Range - Enter the cell reference for the second range of data we want to analyze. The range must consist of a single column or row of data.
 4. Hypothesized Mean Difference - Enter the number we want for the shift in sample means. A value of 0 (zero) indicates that the sample means are hypothesized to be equal.
 5. Variable 1 Variance (known) - Enter the known population variance for the Variable 1 input range.
 6. Variable 2 Variance (known) - Enter the known population variance for the Variable 2 input range.
 7. Labels - Select if the first row or column of our input ranges contains labels. Clear this check box if our input ranges have no labels; Microsoft Excel generates appropriate data labels for the output table.
 9. Alpha - Enter the confidence level for the test. Generally it is 0.05.
 10. Leave the other items at their default selections. Click OK.
- This will give the result of z test for two independent samples.

2. Performing t test for two independent samples using Excel

Example

Two formulations of same drug was placed for stability testing under controlled storage conditions at 37 °C. After 3 months, samples of each formulation were removed and assayed individually. The analytical results are given in following table. Determine whether there is a difference in the mean assay of drug in each formulation following the period of storage.

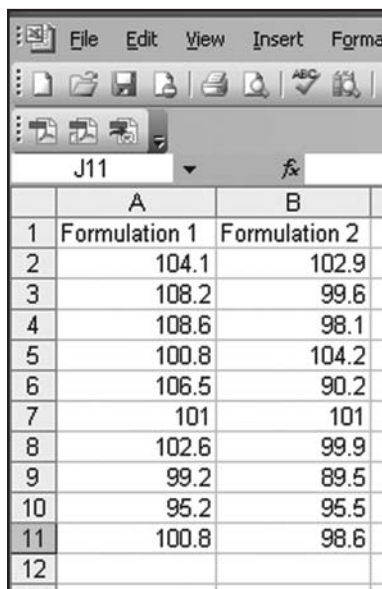
Data:

Formulation 1 104.1, 108.2, 108.6, 100.8, 106.5, 101.0, 102.6, 99.2, 95.2, 100.8

Formulation 2 102.9, 99.6, 98.1, 104.2, 90.2, 101.0, 99.9, 89.5, 95.5, 98.6

Excel Solution:**Step I**

1. Open new MS-Excel file from MS-office.
2. Enter data into Sheet1.
3. In first row put the labels for variables (i.e. in A1 and B1 cell).
4. Enter data for variable 1 i.e formulation 1 in column 'A'
5. Enter data for variable 2 i.e. formulation 2 into column 'B'.



	A	B
1	Formulation 1	Formulation 2
2	104.1	102.9
3	108.2	99.6
4	108.6	98.1
5	100.8	104.2
6	106.5	90.2
7	101	101
8	102.6	99.9
9	99.2	89.5
10	95.2	95.5
11	100.8	98.6
12		

Figure 19.2 Worksheet after data entry

Step II

1. After entering data into the Sheet 1 Select: Tools/ Data Analysis/ t-Test: Two Sample assuming Unequal Variances.
2. Then click on OK button.

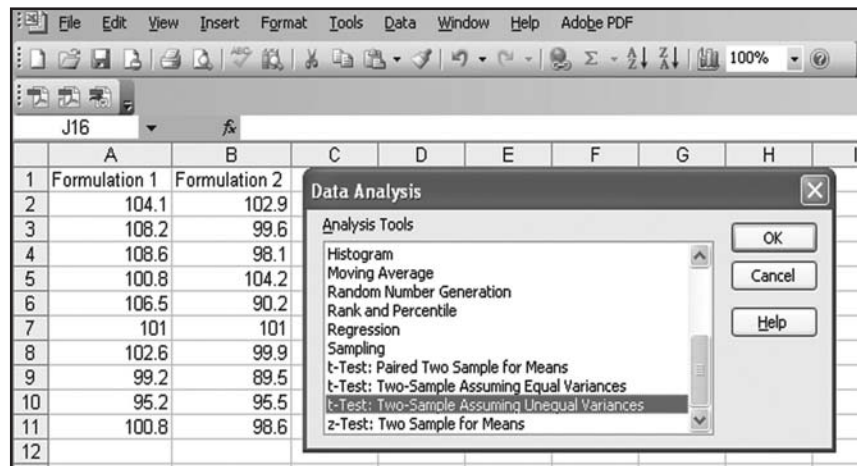


Figure 19.3 Window of Data Analysis

Step III

1. For the Input Range for Variable 1, Select A1:A11 (Label along with values).
2. For the input range for Variable 2, Select B1:B11 (Label along with values).
3. Click on Labels. Leave the other items at their default selections.

This dialog box is shown below. Click on OK button.

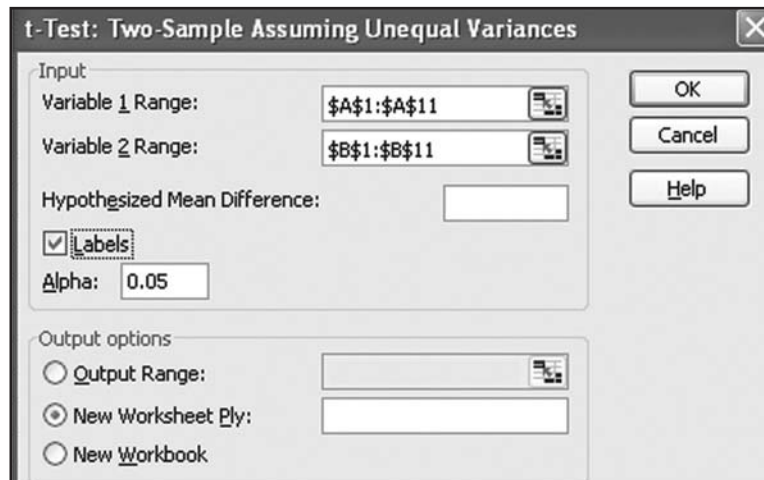


Figure 19.4 Window of t-Test: Two samples assuming unequal variances

Step IV

1. The following output is created in Sheet 4

t-Test: Two-Sample Assuming Unequal Variances

	Formulation 1	Formulation 2
Mean	102.7	97.95
Variance	17.78666667	24.145
Observations	10	10
Hypothesized Mean Difference	0	
df	18	
t Stat	2.319650459	
P(T<=t) one-tail	0.016157567	
t Critical one-tail	1.734063592	
P(T<=t) two-tail	0.032315134	
t Critical two-tail	2.100922037	

Note: Always assume unequal variance.

Interpretation of results

Depending on the hypothesis, level of significance and number of tails associated with design, one can interpret above obtained result for t test.

At 18 degrees of freedom two tail critical value is 2.1 which is less than observed value i.e. 2.32. Therefore, it can be concluded that there is a significant difference in the mean assay of drug in each formulation following the period of storage.

3. Performing Paired t-test for two sample means using Excel**Example**

Systolic blood pressure of 9 normal individuals was taken. Then a known hypotensive drug was given and blood pressure again recorded. Did the hypotensive drug lowers the systolic blood pressure?

Data: Blood pressure of nine healthy volunteers before and after injection of hypotensive drug.

BP Before (X_1)	122	121	120	115	126	130	120	125	128
BPAfter (X_2)	120	118	115	110	122	130	116	124	125

Excel Solution:**Step I**

1. Open new MS-Excel file from MS-office.
2. Enter data into Sheet1.
3. In first row put the labels for variables (i.e. in A1 and B1 cell).
4. Enter data for variable 1 i.e. 'BP Before' into column 'A'.
5. Enter data for variable 2 i.e. 'BPAfter' into column 'B'.

	A	B
1	BP Before	BP After
2	122	120
3	121	118
4	120	115
5	115	110
6	126	122
7	130	130
8	120	116
9	125	124
10	128	125
11		

Figure 19.5 Data entry

Step II

1. To perform a paired t-test, select Tools/ Data Analysis/ t-test: Paired two sample for means.
2. In the dialog box t-test: Paired two sample for means, we have to provide necessary information.
 - a. The Input Range for Variable 1, i.e BP Before, is given as A1:A10.
 - b. The input range for Variable 2, i.e BP Before, is given as B1:B10.
 - c. If selection includes labels in first row: Tick mark Labels.
 - d. For now, leave the other items at their default selections.
3. The dialog box is shown below. Click OK button.

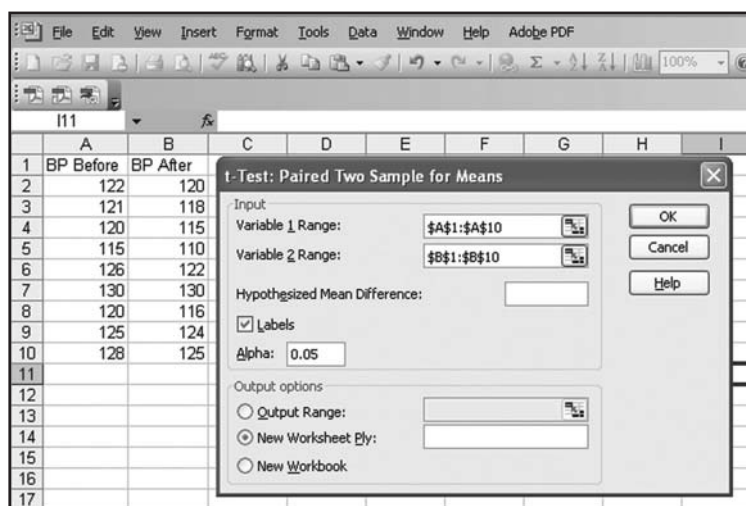


Figure 19.6 Window of t-Test: Paired Two Sample for means

Step III

1. The following output is created in Sheet 4.

t-Test: Paired Two Sample for Means

	BP Before	BP After
Mean	123	120
Variance	21.75	36.25
Observations	9	9
Pearson Correlation	0.979375	
Hypothesized Mean Difference	0	
df	8	
t Stat	5.196152	
P(T<=t) one-tail	0.000413	
t Critical one-tail	1.859548	
P(T<=t) two-tail	0.000826	
t Critical two-tail	2.306004	

Interpretation of Results

Depending on the hypothesis, level of significance and number of tails associated with design, one can interpret above obtained result for paired t test.

At (N-1) = 8 degrees of freedom, 0.05 level of significance critical t value for one tail is 1.86. The calculated t value (5.19) is more than critical value (1.86) and therefore it may be concluded that there was significant difference in blood pressure before and after administration of hypotensive drug. Therefore, reject the null hypothesis and accept the alternative hypothesis that the mean systolic BP of healthy volunteers decreases after hypotensive drug injection.

Summary**Statistical tests for two samples**

Type	Parametric test	Non-parametric test
Independent samples	z test (Sample > 30) t test (sample < 30)	Chi-square test (sample > 20) Fisher exact test (sample < 20) Mann Whitney U test (Metric)
Paired samples	Paired t test	McNemar's test (Categorical) Wilcoxon signed rank test (Metric)

z test for two independent samples

$$z = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)} \quad SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

t test for two independent samples

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{SE(\bar{X}_1 - \bar{X}_2)} \quad SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_p^2}{N_1} + \frac{S_p^2}{N_2}} \quad S_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

Paired t test for two matched samples

$$t = \frac{\bar{x} - O}{SE} = \frac{\bar{x} - O}{SD/\sqrt{N}} \quad SD = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}}$$

Multiple Choice Questions

- Which of the following is used to test significance between two independent samples?
 - z test
 - t test
 - z and t test
 - None of above
- For sample less than 30, which of the following is used to test significance between two independent samples?
 - z test
 - t test
 - z and t test
 - Paired t test
- For sample more than 30, which of the following is used to test significance between two independent samples?
 - z test
 - t test
 - z and t test
 - Paired t test
- Which of the following is used to test significance between paired samples?
 - z test
 - t test
 - z and t test
 - Paired t test
- For categorical data, non-parametric test equivalent to paired t test is _____.
 - Mann Whitney U test
 - Wilcoxon signed rank test
 - Chi-square test
 - McNemar's test
- In t test, the degree of freedom is given by _____.
 - $N - 1$
 - $N_1 - N_2 + 2$
 - $N_1 + N_2 - 2$
 - $N_1 + N_2 + 2$
- In paired t test, the degree of freedom is given by _____.
 - $N - 1$
 - $N_1 - N_2 + 2$
 - $N_1 + N_2 - 2$
 - $N_1 + N_2 + 2$
- A critical region defined by z statistic is _____ region for null hypothesis.
 - acceptance
 - rejection
 - no relation
 - can't say
- Non parametric test used for comparing two independent metric data sets is _____.
 - Chi-square test
 - Fisher exact test
 - Mann Whitney U test
 - McNemar's test
- Non parametric test used for comparing two independent categorical samples ($n < 20$) is _____.
 - Chi-square test
 - Fisher exact test

c. Mann Whitney U test

d. McNemar's test

Exercise

1. In a group of 169 boys in the age group of 12-20 years the mean height was 168 cm with a standard deviation of 14 cm. In a comparable group of 54 girls the mean height was 153 cm with a standard deviation of 8 cm. Is the height differs with sex?

2. In a group of 100 infants of 8 month age, the mean weight was 6.9 kg with a standard deviation of 1.1 kg. In a comparable group of 169 infants the mean weight was 7.3 kg with a standard deviation of 0.91 kg. Test whether mean weights are significantly different.

3. The mean plasma potassium level for 50 adult males with a disease was found to be 3.35 mEq/l with a standard deviation of 0.5 mEq/l. The normal 50 adult male value for plasma potassium is 4.6 mEq/l with a standard deviation of 0.01 mEq/l. Based on above data, can it be concluded that males with disease have lower plasma potassium levels than normal males?

4. Mean systolic blood pressure of 54 normal adults was 75 with a standard deviation of 6. In 31 diseased adults, mean systolic blood pressure was 69 with a standard deviation of 5. Test whether mean systolic blood pressure of two groups differ significantly?

5. In a nutritional study, 13 children (group A) were given a usual diet plus vitamin A and D tablets while the second group (B) of 12 children was taking the usual diet. After 24 months, the gain in weight in kg was noted as given in table below. Can we say that vitamins A and D were responsible for this difference.

Group A	5	3	4	3	2	6	3	2	3	6	7	5	3
Group B	1	3	2	4	2	1	3	4	3	2	2	3	

6. In the experiment, there are two groups of 15 subjects. For the first group, each individual was pre-treated with oral rifampicin (600 mg daily for 10 days). The other group acted as a control, receiving only placebo pre-treatment. All subjects then received an intravenous injection of theophylline (3 mg/kg). A series of blood samples was obtained after the theophylline injections, and analysed for drug content. The efficiency of removal of theophylline was reported as a clearance value. Is increase in clearance of theophylline would be due to the fact that rifampicin may increase the ability of the liver to eliminate this drug? Clearance of theophylline (ml/min/kg) for control subjects and for those pre-treated with rifampicin are given below.

Control	0.81	1.06	0.43	0.54	0.68	0.56	0.45	0.88	0.73	0.43
Treated	1.15	1.28	1.00	0.95	1.06	1.15	0.72	0.79	0.67	1.21
Control	0.46	0.43	0.37	0.73	0.93					
Treated	0.92	0.67	0.76	0.82	0.82					

7. Two groups of hypertensive patients are subjected to two different treatment regimens and systolic blood pressure was recorded after specific time period. Do the results of systolic BP listed below indicate a significant difference between the two therapies at 95% confidence level?

Group I	78	87	75	88	91	82	87	65	80		
Group II	75	88	93	86	84	71	91	79	81	86	89

8. A new natural product of company reported to promote weight loss. To ensure the claim, a clinical study of developed tablet was conducted on ten obese volunteers. Initial weight of volunteers before clinical trial and weight after receiving therapy for 2 months were recorded. Whether the natural product was clinically successful?

Weight before (kg)	141	124	153	120	116	155	151	155	132	116
Weight after (kg)	130	120	149	120	109	150	151	154	125	110

9. The systolic blood pressures of 12 women between the ages of 20 and 35 were measured before and after administration of a newly developed oral contraceptive. Is this increase in systolic blood pressure due to oral contraceptive?

Before	122	126	132	120	142	130	142	137	128	132	128	129
After	127	128	140	119	145	130	148	135	129	137	128	133

10. The extent to which an infant's health is affected by parental smoking is an important public health concern. The following data are the urinary concentrations of cotinine (a metabolite of nicotine); measurements were taken both from a sample of infants who had been exposed to household smoke and from a sample of unexposed infants. Comment.

Unexposed	8	11	12	14	20	43	111	
Exposed	35	56	83	92	128	150	176	208

Answers:

Multiple Choice Questions

1. c 2. b 3. a 4. d 5. d 6. c 7. a 8. a 9. c 10. b

Exercise

1. Yes, height differs with sex. (SE= 1.53; z=9.8)
2. Yes, mean weights are significantly different. (SE=0.13; z=-3.07)
3. It be concluded that males with disease have significantly lower plasma potassium levels than normal males. (SE=0.07; z=-17.67)
4. Yes, mean systolic blood pressure of two groups differ significantly. (SE=1.21; z= 4.94)
5. The calculated t value (2.74) is more than critical value (At 23 df, t critical for two tailed test at 5% level of significance is 2.07) and therefore it can be concluded that vitamins A and D were responsible

for the difference in increase of weight in two groups.

6. The calculated t value (-3.09) is more than critical value (At 28 df, t critical for two tailed test at 5% level of significance is 2.05) and therefore it can be concluded that increase in clearance of theophylline may be due to rifampicin increasing the ability of the liver to eliminate this drug.

7. The calculated t value (-0.74) is less than critical value (At 18 df, t critical for two tailed test at 5% level of significance is 2.1) and therefore it can be concluded that two treatment regimes do not differ significantly.

8. The calculated t value (4.06) is more than critical value (At 9 df, t critical for two tailed test at 5% level of significance is 2.26) and therefore it can be concluded that natural product was clinically successful.

9. The calculated t value (-2.89) is more than critical value (At 11 df, t critical for two tailed test at 5% level of significance is 2.2) and therefore it can be concluded that increase in systolic blood pressure is due to oral contraceptive.

10. The calculated t value (-3.228) is more than critical value (At 13 df, t critical for two tailed test at 5% level of significance is 2.16). Hence, it can be concluded that the infants exposed to smoking are more affected.



Chapter 20

ONE WAY ANALYSIS OF VARIANCE (ANOVA)

Learning objectives

When we have finished this chapter, we should be able to:

1. Understand meaning of ANOVA.
2. Calculate ANOVA by definitional formula.
3. Calculate ANOVA by computational formula.

What is ANOVA?

The ANOVA is used to identify and measure sources of variation within a collection of observations, hence the name analysis of variance. Analysis of variance is a parametric statistical technique that has found extensive applications in scientific research, mainly because of its flexibility. This method may be employed to analyse both paired and independent data and also is used to simultaneously compare large number of variables.

The one-way ANOVA is nothing more than an expansion of the t-test to more than two groups of sample. The analysis of variance involves determining if the observed values belong to the same population, regardless of the group, or whether the observations in at least one of these groups come from a different population.

$$H_0 : \mu_1 = \mu_2 = \mu_3 \dots = \mu_k = \mu$$

To obtain a F value we need two estimates of the population variance. It is necessary to examine the variability (analysis of variance) of observations within groups as well as between groups. The F statistic is computed using a simplified ratio similar to the t-test.

$$F = \frac{\text{Mean squared between (MSB)}}{\text{Mean squared within (MSW)}} \quad \dots 1$$

To calculate the F-statistic for the decision rule either the definitional or computational formulas may be used. With the exception of rounding errors, both methods will produce the same results. In the former case the sample means and standard deviations are used.

Calculations of ANOVA using definitional formula

The following steps are utilised for calculating ANOVA

1. Calculate the denominator of the F statistics, the mean squared within (MSW) in the same way as the pooled variance is calculated for t test.

$$MSW = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 \dots (n_k - 1)s_k^2}{N} \quad \dots 2$$

Where,

n_1, n_2, n_3 = number of observations in group 1, 2 and 3.

k = number of groups.

s_1, s_2, s_3 = standard deviation of group 1, 2 and 3.

N = total number of observations in all groups.

2. Determine the grand mean or pooled mean by using following formula,

$$\bar{X}_G = \frac{(n_1 \bar{X}_1) + (n_2 \bar{X}_2) + (n_3 \bar{X}_3) \dots (n_k \bar{X}_k)}{N} \quad \dots 3$$

\bar{X} = mean

3. Now, the mean squared between (MSB) is calculated similar to a sample variance by squaring the difference between each sample mean and the grand mean, and multiplying by the number of observations associated with each sample mean

$$MSB = \frac{n_1(\bar{X}_1 - \bar{X}_G)^2 + n_2(\bar{X}_2 - \bar{X}_G)^2 + n_3(\bar{X}_3 - \bar{X}_G)^2 \dots n_k(\bar{X}_k - \bar{X}_G)^2}{K - 1} \quad \dots 4$$

4. Finally, F statistics is calculated by formula

$$F = \frac{MSB}{MSW}$$

5. Compare the calculated F value with critical value from F table (Appendix III) and take decision.

6. The greater the spread of the sample observations, the larger the denominator and the smaller the calculated statistic, and thus the lesser the likelihood of rejecting H_0 . The greater the differences between the means, the larger the numerator, the larger the calculated statistic, and the greater the likelihood of rejecting H_0 in favor of H_a .

Example 20.1

The rate of release of three controlled release formulation of Diclofenac sodium after 2hr in % are given in following table. Is there a difference between the release kinetics of these three formulations?

Formulation 1	4.21	5.23	4.01	6.00	5.25	6.41	4.52	4.18	6.05	4.66
Formulation 2	3.69	3.99	4.25	4.08	5.22	2.99	5.66	4.25		
Formulation 3	5.00	6.55	6.02	6.11	4.88	4.29	6.00	5.45	5.04	

Answer: Let us perform ANOVA following the steps given below

1. State the null hypothesis

In this experiment null hypothesis states that the mean rate of release of diclofenac sodium from three formulations is identical.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

2. State the alternative hypothesis

The mean rate of release of diclofenac sodium from the three formulations is not identical.

$$H_a : \mu_1 \neq \mu_2 \neq \mu_3$$

3. State the level of significance

It is assumed that the level of significance is 0.05.

4. State the number of tails

As experiment involves a multiple hypothesis test, so the outcome is two tailed.

5. Select the appropriate statistical test

This is a case where more than two means will be simultaneously compared and hence multiple hypothesis test is required.

There is a one factor i.e. formulation of which there are three sub categories F1, F2 and F3, so one way ANOVA is required.

6. Perform the statistical test

1. Calculate the denominator of the F statistics, the mean squared within (MSW) in the same way as the pooled variance is calculated for t test.

Formula:

$$MSW = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 \dots (n_k - 1)s_k^2}{N}$$

Data: $n_1 = 10$, $n_2 = 8$, $n_3 = 9$ $k = \text{number of groups} = 3$

Calculation for standard deviation

Formula:

$$SD = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}}$$

Data: $s_1 = \text{standard deviation of group 1} = 0.87$

$s_2 = \text{standard deviation of group 2} = 0.84$

$s_3 = \text{standard deviation of group 3} = 0.73$

$$MSW = \frac{(10 - 1)0.87^2 + (8 - 1)0.84^2 + (9 - 1)0.73^2}{27} = 0.57$$

2. Determine the grand mean or pooled mean by using following formula

Formula:

$$\bar{X}_G = \frac{(n_1 \bar{X}_1) + (n_2 \bar{X}_2) + (n_3 \bar{X}_3) \dots (n_k \bar{X}_k)}{N}$$

Calculation of mean

$$\bar{X} = \frac{\sum X}{n}$$

Data: $\bar{X}_1 = \text{mean} = 5.05$; $\bar{X}_2 = \text{mean} = 4.27$; $\bar{X}_3 = \text{mean} = 5.48$

$$\bar{X}_G = \frac{(10 \times 5.05) + (8 \times 4.27) + (9 \times 5.48)}{27} = \frac{133.99}{27} = 4.96$$

3. Now, the mean squared between (MSB) is calculated similar to a sample variance by squaring the difference between each sample mean and the grand mean, and multiplying by the number of observations associated with each sample mean

Formula:

$$MSB = \frac{n_1(\bar{X}_1 - \bar{X}_G)^2 + n_2(\bar{X}_2 - \bar{X}_G)^2 + n_3(\bar{X}_3 - \bar{X}_G)^2 \dots n_k(\bar{X}_k - \bar{X}_G)^2}{K - 1}$$

$$MSB = \frac{10(5.05 - 4.96)^2 + 8(4.27 - 4.96)^2 + 9(5.48 - 4.96)^2}{3 - 1} = 3.2$$

4. Finally, F statistics can be calculated by using formula

$$F = \frac{MSB}{MSW} = \frac{3.2}{0.59} = 5.35$$

5. Compare the calculated F value with critical value from F table (Appendix III) and take decision.

The level of significance is 0.05

The numerator degrees of freedom = number of groups - 1 = K - 1 = 3 - 1 = 2

The number of degrees of freedom of denominator = N - K = 27 - 3 = 24

The critical F value for given degrees of freedom is found to be 3.40 (Appendix III).

7. Decision

If calculated F value is less than the critical F value the null hypothesis is accepted, whereas if the calculated F value is more than or equal to critical F value, the null hypothesis is rejected in favour of alternative hypothesis. The observed F value is greater than the critical F value. It is concluded that there is a significant difference between the rates of release of diclofenac sodium from the three controlled release formulations.

Example 20.2

Four brands of cereal are compared to see if they produce significant weight gain in rats. Four groups of seven rats each were given a diet of the respective cereal brand. At the end of the experimental period, the rats were weighed and the weight was compared to the weight just prior to the start of the cereal diet. Determine whether each brand has a statistically significant effect on the amount of weight gain. The data are provided in the table below.

Brand A	9	7	8	8	7	8	8
Brand B	5	4	6	4	5	7	3
Brand C	2	1	1	2	2	3	2
Brand D	3	8	5	9	2	7	8

Answer:

Let us perform ANOVA following the steps given below

1. State the null hypothesis

In this experiment null hypothesis is the mean weight gain of rats from four brands of cereal

is identical.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

2. State the alternative hypothesis

The mean weight gain of rats from the four brands of cereal is not identical.

$$H_a : \mu_1 \neq \mu_2 \neq \mu_3$$

3. State the level of significance

It is assumed that the level of significance is 0.05.

4. State the number of tails

As experiment involves a multiple hypothesis test, so the outcome is two tailed.

5. Select the appropriate statistical test

This is a case where more than two means will be simultaneously compared and hence multiple hypothesis test is required.

There is a one factor i.e. brands of cereal, of which there are four sub categories A, B, C and D, so one way ANOVA is required.

6. Perform the statistical test

1. Calculate the denominator of the F statistics, the mean squared within (MSW) in the same way as the pooled variance is calculated for t test.

Formula:

$$MSW = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 \dots (n_k - 1)s_k^2}{N}$$

Data: $n_1=7, n_2=7, n_3=7, n_4=7$

$k = \text{number of groups} = 4$

Calculation for standard deviation

Formula:

$$SD = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}}$$

Data: $s_1 = \text{standard deviation of group A} = 0.69$

$s_2 = \text{standard deviation of group B} = 1.34$

$s_3 = \text{standard deviation of group C} = 0.69$

$s_4 = \text{standard deviation of group D} = 2.7$

$$MSW = \frac{(7-1)0.69^2 + (7-1)1.34^2 + (7-1)0.69^2 + (7-1)2.7^2}{28} = 2.16$$

2. Determine the grand mean or pooled mean by using following formula

Formula:

$$\bar{X}_G = \frac{(n_1 \bar{X}_1) + (n_2 \bar{X}_2) + (n_3 \bar{X}_3) \dots (n_k \bar{X}_k)}{N}$$

Calculation of mean

$$\bar{X} = \frac{\sum X}{n}$$

Data: $\bar{X}_1 = \text{mean} = 7.86$; $\bar{X}_2 = \text{mean} = 4.86$; $\bar{X}_3 = \text{mean} = 1.86$; $X_4 = \text{mean} = 6.0$

$$\bar{X}_G = \frac{(7 \times 7.86) + (7 \times 4.86) + (7 \times 1.86) + (7 \times 6)}{28} = 5.14$$

3. Now, the mean squared between (MSB) is calculated similar to a sample variance by squaring the difference between each sample mean and the grand mean, and multiplying by the number of observations associated with each sample mean

Formula:

$$\text{MSB} = \frac{n_1(\bar{X}_1 - \bar{X}_G)^2 + n_2(\bar{X}_2 - \bar{X}_G)^2 + n_3(\bar{X}_3 - \bar{X}_G)^2 \dots n_k(\bar{X}_k - \bar{X}_G)^2}{K - 1}$$

$$\text{MSB} = \frac{7(7.86 - 5.14)^2 + 7(4.86 - 5.14)^2 + 7(1.86 - 5.14)^2 + 7(6 - 5.14)^2}{4 - 1} = 44.28$$

4. Finally, F statistics can be calculated by using formula

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{44.28}{2.16} = 20.47$$

5. Compare the calculated F value with critical value from F table (Appendix III) and take decision.

The level of significance is 0.05

The numerator degrees of freedom = number of groups - 1 = K - 1 = 4 - 1 = 3

The number of degrees of freedom of denominator = N - K = 28 - 4 = 24

The critical F value for given degrees of freedom is found to be 3.01 (Appendix III).

7. Decision

If calculated F value is less than the critical F value the null hypothesis is accepted, whereas if the calculated F value is more than or equal to critical F value, the null hypothesis is rejected in favour of alternative hypothesis. The observed F value is greater than the critical F value. It is concluded that there is a significant difference between the mean weight gain by rats fed on four brands of cereal.

Calculation of ANOVA using computational formula**1. Calculate total sum of squares (SST)**

SST can be calculated using the formula

$$SST = \sum X^2 - \frac{(\sum X)^2}{N} \quad \dots 5$$

Where,

$\sum X^2$ = sum of squares of all observations.

$(\sum X)^2$ = square of summation of observations.

N = total number of observations.

2. Calculate between sum of square (SSB)

The between sum of squares is calculated using the following equation

$$SSB = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} \dots - \frac{(\sum X_T)^2}{N} \quad \dots 6$$

$(\sum X_1)^2$ = Square of summation of observations in group 1

$(\sum X_2)^2$ = Square of summation of observations in group 2

$(\sum X_3)^2$ = Square of summation of observations in group 3

$(\sum X_T)^2$ = Square of summation of total observations

n_1, n_2 and n_3 = Number of observations in group 1, 2 and 3 respectively

3. Calculate within sum of squares (SSW)

The within group sum of squares is calculated as difference between the total sum of squares and the between sum of squares.

$$SSW = SST - SSB$$

4. Calculate between mean sum of squares (MSB) and within mean sum of squares (MSW)

Between mean sum of squares groups, $MSB = SSB/Df$

Where,

Df = degree of freedom = $k - 1$ = (No. of groups - 1)

Within mean sum of squares group, $MSW = SSW/Df$

Where

Df = degree of freedom = No. of observations - No. of groups = $N - k$

SSW = within group sum of squares

5. Calculate F statistics

F statistics can be calculated by using formula

$$F = \frac{MSB}{MSW}$$

6. Prepare ANOVA Table

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	k-1	SSB	MSB	MSB/MSW
Within groups	N-k	SSW	MSW	
Total	N-1	SST		

Example 20.3

The rate of release of three controlled release formulation of Diclofenac sodium after 2 hr in % are given in following table. Is there a difference between the release kinetics of the three formulations?

Formulation 1	4.21	5.23	4.01	6.00	5.25	6.41	4.52	4.18	6.05	4.66
Formulation 2	3.69	3.99	4.25	4.08	5.22	2.99	5.66	4.25		
Formulation 3	5.00	6.55	6.02	6.11	4.88	4.29	6.00	5.45	5.04	

Answer: Let us perform ANOVA following the steps given below:

1. State the null hypothesis

In this experiment null hypothesis is the mean rate of release of diclofenac sodium from three formulation is identical.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

2. State the alternative hypothesis

The mean rate of release of diclofenac sodium from the three formulations is not identical.

$$H_a : \mu_1 \neq \mu_2 \neq \mu_3$$

3. State the level of significance

It is assumed that the level of significance is 0.05.

4. State the number of tails

As experiment involves a multiple hypothesis test, so the outcome is two tailed.

5. Select the appropriate statistical test

This is a case where more than two means will be simultaneously compared and hence multiple hypothesis test is required.

There is a one factor i.e. formulation, of which there are three sub categories F1, F2 and F3, so one way ANOVA is required.

6. Perform the statistical test**1. Calculate total sum of squares (SST)**

SST can be calculated using the formula

Formula:

$$SST = \sum X^2 - \frac{(\sum X)^2}{N}$$

Data: $\sum X^2$ = sum of squares of all observations $(\sum X)^2$ = square of summation of observations

N = total number of observations = 27

			Squared values		
F1	F2	F3	F1	F2	F3
4.21	3.69	5	17.72	13.62	25.00
5.23	3.99	6.55	27.35	15.92	42.90
4.01	4.25	6.02	16.08	18.06	36.24
6	4.08	6.11	36.00	16.65	37.33
5.25	5.22	4.88	27.56	27.25	23.81
6.41	2.99	4.29	41.09	8.94	18.40
4.52	5.66	6	20.43	32.04	36.00
4.18	4.25	5.45	17.47	18.06	29.70
6.05		5.04	36.60		25.40
4.66			21.72		
50.52	34.13	49.34	262.03	150.53	274.80

$$(\sum X)^2 = (50.52 + 34.13 + 49.34)^2$$

$$(\sum X)^2 = (133.99)^2 = 17953.3$$

$$\sum X^2 = (262.03 + 150.53 + 274.8)$$

$$\sum X^2 = 687.36$$

Therefore, the total sum of squares is

$$SST = \sum X^2 - \frac{(\sum X)^2}{N} = 687.36 - \frac{17953.3}{27} = 687.36 - 664.94 = 22.42$$

2. Calculate between sum of squares (SSB)**Formula:**

$$SSB = \frac{\sum (X_1)^2}{n_1} + \frac{\sum (X_2)^2}{n_2} + \frac{\sum (X_3)^2}{n_3} - \frac{\sum (X_T)^2}{N}$$

 $(\sum X_1)^2$ = Square of summation of observations in group 1 $(\sum X_2)^2$ = Square of summation of observations in group 2 $(\sum X_3)^2$ = Square of summation of observations in group 3 $(\sum X_T)^2$ = Square of summation of total observations

$$(\sum X_1)^2 = (50.52)^2 = 2552.27$$

$$(\sum X_2)^2 = (34.13)^2 = 1164.86$$

$$(\sum X_3)^2 = (49.34)^2 = 2434.44$$

$$(\sum X_r)^2 = 17953.3$$

$$n_1 = 10; n_2 = 8; n_3 = 9,$$

$$N = 27$$

Therefore, the between sum of squares is

$$SSB = \frac{2552.27}{10} + \frac{1164.86}{8} + \frac{2434.44}{9} - \frac{17953}{27} = 671.33 - 664.94 = 6.39$$

3. Calculate within group sum of squares (SSW)

$$SSW = SST - SSB$$

$$SSW = 22.42 - 6.39 = 16.03$$

4. Calculate between mean sum of squares (MSB) and within mean sum of squares (MSW)

Mean sum of squares between groups,

$$MSB = SSB/Df$$

$$Df = \text{degree of freedom} = \text{No. of groups} - 1 = k - 1 = 3 - 1 = 2$$

$$MSB = SSB/Df = 6.39/2 = 3.2$$

Mean sum of squares within group,

$$MSW = SSW/Df$$

$$Df = \text{degree of freedom} = \text{No. of observations} - k = 27 - 3 = 24$$

$$MSW = SSW/Df = 16.03/24 = 0.67$$

5. Calculate F statistics

F statistics can be calculated by using formula

$F = \text{Mean sum of squares between groups} / \text{Mean sum of squares within group}$

$$F = MSB/MSW$$

$$F = 3.2/0.67 = 4.78$$

6. Prepare ANOVA Table

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	2	6.39	3.20	4.78
Within groups	24	16.03	0.67	
Total	26	22.42		

7. Decision

If calculated F value is less than the critical F value the null hypothesis is accepted, whereas if the calculated F value is more than or equal to critical F value, the null hypothesis is rejected in favour of alternative hypothesis. The observed F value is greater than the critical F value (3.40). It is concluded that there is a significant difference between the rates of release of diclofenac sodium from the three controlled release formulations.

Example 20.4

Four brands of cereal are compared to see if they produce significant weight gain in rats. Four groups of seven rats each were given a diet of the respective cereal brand. At the end of the experimental period, the rats were weighed and the weight was compared to the weight just prior to the start of the cereal diet. Determine whether each brand has a statistically significant effect on the amount of weight gain. The data are provided in the table below.

Brand A	9	7	8	8	7	8	8
Brand B	5	4	6	4	5	7	3
Brand C	2	1	1	2	2	3	2
Brand D	3	8	5	9	2	7	8

Answer:

Let us perform ANOVA following the steps given below

1. State the null hypothesis

In this experiment null hypothesis is the mean weight gain of rats from four brands of cereal is identical.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

2. State the alternative hypothesis

The mean weight gain of rats from the four brands of cereal is not identical.

$$H_a : \mu_1 \neq \mu_2 \neq \mu_3$$

3. State the level of significance

It is assumed that the level of significance is 0.05

4. State the number of tails

As experiment involves a multiple hypothesis test, so the outcome is two tailed.

5. Select the appropriate statistical test

This is a case where more than two means will be simultaneously compared and hence multiple hypothesis test is required.

There is a one factor i.e. brands of cereal of, which there are four sub categories A, B, C and D, so one way ANOVA is required.

6. Perform the statistical test**1. Calculate total sum of squares (SST)**

SST can be calculated using the formula

Formula:

$$SST = \sum X^2 - \frac{(\sum X)^2}{N}$$

Data: $\sum X^2$ = sum of squares of all observations $(\sum X)^2$ = square of summation of observations

N = total number of observations = 28

				Squared values			
A	B	C	D	A	B	C	D
9	5	2	3	81	25	4	9
7	4	1	8	49	16	1	64
8	6	1	5	64	36	1	25
8	4	2	9	64	16	4	81
7	5	2	2	49	25	4	4
8	7	3	7	64	49	9	49
8	3	2	8	64	9	4	64
55	34	13	42	435	176	27	296

$$(\sum X)^2 = (55+34+13+42)^2$$

$$\sum X^2 = (435+176+27+296)$$

$$(\sum X)^2 = (144)^2 = 20736$$

$$\sum X^2 = 934$$

Therefore, the total sum of squares is

$$SST = \sum X^2 - \frac{(\sum X)^2}{N} = 934 - \frac{20736}{28} = 934 - 740.57 = 193.43$$

2. Calculate between sum of squares (SSB)**Formula:**

$$SSB = \frac{\sum (X_1)^2}{n_1} + \frac{\sum (X_2)^2}{n_2} + \frac{\sum (X_3)^2}{n_3} + \frac{\sum (X_4)^2}{n_4} - \frac{\sum (X_T)^2}{N}$$

 $(\sum X_1)^2$ = Square of summation of observations in Brand A $(\sum X_2)^2$ = Square of summation of observations in Brand B $(\sum X_3)^2$ = Square of summation of observations in Brand C $(\sum X_4)^2$ = Square of summation of observations in Brand D $(\sum X_T)^2$ = Square of summation of total observations

$$(\sum X_1)^2 = (9+7+8+8+7+8+8)^2 = (55)^2 = 3025$$

$$(\sum X_2)^2 = (5+4+6+4+5+7+3)^2 = (34)^2 = 1156$$

$$(\sum X_3)^2 = (2+1+1+2+2+3+2)^2 = (13)^2 = 169$$

$$(\sum X_4)^2 = (3+8+5+9+2+7+8)^2 = (42)^2 = 1764$$

$$(\sum X_T)^2 = 20736$$

$$n_1 = 7, \quad n_2 = 7, \quad n_3 = 7, \quad n_4 = 7$$

$$N=28$$

Therefore, the between sum of squares is

$$SSB = \frac{3025}{7} + \frac{1156}{7} + \frac{169}{7} + \frac{1764}{7} - \frac{20736}{28} = 873.43 - 740.57 = 132.86$$

3. Calculate within group sum of squares (SSW)

$$SSW = SST - SSB$$

$$SSW = 193.43 - 132.86 = 60.47$$

4. Calculate between mean sum of squares (MSB) and within mean sum of squares (MSW)

Mean sum of squares between groups,

$$MSB = SSB/Df$$

$$Df = \text{degree of freedom} = \text{No. of groups} - 1 = k - 1 = 4 - 1 = 3$$

$$MSB = SSB/Df = 132.86/3 = 44.28$$

Mean sum of squares within group,

$$MSW = SSW/Df$$

$$Df = \text{degree of freedom} = \text{No. of observations} - k = 28 - 4 = 24$$

$$MSW = SSW/Df = 60.47/24 = 2.52$$

5. Calculate F statistics

F statistics can be calculated by using formula

$$F = \text{Mean sum of squares between groups} / \text{Mean sum of squares within group}$$

$$F = MSB/MSW$$

$$F = 44.28/2.52 = 17.57$$

6. Prepare ANOVA Table

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	3	132.86	44.28	17.57
Within groups	24	60.57	2.52	
Total	27	22.42		

7. Decision

If calculated F value is less than the critical F value the null hypothesis is accepted, whereas if the calculated F value is more than or equal to critical F value, the null hypothesis is rejected in favour of alternative hypothesis. The observed F value is greater than the critical F value (3.01). Therefore, it

is concluded that there is a significant difference between the mean weight gain by rats fed from four brands of cereal.

One Way ANOVA Using Microsoft Excel

The rate of release of three controlled release formulation of Diclofenac sodium after 2 hr in % are given in following table. Is there a difference between the release kinetics of the three formulations?

Formulation 1	4.21	5.23	4.01	6.00	5.25	6.41	4.52	4.18	6.05	4.66
Formulation 2	3.69	3.99	4.25	4.08	5.22	2.99	5.66	4.25		
Formulation 3	5.00	6.55	6.02	6.11	4.88	4.29	6.00	5.45	5.04	

Excel Solution

Step 1:

Open new file in MS-Excel as Book 1. Enter the data into an Excel datasheet (Sheet 1). Worksheet will appear as follows:

	A	B	C	D
1	Formulation 1	Formulation 2	Formulation 3	
2	4.21	3.69	5	
3	5.23	3.99	6.55	
4	4.01	4.25	6.02	
5	6	4.08	6.11	
6	5.25	5.22	4.88	
7	6.41	2.99	4.29	
8	4.52	5.66	6	
9	4.18	4.25	5.45	
10	6.05		5.04	
11	4.66			
12				

Figure 20.1 Worksheet after data entry

Step 2:

In MS-Excel, select Tools menu from Menu bar. Then, it will display pull down menus. From pull down menus, select Data Analysis option. Instantly, Data Analysis dialog box will appear.

Step 3:

Select Anova: Single Factor option from Analysis Tools.

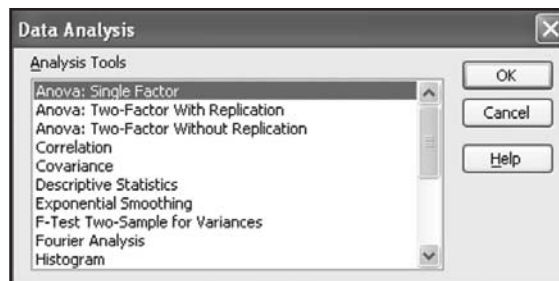


Figure 20.2 Window of Data Analysis

Step 4:

In the following Dialog box, enter the input range that corresponds to the data columns (\$A\$1:\$C\$11) and click OK. Check the option "Labels in First Row".

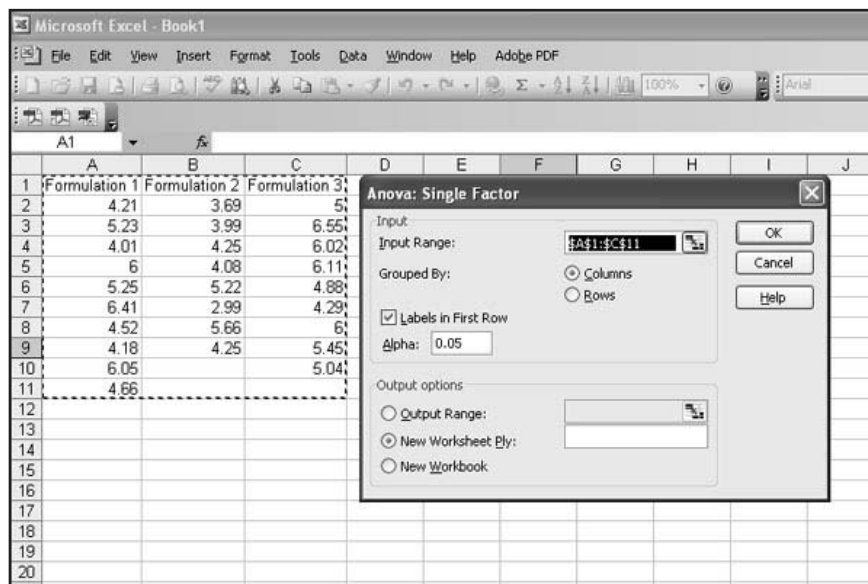


Figure 20.3 Window of Anova: Single Factor

The results appear in a new worksheet, as shown here:

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Formulation 1	10	50.52	5.052	0.755729
Formulation 2	8	34.13	4.26625	0.703513
Formulation 3	9	49.34	5.482222	0.538094

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	6.38922	2	3.194608	4.782674	0.01786	3.40283
Within Groups	16.0309	24	0.667954			
Total	22.4201	26				

In this output, the test statistic, F, is reported in the analysis of variance table, $F(2, 24) = 4.78$. The p-value for this statistics is reported in the table as 0.017. As observed F value is greater than

critical F value at 5 % level of significance, it can be concluded that there is a significant difference between the rates of release of diclofenac sodium from the three controlled release formulations.

Summary

An independent group ANOVA is an extension of the independent group t-test where we have more than two groups. This test is used to compare the means of more than two independent groups and is also called a One Way Analysis of Variance.

ANOVA using definitional formula

$$MSW = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 \dots (n_k - 1)s_k^2}{N}$$

$$\bar{X}_G = \frac{(n_1 \bar{X}_1) + (n_2 \bar{X}_2) + (n_3 \bar{X}_3) \dots (n_k \bar{X}_k)}{N}$$

$$MSB = \frac{n_1(\bar{X}_1 - \bar{X}_G)^2 + n_2(\bar{X}_2 - \bar{X}_G)^2 + n_3(\bar{X}_3 - \bar{X}_G)^2 \dots n_k(\bar{X}_k - \bar{X}_G)^2}{K - 1}$$

$$F = \frac{MSB}{MSW}$$

ANOVA using computational formula

$$SST = \sum X^2 - \frac{(\sum X)^2}{N}$$

$$SSB = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} - \frac{(\sum X_T)^2}{N}$$

$$SSW = SST - SSB$$

$$MSB = SSB/Df$$

$$MSW = SSW/Df$$

$$F = \text{Mean sum of squares between groups (MSB) / Mean sum of squares within group (MSW)}$$

Multiple Choice Questions

1. Which of the following is not a necessary assumption underlying the use of the Analysis of Variance technique?

- The samples are independent and randomly selected.
- The populations are normally distributed.
- The variances of the populations are the same.
- The means of the populations are equal.

2. A statistical test used to compare 2 or more group means is known as _____.
 - a. One-way analysis of variance
 - b. Post hoc test
 - c. t-test for correlation coefficients
 - d. Simple regression
3. One way ANOVA is extension of _____ to more than 2 groups
 - a. Chi square
 - b. t test
 - c. paired t test
 - d. z test
4. One way ANOVA is _____ test.
 - a. parametric
 - b. non-parametric
 - c. categorical
 - d. none of these
5. ANOVA is used to identify and measure _____.
 - a. sources of variation
 - b. location of mean
 - c. weighted mean of groups
 - d. none of these
6. The greater the spread of sample observations, _____ will be the denominator and smaller the calculated statistic.
 - a. smaller
 - b. larger
 - c. lesser
 - d. none of above
7. The greater the differences between the means, _____ will be the numerator and larger the calculated statistic.
 - a. smaller
 - b. larger
 - c. lesser
 - d. none of above
8. The greater the spread of sample observations, _____ the likelihood of rejecting null hypothesis.
 - a. greater
 - b. lesser
 - c. no relation
 - d. none of above
9. The greater the differences between the means, _____ the likelihood of rejecting null hypothesis.
 - a. greater
 - b. lesser
 - c. no relation
 - d. none of above
10. F statistic is also called as _____.
 - a. deviation ratio
 - b. mean ratio
 - c. mean squared
 - d. variance ratio

Exercise

1. During the manufacture of a diclofenac coated tablet, samples were periodically selected from production lines at three different facilities. Weights were taken for 10 tablets and their average weights listed in table. Is there any significant difference in weights of the tablets between the three facilities? F Critical at 2 degrees of freedom in numerator and 27 degrees of freedom in denominator is 3.35 at 5% level of significance.

Facility A	77.3	80.3	79.1	75.2	73.6	76.7	81.7	78.7	78.4	72.9
Facility B	71.6	74.8	71.2	77.6	74.5	75.7	76.1	75.9	75.5	74.0
Facility C	75.5	74.2	67.5	74.2	70.5	84.4	75.6	77.1	72.2	73.4

2. Four brands of diclofenac sodium were selected and assayed according to IP. Results of assay in terms of percent labeled amount of drug are listed in table. Was there any significant difference based on brands in terms of labeled claim? F Critical at 3 degrees of freedom in numerator and 36 degrees of freedom in denominator is 2.86 at 5% level of significance.

Brand A	100	99.8	99.5	100.1	99.7	99.9	100.4	100	99.7	99.9
Brand B	99.5	100	99.3	99.9	100.3	99.5	99.6	98.9	99.8	100.1
Brand C	99.6	99.3	99.5	99.1	99.7	99.6	99.4	99.5	99.5	99.9
Brand D	99.8	100.5	100	100.1	99.4	99.6	100.2	99.9	100.4	100.1

3. Four brands of Atenolol were under investigation for their antihypertensive efficacy. Four groups were selected comprising of 10 volunteers in each group. According to designed protocol drug was given to each group and mean systolic blood pressure was listed in table. Determine if there is significant difference in mean systolic blood pressure in order to assess the role of different brands of Atenolol. F Critical at 3 degrees of freedom in numerator and 36 degrees of freedom in denominator is 2.86 at 5% level of significance.

Brand A	125	130	135	120	115	120	130	135	140	135
Brand B	120	122	115	110	125	122	120	120	126	120
Brand C	120	115	115	130	120	125	122	115	126	118
Brand D	118	120	118	120	120	115	125	125	120	115

4. Four treatments are given to four groups of patients with anaemia. Increase in Hb% level was noted after one month. Find whether the difference in improvement in four groups is significant or not.

Group A	3	1	2	0	1	2	2
Group B	3	2	2	3	1	3	2
Group C	3	4	5	4	2	2	4
Group D	4	2	3	1	4	5	1

F Critical at 3 degrees of freedom in numerator and 24 degrees of freedom in denominator is 3.01 at 5% level of significance.

5. Pharmaceutical company suspected that four filling machines were not filling the bottles in a uniform way. An experiment on four machines were performed and recorded in table. Find whether there is a significant difference in the filling performance of four machines? F Critical at 3 degrees of freedom in numerator and 19 degrees of freedom in denominator is 3.24 at 5% level of significance.

Machine A	52.05	52.07	52.04	52.04	51.99
Machine B	51.98	52.05	52.06	52.02	51.99
Machine C	52.04	52.03	52.03	52.00	51.96
Machine D	52.00	51.97	52.98	51.99	51.96

6. Dissolution test was performed for different brands of paracetamol available in India. Dissolution efficiencies were calculated and reported in table. Find whether there is a significant difference in the dissolution efficiencies of six brands of paracetamol? F Critical at 5 degrees of freedom in numerator and 30 degrees of freedom in denominator is 2.53 at 5% level of significance.

Brand A	63.33	56.67	56.67	60.00	56.67	53.33
Brand B	63.33	63.33	60.00	66.67	60.00	73.33
Brand C	50.00	46.67	53.33	50.00	46.67	53.33
Brand D	53.33	56.67	46.67	50.00	45.00	46.67
Brand E	45.00	46.67	48.33	45.00	46.67	48.33
Brand F	50.00	53.33	50.00	46.67	56.67	60.00

7. The students are doing titrimetric estimation of certain compound in the laboratory. Three groups have reported their analysis results as given below.

Group A	18.0	16.4	15.7	19.6	16.5	18.2
Group B	21.1	17.8	18.6	20.8	17.9	19.0
Group C	16.1	17.8	16.5			

Perform ANOVA at 0.05 level of significance to test whether the differences among the sample means are significant.

8. Three chemicals A, B and C show the cleaning efficiency as given below. Find whether the differences among them are significant at 5% significance level.

Chemical A	80	77	76	81	71
Chemical B	70	58	72	66	74
Chemical C	77	80	82	85	76

9. If five different liquid filling machines fill 30 ml medicines into a container as shown below, find whether the differences amongst the five machines is significant at 0.01 significance level by performing ANOVA.

Machine A	30	25	27	26
Machine B	29	28	26	29
Machine C	37	32	32	35
Machine D	32	33	34	29
Machine E	31	26	27	32

10. In a pharmaceutical company, three different brands of lubricants were used during tableting. The quantity required for three brands is given below. Test at the 0.01 level of significance whether the differences among the three sample means are significant.

Lubricant A	8	14	10	10	13	12	13	12			
Lubricant B	11	7	6	8	9	11	8	9	12	8	9
Lubricant C	7		7		5		6		10	4	
										8	9

Answers:

Multiple Choice Questions

1. d 2. a 3. b 4. a 5. a 6. b 7. b 8. b 9. a 10. d

Exercise

1. The observed F value is less than the critical F value (3.35). It is concluded that there is a no significant difference in weights of the tablets between the three facilities.

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	2	53.09	26.55	2.49
Within groups	27	287	10.63	
Total	29	340.1		

2. The observed F value is more than the critical F value (2.86). It is concluded that there is a significant difference in terms of labeled claim of four brands.

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	3	1.437	0.479	4.81
Within groups	36	3.578	0.099	
Total	39	5.015		

3. The observed F value is more than the critical F value (2.86). It is concluded that there is a significant difference in mean systolic blood pressure produced by different brands of Atenolol.

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	3	535.47	179.49	5.65
Within groups	36	1143.3	31.75	
Total	39	1681.77		

4. The observed F value is more than the critical F value (3.01). It is concluded that there is a significant difference in mean %Hb in four groups.

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	3	13.25	4.41	3.34
Within groups	24	31.71	1.32	
Total	27	44.96		

5. The observed F value is less than the critical F value (3.24). It is concluded that there is no significant difference in the filling performance of four machines.

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	3	0.094	0.031	0.615
Within groups	16	0.81	0.051	
Total	19	0.91		

6. The observed F value is more than the critical F value (2.53). It is concluded that there is a significant difference in the dissolution efficiencies of six brands of paracetamol.

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	5	1199.83	239.96	15.43
Within groups	30	466.47	15.54	
Total	35	1666.3		

7. The observed F value is more than the critical F value (3.89). It is concluded that there is a significant difference among sample means of three groups.

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	2	15.12	7.58	4.06
Within groups	12	22.34	1.86	
Total	14	37.46		

8. The observed F value is more than the critical F value (3.89). It is concluded that there is significant difference in the cleaning efficiency of chemicals A, B and C.

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	2	390	195	8.48
Within groups	12	276	23	
Total	14	666		

9. The observed F value is more than the critical F value (6.42). It is concluded that there is a significant difference in filling performance of five liquid filling machines at 0.01 level of significance.

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	4	136	34	6.54
Within groups	15	78	5.2	
Total	19	214		

10. The observed F value is more than the critical F value (5.61). It is concluded that there is a significant difference in three brands of lubricants at 0.01 level of significance.

Source	Degree of freedom	Sum of squares	Mean squares	F
Between groups	2	82	41	11.05
Within groups	24	89	3.70	
Total	26	171		



Chapter 21

CORRELATION

Learning objectives

When we have finished this chapter, we should be able to:

1. Understand meaning of correlation.
2. Understand types of correlation.
3. Calculate the correlation coefficient.

What is correlation?

The relationship between two metric continuous variables is called correlation. The easiest way to visualise the relationship between two continuous variables is graphically, using a scatter plot. One variable is labeled X and plotted on the x-axis of the graph or the abscissa. The second variable, Y is plotted on the vertical y-axis or the ordinate. The first role of correlation is to determine the strength of the relationship between the two variables represented on the x-axis and the y-axis. The measure of this magnitude is called the correlation coefficient. This index measures both the magnitude and the direction of the relationships:

- + 1.0 perfect positive correlation
- 0.0 no correlation
- 1.0 perfect negative correlation

Types of correlation

1. Perfectly Positive Correlation
2. Perfectly Negative Correlation
3. Moderately Positive Correlation
4. Moderately Negative Correlation
5. Absolutely no correlation

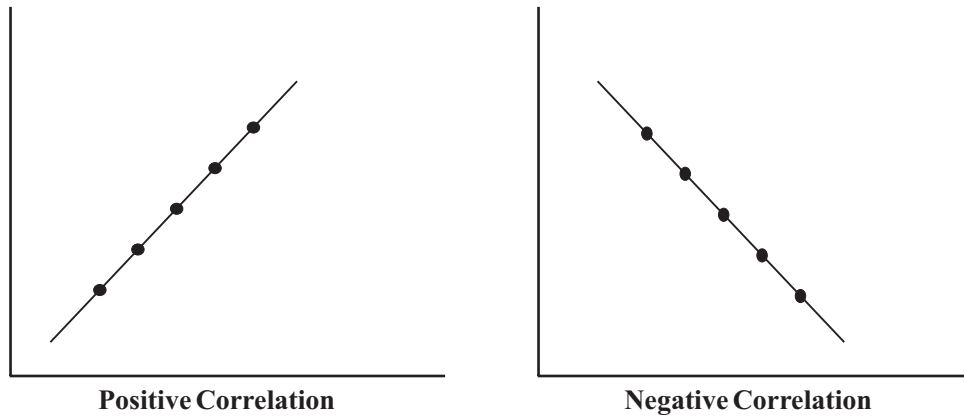
1. Perfectly Positive Correlation

In this the variables are directly proportional to each other. If one variable rises another also rises or if one variable falls another variable also falls. In a perfect +ve relationship (coefficient of +1), all of the data points would fall on a straight line. Here this straight line runs from the lower left of the scatter plot to upper right.

2. Perfectly Negative Correlation

If one variable increases, other variable decreases but the fall is directly proportional to each other. Here also the points fall along a straight line, but the straight line runs from upper left side to

lower right of the scatter plot. This is called as negative correlation.



3. Moderately Positive Correlation

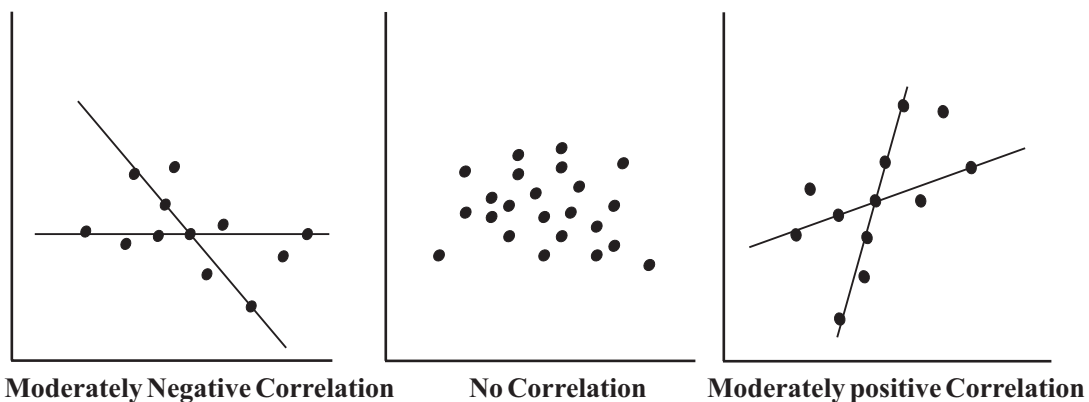
If the points do not fall on straight line but an orientation of points is from the lower left to the upper right, as shown in figure then the moderately positive correlation exists. Here the correlation coefficient ranges from 0 to +1.

4. Moderately Negative Correlation

If the points do not fall on straight line but an orientation of points is from the upper left to the lower right, as shown in figure, then the moderately negative correlation exists. Here the correlation coefficient ranges from 0 to -1.

5. Absolutely No Correlation

If the points fall within the circle there is no correlation. Here the two variables are not related to one another.



Correlation Coefficient

The strength of the relationship (correlation coefficient) can be calculated by using Pearson correlation coefficient for parametric data while Spearman rank correlation is used for non parametric procedures.

Significance of a correlation coefficient

A positive or negative correlation between two variables shows that a relationship exists. Whether it is strong or weak correlation, it can be roughly interpreted as given in following table.

< 20	Slight correlation
0.20 - 0.40	Low correlation
0.40 - 0.70	Moderate correlation
0.70 - 0.90	High correlation
> 90	Very high correlation

Calculation of Pearson Product Moment Coefficient (r) by definitional formula

The following mathematical formula is used for determining Pearson Correlation coefficient (r)

$$r = \frac{\sum (x - \bar{X})(y - \bar{Y})}{\sqrt{\sum (x - \bar{X})^2 \sum (y - \bar{Y})^2}} \quad \dots 1$$

Where

x = value of each measurement on x axis

y = value of each measurement on y axis

\bar{X} = Mean for variables on x axis

\bar{Y} = Mean for variables on y axis

1. Calculate the mean for both x variable (\bar{X}) and y variable (\bar{Y}).

$$\bar{X} = \frac{\sum x}{N} \quad \dots 2$$

$$\bar{Y} = \frac{\sum y}{N} \quad \dots 3$$

Where

x = value of each measurement on x axis

y = value of each measurement on y axis

N = number of observations

2. Now, calculate deviations of individual scores x and y about means; (x - \bar{X}) and (y - \bar{Y}).
3. Then, multiply these deviations (x - \bar{X}) (y - \bar{Y}) and find the sum of the product of these deviations $\sum (x - \bar{X})(y - \bar{Y})$. This is used in the numerator of the equation.
4. The deviations (x - \bar{X}) of x variables and deviations (y - \bar{Y}) of y variables are squared, which

gives $\Sigma (x - \bar{X})^2 (y - \bar{Y})^2$

5. The table shown below is developed for computation of correlation coefficient.

Data layout for computation for Pearson Product Moment Correlation Coefficient formula

x	y	(x- \bar{X})	(y- \bar{Y})	(x- \bar{X})(y- \bar{Y})	(x- \bar{X}) ²	(y- \bar{Y}) ²
x ₁	y ₁
x ₂	y ₂
x ₃	y ₃
...
x _n	y _n
				$\Sigma (x - \bar{X})(y - \bar{Y})$	$\Sigma (x - \bar{X})^2$	$\Sigma (y - \bar{Y})^2$

6. Calculate, correlation coefficient, r by using following formula

$$r = \frac{\Sigma (x - \bar{X})(y - \bar{Y})}{\sqrt{\Sigma (x - \bar{X})^2 \Sigma (y - \bar{Y})^2}}$$

Example 21.1

Samples of drug products are stored in their original containers under normal conditions and sampled periodically to analyse the content of the medication. Determine correlation coefficient.

Data:

Time (months)	6	12	18	24	36	48
Content (mg)	995	984	973	960	952	948

Solution:

1. Construct the following table

x	y	(x- \bar{X})	(y- \bar{Y})	(x- \bar{X})(y- \bar{Y})	(x- \bar{X}) ²	(y- \bar{Y}) ²
6	995	-18	26.33	-474	324	693.44
12	984	-12	15.33	-184	144	235.11
18	973	-6	4.33	-26	36	18.78
24	960	0	-8.67	0	0	75.11
36	952	12	-16.67	-200	144	277.78
48	948	24	-20.67	-496	576	427.11
144	5812	0	00	-1380	1224	1727.33
Σx	Σy	$\Sigma (x - \bar{X})$	$\Sigma (y - \bar{Y})$	$\Sigma (x - \bar{X})(y - \bar{Y})$	$\Sigma (x - \bar{X})^2$	$\Sigma (y - \bar{Y})^2$

2. Calculate the mean for both x variable (\bar{X}) and y variable (\bar{Y}).

$$\bar{X} = 144/6 = 24$$

$$\bar{Y} = 5812/6 = 968.66$$

$$\Sigma (x - \bar{X})(y - \bar{Y}) = -1380$$

$$\Sigma (x - \bar{X})^2 = 1224$$

$$\Sigma (y - \bar{Y})^2 = 1727.33$$

3. Calculate, correlation coefficient, r by using following formula

$$r = \frac{\Sigma (x - \bar{X})(y - \bar{Y})}{\sqrt{\Sigma (x - \bar{X})^2 \Sigma (y - \bar{Y})^2}}$$

$$r = \frac{-1380}{\sqrt{1224 \times 1727.33}} = -0.9490$$

Correlation coefficient is **-0.949**.

Calculation of Pearson Correlation Coefficient, r by Computational Formula

Following mathematical formula is used for calculation of Pearson Correlation

$$r = \frac{n \Sigma xy - \Sigma x \Sigma y}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}} \quad \dots 4$$

Where

Σx = Summation of x

Σy = Summation of y

Σxy = Summation of product of x and y

$(\Sigma x)^2$ = Square of summation of x

Σx^2 = Summation of square of x

$(\Sigma y)^2$ = Square of summation of y

Σy^2 = Summation of square of y

n = number of pairs of data

The following steps should be used for calculating correlation coefficient by computational formula

1. Develop the table as shown below

x	y	x^2	y^2	xy
x_1	y_1
x_2	y_2
x_3	y_3
...
x_n	y_n
Σx	Σy	Σx^2	Σy^2	Σxy

2. Write the values of x variable in column 1 of the table while the values of y variable in column 2 of the table.
3. Write the individual x values squared (x^2) in column 3 while individual y values squared (y^2) in column 4 of the table.
4. Write the product of x and y for each data point in column 5 of the table.
5. Now using the formula for Pearson correlation, find the value of r.

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

Example 21.2

Samples of drug products are stored in their original containers under normal conditions and sampled periodically to analyse the content of the medication. Determine correlation coefficient.

Data:

Time (months)	6	12	18	24	36	48
Content (mg)	995	984	973	960	952	948

Solution:

1. Develop the table as shown below

x	y	x^2	y^2	xy
6	995	36	990025	5970
12	984	144	968256	11808
18	973	324	946729	17514
24	960	576	921600	23040
36	952	1296	906304	34272
48	948	2304	898704	45504
Σx	Σy	Σx^2	Σy^2	Σxy
144	5812	4680	5631618	138108

2. Calculate, correlation coefficient, r by using following formula

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

$$r = \frac{(6 \times 138108) - (144 \times 5812)}{\sqrt{(6 \times 4680 - (144)^2)} \sqrt{(6 \times 5631618 - (5812)^2)}}$$

$$r = -0.9490$$

Coefficient of correlation r is **-0.949**.

Example 21.3.

Following data were obtained for plotting calibration curve of drug. Determine correlation coefficient by both methods.

Data:

Concentration (mg/l)	5	10	15	20	25	30
Absorbance	0.11	0.2	0.31	0.4	0.51	0.62

1. Definitional formula**Solution:**

1. Construct the following table

x	y	(x- \bar{X})	(y- \bar{Y})	(x- \bar{X})(y- \bar{Y})	(x- \bar{X}) ²	(y- \bar{Y}) ²
5	0.11	-12.5	-0.248	3.1	156.25	0.0616
10	0.20	-7.5	-0.158	1.18	56.25	0.025
15	0.31	-2.5	-0.048	0.12	6.25	0.0023
20	0.40	2.5	0.042	0.10	6.25	0.0017
25	0.51	7.5	0.152	1.14	56.25	0.0023
30	0.62	12.5	0.262	3.27	156.25	0.0684
105	2.15	00	00	8.925	437.5	0.1822
Σx	Σy	$\Sigma (x- \bar{X})$	$\Sigma (y- \bar{Y})$	$\Sigma (x- \bar{X})(y- \bar{Y})$	$\Sigma (x- \bar{X})^2$	$\Sigma (y- \bar{Y})^2$

2. Calculate the mean for both x variable (\bar{X}) and y variable (\bar{Y}).

$$\bar{X} = 105/6 = 17.5$$

$$\bar{Y} = 2.15/6 = 0.358$$

$$\Sigma (x- \bar{X})(y- \bar{Y}) = 8.925$$

$$\Sigma (x- \bar{X})^2 = 437.5$$

$$\Sigma (y- \bar{Y})^2 = 0.1822$$

3. Calculate, correlation coefficient, r by using definitional formula

$$r = \frac{\Sigma (x- \bar{X})(y- \bar{Y})}{\sqrt{\Sigma (x- \bar{X})^2 \Sigma (y- \bar{Y})^2}}$$

$$r = \frac{8.925}{\sqrt{437.5 \times 0.1822}} = 0.999$$

Correlation coefficient is **0.999**.

Computational Formula**Solution:**

1. Develop the table as shown below

x	y	x ²	y ²	xy
5	0.11	25	0.0121	0.55
10	0.2	100	0.04	2
15	0.31	225	0.0961	4.65
20	0.4	400	0.16	8
25	0.51	625	0.2601	12.75
30	0.62	900	0.3844	18.6
Σx	Σy	Σx^2	Σy^2	Σxy
105	2.15	2275	0.9527	46.55

2. Calculate, correlation coefficient, r by using following formula

$$r = \frac{n \Sigma xy - \Sigma x \Sigma y}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}}$$

$$r = \frac{(6 \times 46.55) - (105 \times 2.15)}{\sqrt{(6 \times 2275 - (105)^2)} \sqrt{(6 \times 0.9527 - (2.15)^2)}}$$

$$r = 0.999$$

Coefficient of correlation r is **0.999**

Estimation of Correlation coefficient by using Excel**Example 21.1**

Samples of drug products are stored in their original containers under normal conditions and sampled periodically to analyse the content of the medication. Determine correlation coefficient.

Data:

Time (months)	6	12	18	24	36	48
Content (mg)	995	984	973	960	952	948

Solution:**Step 1:**

Open new file in MS-Excel as Book 1.

Enter the data into an Excel datasheet (Sheet 1).

Worksheet will appear as follows:

	A	B	
1	Time (X)	Content Remaining (Y)	
2	6	995	
3	12	984	
4	18	973	
5	24	960	
6	36	952	
7	48	948	

Figure 21.1 Data entry

Step 2:

In MS-Excel, select Tools menu from Menu bar. Then, it will display pull down menus. From pull down menus, select Data Analysis option. Instantly, Data Analysis dialog box will appear.

Step 3:

Select Correlation option from Analysis Tools and then click on Ok.

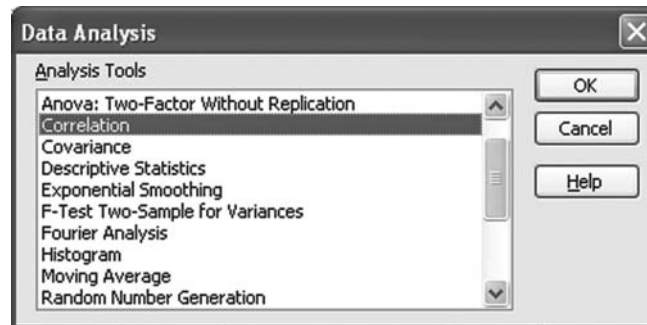


Figure 21.2 Window of Data Analysis

Step 4: In the following Dialog box, enter the input range that corresponds to the data columns (\$A\$1:\$B\$7). Click on Labels and then click on OK.

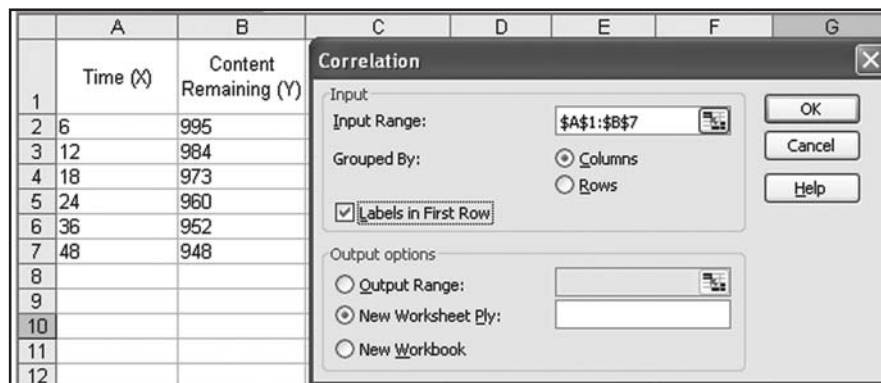


Figure 21.3 Window of Correlation

The results will appear in a new worksheet, as shown here:

	<i>Time (X)</i>	<i>Content Remaining (Y)</i>
Time (X)		1
Content Remaining (Y)	-0.949074491	1

So, correlation coefficient is **-0.949**.

Summary

Correlation

The relationship between two metric continuous variables is called correlation.

Types of correlation

- + 1.0 perfect positive correlation
- 0.0 no correlation
- 1.0 perfect negative correlation

Significance of correlation

- < 0.2 Slight correlation
- 0.20 - 0.40 Low correlation
- 0.40 - 0.70 Moderate correlation
- 0.70 - 0.90 High correlation
- > 0.9 Very high correlation

Computation of correlation coefficient

Using definitional formula

$$r = \frac{\sum (x - \bar{X})(y - \bar{Y})}{\sqrt{\sum (x - \bar{X})^2 \sum (y - \bar{Y})^2}}$$

Using computational formula

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

Multiple Choice Questions

- The relationship between two metric continuous variables is called as _____.
 - Association
 - Correlation
 - Regression
 - Relation
- Correlation coefficient for parametric data can be calculated by using _____.
 - Spearman rank
 - Pearson
 - a and b
 - None of above
- The correlation coefficient provides:
 - measure of the extent to which changes in one variable cause changes in another variable.
 - a measure of the strength of the linear association between two categorical variables.
 - a measure of the strength of the association between two categorical variables.
 - a measure of the strength of the linear association between two quantitative variables.

4. Correlation involves
- Independent variable
 - Response variable
 - a and b
 - None of above
5. A statistical test used to determine whether a correlation coefficient is statistically significant is called the _____.
- One-way analysis of variance
 - t-test for independent samples
 - Chi-square test for contingency tables
 - t-test for correlation coefficients
6. In this the variable are directly proportional to each other.
- perfect positive correlation
 - low correlation
 - moderate correlation
 - no correlation
7. The correlation coefficient value of 0.70-0.90 means _____.
- low correlation
 - high correlation
 - moderate correlation
 - very high correlation
8. If one variable increases, other variable decreases but the fall is directly proportional to each other. This is _____ correlation.
- perfect positive
 - moderate
 - perfect negative
 - no
9. For non parametric data, correlation is measured by _____.
- Spearman rank
 - Pearson
 - a and b
 - none of above
10. The value of correlation coefficient < 0.2 suggests _____.
- low correlation
 - moderate correlation
 - slight correlation
 - high correlation

Exercise

1. Following area under curves were observed after clinical trials of different formulations of same drug. Calculate correlation coefficient between dose and AUC

Dosage	100	300	600	900	1200
AUC	1.07	5.82	15.85	25.18	33.12

2. Following data were obtained after diffusion experiment of drug. Calculate correlation coefficient.

Time (h)	1	2	3	4	5	6	7
Amount in donor compartment (mg)	248.7	246.6	244.9	243.4	242.1	241.9	240.1

3. Data given are the values for age (X, in years) and systolic blood pressure (Y, in mm/Hg) for 15 women. Calculate correlation coefficient between age and systolic BP.

X	42	46	42	71	80	74	70	80	85	72	64	81	41	61	75
Y	130	115	148	100	156	162	151	156	162	158	155	160	125	150	165

4. The following are the height (cm) and the weight (kilogram) of 10 men. Calculate correlation coefficient between height and weight.

Height	162	168	174	176	180	180	182	184	186	186
Weight	65	65	84	63	75	76	82	65	80	81

5. Following data were obtained during laboratory experiment of muscular contraction of a rabbit intestine. The height of the curve was considered as the response to the drug. Calculate correlation coefficient between dose and response.

Dose of drug (mcg)	0.3	0.4	0.6	0.8	0.9	1.2
Response (mm)	54	59	60	65	70	75

6. In a study on the elimination of a drug in man, the following data were recorded. Calculate correlation coefficient between time and drug concentration.

Time (h)	0.5	1	2	3	4	5
Drug Concentration (g/ml)	0.39	0.34	0.27	0.2	0.16	0.1

7. An experiment was conducted to study the effect on sleeping time of increasing the dosage of a barbiturate. Three readings were made at each of three dose levels. Calculate correlation coefficient between dose and sleeping time.

Dosage ($\mu\text{M/kg}$)	3	3	3	10	10	10	15	15	15
Sleeping Time (Hrs)	4	6	5	9	8	7	13	11	9

8. Calculate correlation coefficient for following data of variables.

x	43	62	52	41	53	51	59	46
y	50	56	53	46	48	57	55	42

9. Find if any correlation exists between two variables x and y

x	10	09	11	12	13	09	11	12	10	11	12	14
y	12	14	11	14	11	10	10	14	16	15	12	13

10. Find the value of Karl Pearson's coefficient of correlation

x	12	9	8	10	11	13	7
y	14	8	6	9	11	12	3

Answers:

Multiple Choice Questions

1. b 2. b 3. d 4. b 5. d 6. a 7. b 8. c 9. a 10. c

Exercise

- | | |
|--------------|---------------|
| 1. $r=0.998$ | 2. $r=-0.984$ |
| 3. $r=0.564$ | 4. $r=0.514$ |
| 5. $r=0.981$ | 6. $r=-0.993$ |
| 7. $r=0.900$ | 8. $r=0.680$ |
| 9. $r=-0.01$ | 10. $r=0.948$ |



Chapter 22

LINEAR REGRESSION

Learning objectives

When we have finished this chapter, we should be able to

1. Differentiate between correlation and regression.
2. Understand meaning of regression.
3. Determine regression coefficient (b).

Linear Regression

Linear regression is a statistical method to evaluate how one or more independent variables influence outcomes for one continuous dependent variable through a linear relationship. For example, we can predict the weight of children (up to 10 yrs), based on their age. Here age is known as predictor and also known as independent variable. That which is predicted (weight) is referred to as dependent variable. By the use of a regression equation, we can predict scores on the dependent variable from those of independent variable. This is done by finding another constant called 'Regression Coefficient'.

Let us understand difference between correlation and regression as both describe the strength of the relationship between two or more continuous variables.

Correlation

1. Relationship between two variables is established but response for dependent variable (y) cannot be predicted based on independent variable (x).
2. Correlation involves only dependent or response variables.
3. Correlation simply describes the strength and direction of the relationship.
4. Correlation coefficient gives the idea of relationship between two or more continuous variable.

Linear Regression

1. Relationship between two variables is established but response for dependent variable (y) can be measured based on independent variable (x).
2. Regression involves at least one independent variable which is under researchers control.
3. Regression in addition provides a method for describing the nature of relationship between two or more continuous variable.
4. Regression coefficient is useful for predicting the value of y from the value of x, or vice versa.

Meaning of Regression

Regression is based on the relationship or association between two (or more) variables. In this analysis we have known variables which are used to predict the other variable. The known variable (or variables) is called the independent variable (or variables). The variable we are trying to predict is the dependent variable. In regression, we have only one dependent variable in our regression equation. However, we can use more than one independent variable.

Regression equation can be expressed as,

$$Y = a + bX \quad \dots 1$$

Where,

Y = Dependent variable

X = Independent variable

a = Y intercept

b = Slope of line

The above equation is linear in X and Y. Also graphically it represents a straight line. So straight line is called as regression line.

The regression analysis confined to study of only one independent variable is called the simple regression. If we are interested in studying relationship between a dependent variable with more than one independent variable then it is called as multiple regression.

Regression coefficient (b)

Regression coefficient is a measure of the change in one dependent character (Y) with one unit change in the independent character (X).

Calculation of Regression coefficient by least square method

Formula for calculating regression coefficient is

$$b = \frac{N \sum xy - (\sum x)(\sum y)}{N \sum x^2 - (\sum x)^2} \quad \dots 2$$

Where

$\sum xy$ = Summation of product of x and y.

$\sum y$ = Summation of y

$\sum x$ = Summation of x

$\sum x^2$ = Summation of x^2

$(\sum x)^2$ = Square of summation of x

N = Number of observations

b = Regression coefficient (Slope of line)

Steps for calculating regression coefficient

1. Develop a table as shown below

x	y	x^2	y^2	xy
x_1	y_1	x_1^2	y_1^2	x_1y_1
x_2	y_2	x_2^2	y_2^2	x_2y_2
x_3	y_3	x_3^2	y_3^2	x_3y_3
x_3	y_3	x_3^2	y_3^2	x_4y_4
x_n	y_n	x_n^2	y_n^2	x_ny_n
$\Sigma x =$	$\Sigma y =$	$\Sigma x^2 =$	$\Sigma y^2 =$	$\Sigma xy =$

- Write the corresponding values of x and y in column 1 and 2 respectively and calculate their sums (Σx and Σy)
- Write squared values of x and y in column 3 and calculate its sum.
- Write the squared values of y in column 4 and calculate its sum.
- Write the product of values of x and y in column 5 and calculate its sum .
- Using formula of regression coefficient find the value of 'b'.

$$b = \frac{N \Sigma xy - (\Sigma x)(\Sigma y)}{N \Sigma x^2 - (\Sigma x)^2}$$

Where

Σxy = Summation of product of x and y.

Σy = Summation of y

Σx = Summation of x

Σx^2 = Summation of x^2

$(\Sigma x)^2$ = Square of summation of x

N = Number of observations

b = Regression coefficient (Slope of line)

7. Now the y intercept (a) can be calculated by using formula

$$a = (\Sigma y - b \Sigma x) / N \quad \dots 3$$

8. Now, put the values of a and b in regression equation $y = a + bx$ and find the value of y corresponding value of x.

Example 22.1

Samples of drug products are stored in their original containers under normal conditions and sampled periodically to analyse the content of the medication. Determine slope and intercept by least square method.

Data:

Time (months)	6	12	18	24	36	48
Content (mg)	995	984	973	960	952	948

Solution:

1. Develop a table as shown below

x	y	x ²	y ²	xy
6	995	36	990025	5970
12	984	144	968256	11808
18	973	324	946729	17514
24	960	576	921600	23040
36	952	1296	906304	34272
48	948	2304	898704	45504
144	5812	4680	5631618	138108
Σx	Σy	Σx^2	Σy^2	Σxy

2. Using formula of regression coefficient find the value of 'b'.

$$b = \frac{N \Sigma xy - (\Sigma x)(\Sigma y)}{N \Sigma x^2 - (\Sigma x)^2} = \frac{(6 \times 138108) - (144 \times 5812)}{(6 \times 4680) - (144)^2}$$

$$b = \frac{828648 - 836928}{28080 - 20736} = \frac{-8280}{7344} = -1.127$$

3. Now the y intercept can be calculated by using formula

$$a = \frac{(\Sigma y - b \Sigma x)}{N} = \frac{(5812 - (-1.127 \times 144))}{6} = \frac{5812 + 162.28}{6} = 995.72$$

4. Now, $y = a + bx$. Once the values of a and b are known, then the value of y corresponding to x can be calculated. For example, if value of y for corresponding value of $x = 18$ is to be determined, then

$$\begin{aligned} y &= a + bx \\ y &= 995.72 - 1.127(18) \\ &= 995.72 - 20.286 \\ &= 975.4 \end{aligned}$$

This can be used to cross check whether values of a and b are correct.

Example 22.2

The values for the birthweight (X, in kgs) and the increase in weight between 70 and 100 days of life, expressed as a percentage of the birth weight (Y) for 9 infants. Perform linear regression

analysis and determine slope and intercept. Report graphical display of data.

Data:

X	112	111	107	119	80	81	84	106	94
Y	63	66	72	52	118	120	114	72	91

Solution:

1. Develop a table as shown below

x	y	x ²	y ²	xy
112	63	12544	3969	7056
111	66	12321	4356	7326
107	72	11449	5184	7704
119	52	14161	2704	6188
80	118	6400	13924	9440
81	120	6561	14400	9720
84	114	7056	12996	9576
106	72	11236	5184	7632
94	91	8836	8281	8554
894	768	90564	70998	73196
Σx	Σy	Σx^2	Σy^2	Σxy

2. Using formula of regression coefficient find the value of 'b'.

$$b = \frac{N \Sigma xy - (\Sigma x)(\Sigma y)}{N \Sigma x^2 - (\Sigma x)^2} = \frac{(9 \times 73196) - (894 \times 768)}{(9 \times 90564) - (894)^2}$$

$$b = \frac{658764 - 686592}{815076 - 799236} = \frac{-27828}{15840} = -1.76$$

3. Now the y intercept can be calculated by using formula

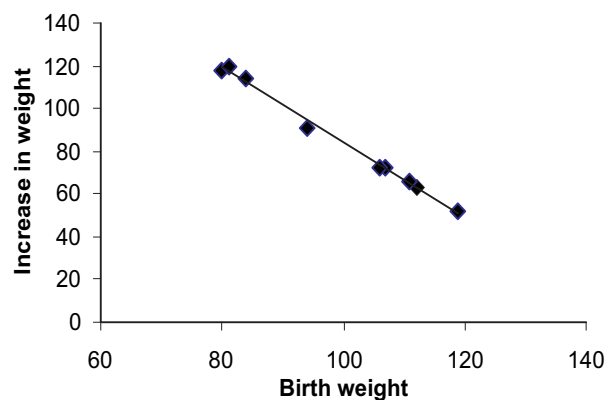
$$a = \frac{(\Sigma y - b \Sigma x)}{N} = \frac{(768 - (-1.76 \times 894))}{9} = \frac{768 + 1573.44}{9} = 259.84$$

4. Now, $y = a + bx$. Once the values of a and b are known, then the value of y corresponding to x can be calculated.

$$y = a + bx$$

$$y = 259.84 - 1.76(x)$$

5. Graphical display of data

**Estimation of Linear Regression using Excel****Example 22.1**

Samples of drug products are stored in their original containers under normal conditions and sampled periodically to analyse the content of the medication. Determine slope and intercept by least square method. Determine content after 20 months.

Data:

Time (months)	6	12	18	24	36	48
Content (mg)	995	984	973	960	952	948

Excel Solution**Step 1:**

Open new file in MS-Excel as Book 1. Enter the data into an Excel datasheet (Sheet 1). Worksheet will appear as follows:

	A	B	
1	Time (X)	Content Remainin g (Y)	
2	6	995	
3	12	984	
4	18	973	
5	24	960	
6	36	952	
7	48	948	
8			

Figure 22.1 Data entry

Step 2:

In MS-Excel, select Tools menu from Menu bar. Then, it displays pull down menus. From pull down menus, select Data Analysis option. Instantly, Data Analysis dialog box will appear.

Step 3:

Select Regression option from Analysis Tools and then click on Ok.

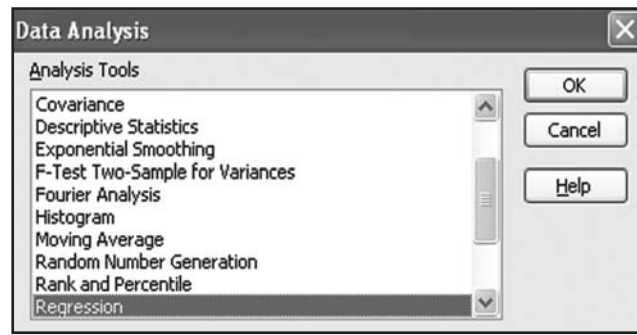


Figure 22.2 Window of Data analysis

Step 4: In the following Dialog box, enter the input range that corresponds to the data columns.

For Input Y Range box, type: \$B\$1:\$B\$7

For Input X Range box, type: \$A\$1:\$A\$7

Also you can select ranges

Click on Labels and

Then click on OK.

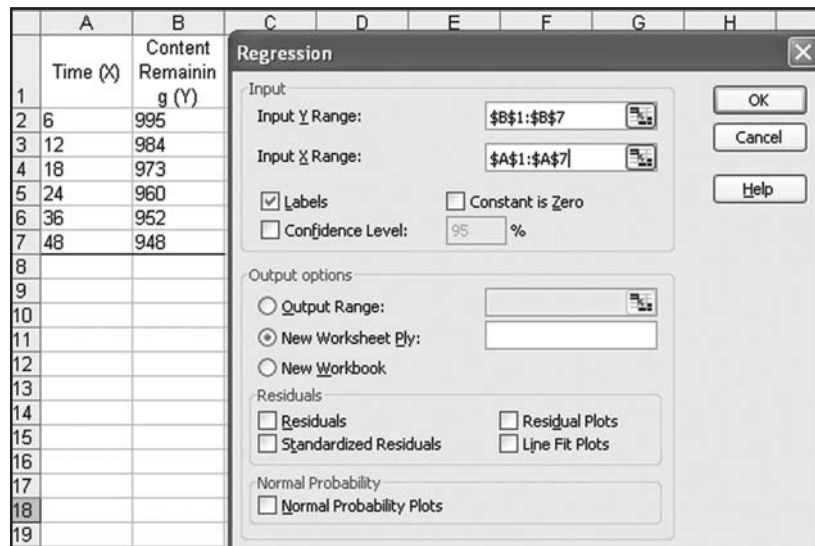


Figure 22.3 Window of Regression

The results will appear in a new worksheet, as shown here:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9490745
R Square	0.9007424
Adjusted R Square	0.875928
Standard Error	6.5469646
Observations	6

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	1555.882	1555.882	36.29918	0.003824076
Residual	4	171.451	42.86275		
Total	5	1727.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	995.72549	5.226327	190.5211	4.55E-09	981.21488	1010.236	981.2149	1010.236
Time (X)	-1.127451	0.187133	-6.02488	0.003824	-1.647014168	-0.607888	-1.647014	-0.607888

In above results, $b = -1.127$; $a = 995.72$

Regression equation

$$y = a + bx$$

$$y = 995.72 - 1.127x$$

Summary

Regression

Regression is based on the relationship or association between two (or more) variables.

Regression equation, $Y = a + bX$

Regression coefficient (b)

Regression coefficient is a measure of the change in one dependent character (Y) with one unit change in the independent character (X)

$$b = \frac{N \sum xy - (\sum x)(\sum y)}{N \sum x^2 - (\sum x)^2}$$

$$a = (\sum y - b \sum x) / N$$

Multiple choice questions

1. Given that we have collected pairs of observations on two variables X and Y, we would consider fitting a straight line with X as an explanatory variable if:

- the change in Y is an additive constant.
- the change in Y is a constant for each unit change in X
- the change in Y is a fixed percent of Y
- the change in Y is exponential

2. The least squares regression line is the line which:
 - a. is determined by use of a function of the distance between the observed Y 's and the predicted Y 's.
 - b. has the smallest sum of the squared residuals of any line through the data values.
 - c. for which the sum of the residuals about the line is zero.
 - d. has all of the above properties
3. Regression equation can be expressed as,
 - a. $Y = a + bX$
 - b. $X = a + bY$
 - c. $Y = abX$
 - d. $X = abY$
4. Regression is based on
 - a. the relationship or association between two (or more) variables.
 - b. the comparison between two (or more) variables
 - c. a & b
 - d. none of above
5. _____ is the set of procedures used to explain or predict the values of a dependent variable based on the values of one or more independent variables.
 - a. Regression analysis
 - b. Regression coefficient
 - c. Regression equation
 - d. Regression line
6. _____ involves at least one independent variable which is under researchers control.
 - a. Correlation
 - b. Association
 - c. Regression
 - d. None of above
7. _____ is useful for predicting the value of y from the value of x .
 - a. Correlation coefficient
 - b. Regression coefficient
 - c. Pearson coefficient
 - d. Association coefficient
8. The regression analysis confined to study of only one independent variable is called _____.
 - a. simple regression
 - b. multiple regression
 - c. simple correlation
 - d. multiple correlation
9. The known variable which is used to predict other variable is _____.
 - a. independent variable
 - b. dependent variable
 - c. regression variable
 - d. response variable
10. _____ involves only dependent or response variable.
 - a. Simple regression
 - b. Multiple regression
 - c. Association
 - d. Correlation

Exercise

1. In an experiment investigating the breakdown of aspirin in a pharmaceutical product stored at 25°C is given below. Perform linear regression analysis.

Time	18	35	51	68	85	101	118	136	166
Aspirin remaining	603.6	601.1	597.9	594.3	591.4	587.3	581.8	580	578.4

Time	197	229	260	292	326	355
Aspirin remaining	570.5	561.4	557.4	549.2	546	541

2. Following area under curves were observed after clinical trials of different formulations of same drug. Calculate slope and intercept.

Dosage	100	300	600	900	1200
AUC	1.07	5.82	15.85	25.18	33.12

3. Following data were obtained after diffusion experiment of drug. Determine slope and present data graphically.

Time (h)		1	2	3	4	5	6	7
Amount in donor compartment (mg)		248.7	246.6	244.9	243.4	242.1	241.9	240.1

4. Following data were obtained during laboratory experiment of muscular contraction of a rabbit intestine. The height of the curve was considered as the response to the drug. Perform linear regression analysis.

Dose of drug (mcg)	0.3	0.4	0.6	0.8	0.9	1.2
Response (mm)	54	59	60	65	70	75

6. Data give the values for age (x, in years) and systolic blood pressure (y, in mm/Hg) for 15 women. Perform linear regression analysis.

X	42	46	42	71	80	74	70	80	85	72	64	81	41	61	75
Y	130	115	148	100	156	162	151	156	162	158	155	160	125	150	165

7. An experiment was conducted to study the effect on sleeping time of increasing the dosage of a certain barbiturate. Three readings were made at each of three dose levels. Perform linear regression analysis.

Dosage ($\mu\text{M/kg}$)	3	3	3	10	10	10	15	15	15
Sleeping Time (Hrs)	4	6	5	9	8	7	13	11	9

8. In a study on the elimination of a drug in man, the following data were recorded. Perform linear regression analysis.

Time (h)		0.5	1	2	3	4	5
Drug Concentration (g/ml)		0.39	0.34	0.27	0.2	0.16	0.1

9. Perform regression analysis of following data

Age (years)	66	38	56	72	42	45	55	47	36	63
Blood pressure (mm of Hg)	145	124	147	160	125	124	150	128	118	149

10. Obtain the regression equation of sale of medicine and advertising expenses.

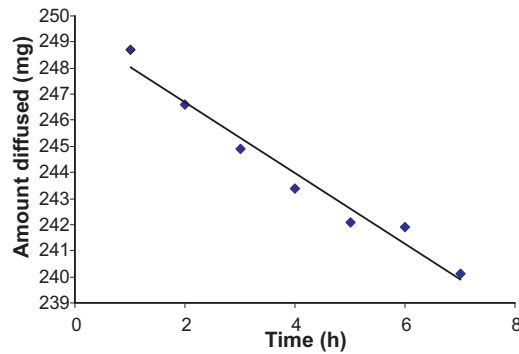
Sales of Medicines	190	240	250	300	310	335	300
Advertising expenses	5	10	15	20	20	30	30

Answers:**Multiple Choice Questions**

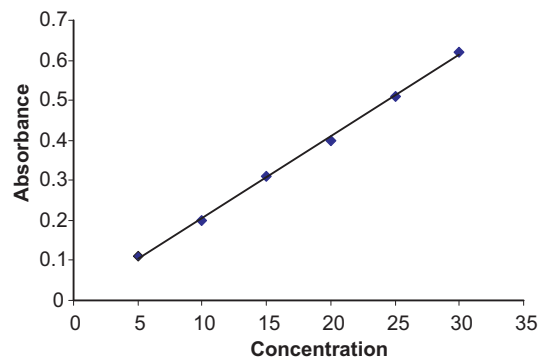
1. b 2. b 3. a 4. c 5. a 6. c 7. b 8. a 9. a 10. d

Exercise

1. Intercept 606.98, Slope -0.19.
2. Intercept -2.3, Slope 0.03.
3. Intercept 249.38, Slope -1.36.



4. Intercept 47.96, Slope 22.68.
5. Intercept 0.0013, Slope 0.02.



6. Intercept 99.96, Slope 0.7.
7. Intercept 3.37, Slope 0.495.
8. Intercept 0.406, Slope -0.063.
9. Intercept 79.03, Slope 1.114
10. $y = a + bx$ $y = -28.62 + 0.171x$



Important Points & Formulaes At A Glance

Presentation of data

Presentation of data	Nominal	Ordinal	Metric Continuous	Metric Discrete
Graphical	Pie chart, Bar chart	Bar chart	Histogram, frequency polygon, Box & Whisker, Stem & Leaf plot Ogive, Scatter plot	Bar chart, Line plot, Point plot

Measures of Location & Dispersion

Data	Ungrouped data	Grouped data	Metric discrete data
Mode	Most frequently occurring score	$\text{Mode} = l_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times i$	
Median	Odd number = Size or value of $\left(\frac{n+1}{2} \right)$ th observation Even number = Size or value of $\frac{n}{2}$ th + $\left(\frac{n}{2} + 1 \right)$ th $\frac{1}{2}$ observation	$\text{Median} = l_1 + \frac{(n/2) - \text{c.f.}}{f} \times i$	Median = value of (n/2)th observation
Mean	$\text{Mean } (\bar{X}) = \frac{\sum X}{N}$	$\text{Mean } (\bar{X}) = A + \frac{\sum f_i d_i}{N} \times i$	$\text{Mean } (\bar{X}) = \frac{\sum f_i X_i}{N}$
Percentile	$\text{Pth Percentile} = \frac{P}{100} (N+1)\text{th value}$		
Range	Range = Lowest value to Highest value		
IQR	Interquartile range = Q1 to Q3		
SD	$s = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}}$	$s = \sqrt{\frac{\sum fd^2 - \frac{(\sum fd)^2}{N}}{N-1}} \times i$	$s = \sqrt{\frac{\sum f_i X_i^2 - \frac{(\sum f_i X_i)^2}{N}}{N-1}}$
Coefficient of variation (CV) = $\frac{\text{SD}}{\text{mean}}$ Relative Standard Deviation = CV x 100			

Probability

$$\text{Probability } [P^E] = \frac{\text{Number of outcome that favour the event (m)}}{\text{Total number of outcomes (N)}}$$

$$\text{Probability of composite outcomes: } P(E_i \text{ or } E_j) = P(E_i) + P(E_j)$$

$$\text{Probability of complementary event: } p(\bar{E}) = 1 - p(E)$$

$$\text{Probability of intersect: } p(A \text{ and } B) = p(A \cap B) = p(A) \times p(B)$$

$$\text{Probability of conjoint: } p(A \text{ or } B) = p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

$$\text{Conditional probability: } p(A) \text{ given } B = p(A|B) = p(A \cap B) / p(B)$$

Probability of binomial distribution:

$$P(X) = \binom{n}{X} p^X q^{n-X} \quad \binom{n}{X} = \frac{n!}{X!(n-X)!} \quad P(X) = \frac{n!}{X!(n-X)!} p^X q^{n-X}$$

Poisson distribution:

$$P(X) = \frac{e^{-\mu} \mu^X}{X!} = \frac{\mu^X}{e^\mu X!}$$

Estimation of Confidence Interval**Standard error of Mean**

$$SEM = \frac{SD}{\sqrt{N}}$$

At 95% confidence, Population mean = Sample mean \pm 1.96 x SE

Confidence interval

$$p\% = \bar{X} \pm \frac{Z_{1-\alpha/2} \sigma}{\sqrt{N}}$$

Key stages in hypothesis testing

- | | |
|-------------------------------------|-------------------------------------|
| 1. State the null hypothesis | 2. State the alternative hypothesis |
| 3. Select the level of significance | 4. Select number of tails |
| 5. Test the statistics | 6. Compare table and observed value |
| 7. Decision | |

Decision rule

1. Determine p value
2. Compare it with the critical value, usually 0.05.
3. If the obtained p value is less than critical value, reject null hypothesis; otherwise accept it.

Types of error

Type I error is the probability of rejecting a true null hypothesis

Type II error is the probability of accepting a false H_0 .

Choice of statistical test

	z-test	t-test
Single sample	$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{N}}$	$t_{(N-1)df} = \frac{\bar{X} - \mu_0}{S / \sqrt{N}}$
Two independent samples	$z = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}$ $SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$	$t = \frac{(\bar{X}_1 - \bar{X}_2)}{SE(\bar{X}_1 - \bar{X}_2)} \quad SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_p^2}{N_1} + \frac{S_p^2}{N_2}}$ $S_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$
Two paired samples		$t = \frac{\bar{x} - O}{SE} = \frac{\bar{x} - O}{SD / \sqrt{N}} \quad SD = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}}$

	ANOVA
Definitional formula	$MSW = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 \dots (n_k - 1)s_k^2}{N}$ $\bar{X}_G = \frac{(n_1 \bar{X}_1) + (n_2 \bar{X}_2) + (n_3 \bar{X}_3) \dots (n_k \bar{X}_k)}{N}$ $MSB = \frac{n_1(\bar{X}_1 - \bar{X}_G)^2 + n_2(\bar{X}_2 - \bar{X}_G)^2 + n_3(\bar{X}_3 - \bar{X}_G)^2 \dots n_k(\bar{X}_k - \bar{X}_G)^2}{K - 1}$ $F = \frac{MSB}{MSW}$
Computational formula	$SST = \sum X^2 - \frac{(\sum X)^2}{N}$ $SSB = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} - \frac{(\sum X_T)^2}{N}$

	Correlation Coefficient	Regression Coefficient
Definitional formula	$r = \frac{\sum(x - \bar{X})(y - \bar{Y})}{\sqrt{\sum(x - \bar{X})^2 \sum(y - \bar{Y})^2}}$	Regression equation, $Y = a + bX$ $b = \frac{N \sum xy - (\sum x)(\sum y)}{N \sum x^2 - (\sum x)^2}$ $a = (\sum y - b \sum x) / N$
Computational formula	$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$	

Appendix I

Normal-Curve Areas

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Also, for $z = 4.0$, 5.0 and 6.0 , the areas are 0.49997 , 0.4999997 , and 0.499999999 .

Appendix 2

Critical values of the t distribution

df	Two-tailed test			One-tailed test		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
1	6.314	12.706	63.657	3.078	6.314	31.821
2	2.920	4.303	9.925	1.886	2.920	6.965
3	2.353	3.182	5.841	1.638	2.353	4.541
4	2.132	2.776	4.604	1.533	2.132	3.747
5	2.015	2.571	4.032	1.476	2.015	3.365
6	1.943	2.447	3.707	1.440	1.943	3.143
7	1.895	2.365	3.499	1.415	1.895	2.998
8	1.860	2.306	3.355	1.397	1.860	2.896
9	1.833	2.262	3.250	1.383	1.833	2.821
10	1.812	2.228	3.169	1.372	1.812	2.764
11	1.796	2.201	3.106	1.363	1.796	2.718
12	1.782	2.179	3.055	1.356	1.782	2.681
13	1.771	2.160	3.012	1.350	1.771	2.650
14	1.761	2.145	2.977	1.345	1.761	2.624
15	1.753	2.131	2.947	1.341	1.753	2.602
16	1.746	2.120	2.921	1.337	1.746	2.583
17	1.740	2.110	2.898	1.333	1.740	2.567
18	1.734	2.101	2.878	1.330	1.734	2.552
19	1.729	2.093	2.861	1.328	1.729	2.539
20	1.725	2.086	2.845	1.325	1.725	2.528
21	1.721	2.080	2.831	1.323	1.721	2.518
22	1.717	2.074	2.819	1.321	1.717	2.508
23	1.714	2.069	2.807	1.319	1.714	2.500
24	1.711	2.064	2.797	1.318	1.711	2.492
25	1.708	2.060	2.787	1.316	1.708	2.485
26	1.706	2.056	2.779	1.315	1.706	2.479
27	1.703	2.052	2.771	1.314	1.703	2.473
28	1.701	2.048	2.763	1.313	1.701	2.467
29	1.699	2.045	2.756	1.311	1.699	2.462
30	1.697	2.042	2.750	1.310	1.697	2.457
40	1.684	2.021	2.704	1.303	1.684	2.423
60	1.671	2.000	2.660	1.296	1.671	2.390
120	1.658	1.980	2.617	1.289	1.658	2.358
∞	1.645	1.960	2.576	1.282	1.645	2.326

α denotes the level of significance and df the number of degrees of freedom.

Appendix 3

Values of F 0.05		Degrees of freedom for numerator																			
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞	
Degrees of freedom for denominator																					
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254		
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5		
3	10.10	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53		
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63		
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37		
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67		
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23		
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93		
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71		
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54		
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40		
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30		
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21		
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13		
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07		
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01		
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.29	2.15	2.10	2.06	2.01	1.96		
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92		
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88		
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84		
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81		
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78		
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76		
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73		
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71		
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.76	1.74	1.68	1.62		
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51		
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39		
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25		
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00		

Appendix 3
Values of F 0.01

Degrees of freedom for numerator

Critical Values of F Distribution																				
		Degrees of freedom for denominator																		
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	5000	5403	5625	5764	5859	5928	5982	6023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366	
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1	
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5	
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02	
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.05	6.97	6.88	
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65	
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86	
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31	
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91	
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60	
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36	
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17	
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00	
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87	
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75	
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65	
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57	
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49	
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42	
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36	
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31	
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26	
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21	
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17	
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01	
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80	
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60	
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38	
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00	

BIBLIOGRAPHY

- ❑ Belle GV, Fisher LD, Heagerty PJ, Lumley T. Biostatistics A methodology for the health sciences. John Wiley & Sons, Inc. America, 2004.
- ❑ Bowers D. Medical Statistics from Scratch An Introduction for Health Professionals. John Wiley & Sons, Ltd., England, 2008.
- ❑ David J. Pharmaceutical Statistics. Pharmaceutical Press, London, 2002.
- ❑ De Muth JE. Basic statistics and Pharmaceutical applications. Marcel Dekker Inc, New York, 1999.
- ❑ Freund JE. Modern elementary statistics. Prentice Hall International Inc, London, 1984.
- ❑ Gupta V. Statistical analysis with Excel. VJ Books Inc, Canada, 2002.
- ❑ Kothari CR. Research methodology methods and techniques. New Age International Pvt Ltd., New Delhi, 2004.
- ❑ Le CT. Introductory statistics. John Wiley & Sons, Inc., America, 2003.
- ❑ Mahajan BK. Methods in Biostatistics. Jaypee, New Delhi, 2010.
- ❑ Mithal P, Goel R. Computer fundamentals. Paragon International Publishers, New Delhi, 2007.
- ❑ Nagpal DP. Computer fundamentals. S. Chand & Company Ltd, New Delhi, 2008.
- ❑ Oka MM. Computer Fundamentals. 6th Edition. Everest Publishing House, Pune, 2001.
- ❑ O'Leary TJ, O'Leary LI. Computer Essentials. Tata McGraw-Hill, New Delhi, 2002.
- ❑ Po WAL. Statistics for pharmacists. Blackwell Publishing Ltd, India, 2006.
- ❑ Rajaraman V. Fundamentals of Computers. 4th Edition. Prentice-Hall of India Pvt Ltd., New Delhi, 2004.
- ❑ Ram B. Computer fundamentals architecture and organization. New Age International Publishers, Delhi, 2007.
- ❑ Rao BT. Methods of Biostatistics. Paras Medical Publishers, Hyderabad, 2010.
- ❑ Row P. Essential statistics for the pharmaceutical sciences. John Wiley & Sons, Ltd., England, 2007.
- ❑ Thakur PS, Manchanda R, Nand P. Computer in Pharmacy. 2nd Edition. Birla Publications Pvt Ltd., New Delhi, 2002.
- ❑ Shah YI, Paradkar AR, Dhayagude MG. Introduction to biostatistics and computer applications. Nirali Prakashan, Pune, 2007.
- ❑ Sinha PK, Sinha PP. Computer Fundamentals. 4th Edition. BPB Publication, New Delhi, 2007.
- ❑ Tipnis HP, Bajaj A. Clinical Pharmacy. Career Publications, Nasik, 2009.