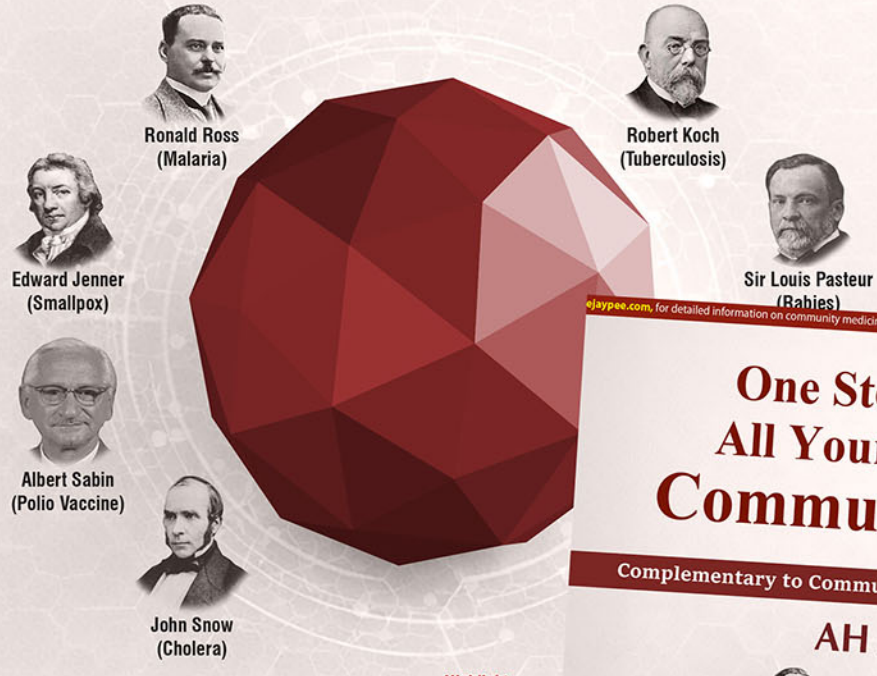
tion on community medicine books, visit our website www.ejaypee.com, for detailed information on community medicine books, visit our website www.ejaypee.com,

COMMUNITY MEDICINE *with* Recent Advances

Complementary Book Included

AH Suryakantha



Highlights

• Prudent diet • Aarogya Mitra • Umbrella ICDS—Scheme • Steps taken by the GoI fo
Ayushman Arogya Mandir • Polio Eradication Strategy, 2022-2026 • Mission Sampark • Populat

As per the Revised Competency Based Medica

9th Edition



Dr. Wise
Innovative & Geniu
AI Chatbot
access code inside

ejaypee.com, for detailed information on community medicine books, visit our website www.ejaypee.com, for detailed information on community medicine books

One Stop Resource for All Your Exam Needs in Community Medicine

Complementary to Community Medicine with Recent Advances, 9/e

AH Suryakantha



Long and short answer questions (last 5 years) • Case scenario-based questions
Reasoning type questions • 1,000+ MCQs (Including case based & reasoning types)

diginerve
A Jaypee Initiative

JAYPEE

Contents

Color Plates

SECTION I: Basic Concepts of Community Medicine

Chapter 1: Introduction to Community Medicine 3
- Hygiene *3*
- Preventive Medicine *3*
- Social Medicine *3*
- Public Health *4*
- Community Diagnosis *6*

Chapter 2: Indigenous System of Medicine in India 9
- Ayurveda *9*
- Yoga *9*
- Unani *10*
- Siddha *10*
- Homeopathy *10*
- Naturopathy *10*

Chapter 3: Concept of Health 13
- Changing Concepts of Health *13*
- Definition *13*
- Physical Dimension *13*
- Mental Dimension *13*
- Social Dimension *13*
- Spiritual Dimension *14*
- Emotional Dimension *14*
- Vocational Dimension *14*
- Positive Health *14*
- Well-being *14*
- Human Development Index *14*
- Gross Domestic Product *15*
- Spectrum of Health *15*
- Iceberg of Health *15*
- Determinants of Health *15*
- Barriers to Health *17*
- Indicators of Health *17*
- Holistic Health *22*
- Digi Health *23*

Chapter 4: Concept of Disease 24
- Natural History of Disease *25*
- Iceberg Phenomenon of Disease *27*

Chapter 5: Concept of Prevention 28
- Primary Prevention *28*
- Primordial Prevention *28*
- Jakarta Declaration on Health Promotion in the 21st Century *29*
- Secondary Prevention *30*
- Tertiary Prevention *31*

SECTION II: Environment and Health

Chapter 6: Environment and Water 35
- Water *35*
- Water Cycle *36*
- Well *36*
- Health Hazards of Water Contamination *38*
- Purification of Water *38*
- Kangen Water *50*
- Hardness of Water *51*
- Conservation of Water Resources *52*
- Laboratory Examination of Water *52*
- Water Quality Standards *56*

Chapter 7: Air and Ventilation 58
- Changes in the Air Due to Human Occupancy *58*
- Effects of Vitiated Air *58*
- Indicators of Thermal Comfort *58*
- Air Pollution *59*
- Ventilation *62*

Chapter 8: Noise 65
- Noise Level Values *65*
- Instruments Used in the Study of Noise *65*
- Hazards of Noise Pollution *65*
- Prevention and Control of Noise Pollution *66*

Chapter 9: Light 67
- Measurement of Light *67*
- Natural Lighting *67*
- Artificial Lighting *67*
- Health Hazards of Lighting *68*

Chapter 10: Radiation 69
- Nonionizing Radiations *69*
- Ionizing Radiations *70*
- Prevention and Control of Radiation Hazards *72*

Chapter 11: Housing 74
- Requirements of a House *74*
- Housing Standards *74*
- Housing and Health *75*

- Indicators of Housing *76*
- Housing Problem *76*
- Slaughter House *76*

Chapter 12: Meteorology 79
- Atmospheric Pressure *79*
- Air Temperature *80*
- Humidity *82*
- Air Movement *83*

Chapter 13: Disposal of Wastes 85
- Refuse *85*
- Methods of Refuse Disposal *85*
- Recycling of Refuse *87*
- Disposal of Excreta *87*
- Swachh Bharat Abhiyan *93*
- E-Toilets *93*
- Bio-Toilet *94*
- Disposal of Sullage *95*
- Disposal of Sewage (Sewage Treatment) *95*
- Great Sanitary Awakening *98*

Chapter 14: Management of Hospital Waste 100
- Characteristics of Healthcare Waste/Biomedical Waste *100*
- Objectives of the Waste Management System *101*
- Biomedical Waste Management Rules, 2016 *102*
- Schedule I: Part – 1 *103*
- Strategies Adopted for Hospital Waste Management *104*
- Available Treatment and Disposal Technologies *106*
- Decontamination Procedure *112*
- Plastic Waste Management *112*

Chapter 15: Electronic Waste Management 114
- E-waste Burden Globally *115*
- Burden in India *115*
- Sources of E-waste *115*
- Impacts/Hazards of E-waste *115*
- Basel Convention *115*
- Waste Minimization Techniques *116*
- Management Options *116*
- E-waste Disposal *117*

Chapter 16: Medical Entomology 118
- Classification of Arthropods *118*
- Arthropod-borne Diseases *119*
- Common Terms Used in Entomology *119*
- Principles of Arthropod Control *119*
- Arthropods of Public Health Importance *120*
- Disinsection *152*
- Insecticide Toxicity *154*
- Rodents *154*

SECTION III: Nutrition and Health

Chapter 17: Nutrition and Health 159
- Classification of Foods *159*
- Nutrients *159*
- Antioxidants *171*
- Food Groups *171*
- Energy Requirements *177*
- Balanced Diet *178*
- Food Hygiene *182*
- Food-related Diseases (Public Health Nutrition Problems) *190*
- Anemia Mukt Bharat (Intensified National Iron Plus Initiative) *198*
- Iodine-Deficiency Disorders *201*
- Assessment of Iodine Deficiency Disorders *203*
- Global Hunger Index *205*
- National Nutrition Policy *217*
- National Programs on Nutrition *218*
- Community Nutrition Programmes *218*
- Annexure *219*

SECTION IV: Occupational Health

Chapter 18: Occupational Health 223
- Burden of Occupational Diseases *223*
- Ergonomics *223*
- Occupational Hazards *225*
- Pneumoconioses *226*
- Lead Poisoning (Plumbism) *230*
- Metal Fume Fever (Brass Founder's Ague) *232*
- Occupational Cancers *232*
- Occupational Dermatoses *234*
- Occupational Health in Agricultural Industry *234*
- Hazards Due to Industrialization and Urbanization *235*
- Accidents in Industries *236*
- Sickness Absenteeism *236*
- Women in Industry *237*
- Health Status of Industry *238*
- Prevention and Control of Occupational Hazards *238*
- Legislation *241*
- Social Security *244*
- National Health Insurance Schemes (National Health Portal of India) *246*

SECTION V: Epidemiology

Chapter 19: Principles and Scope of Epidemiology 253
- Epidemiology *253*
- Mortality Rates Versus Mortality Ratios *259*
- Medical Certificate of Cause of Death *260*
- Verbal Autopsy *262*
- Epidemiological Studies (Methods) *262*
- Association and Causation *275*

Chapter 20: Epidemiology of Infectious Diseases 278
- Commonly Used Terms *278*
- Eradication of a Disease *280*
- Surveillance *281*
- Sentinel Surveillance *282*
- Dynamics of Disease Transmission *284*
- Incubation Period *287*
- Investigation of an Outbreak of a Disease *289*
- Prevention and Control of an Epidemic Disease *291*

- Investigations of Adverse Events Following Immunization *297*
- Immunization Program *297*
- Cold Chain *307*
- Reverse Cold Chain *310*
- Screening for Disease *313*
- Screening of Blood in Blood Bank *317*
- Disinfection *318*
- Healthcare-associated Infections (Nosocomial Infections) *324*
- Health Hazards Among Healthcare Workers *328*
- Emporiatrics *330*

Chapter 21: Epidemiology of Communicable Diseases..... 334

Air-borne Diseases 334
- Story of Smallpox (Variola Major) (Obituary) *334*
- Chickenpox (Varicella; Waterpox) *335*
- Acute Respiratory Infections *337*
- Pneumococcal Vaccines *339*
- Measles (Morbilli; Rubeola) *341*
- Global Elimination of Measles and Rubella Strategic Framework, 2021–2030 *348*
- Mumps *349*
- Influenza *350*
- Diphtheria *358*
- Whooping Cough (Pertussis) *363*
- Meningococcal Meningitis (Cerebrospinal Fever or Cerebrospinal Meningitis) *365*
- Tuberculosis *367*
- Tuberculosis and HIV *379*
- Severe Acute Respiratory Syndrome *379*

Water-borne Diseases *382*
- Typhoid Fever *382*
- Acute Diarrheal Diseases *386*
- Cholera *392*
- Viral Hepatitis *397*
- Poliomyelitis *406*
- Dracunculiasis *412*
- Amebiasis *413*
- Giardiasis *414*

Soil-borne Diseases 416
- Ascariasis *416*
- Ancylostomiasis *418*
- Tetanus *422*
- Neonatal Tetanus *426*

Vector-borne Diseases 428
- Malaria *428*
- Lymphatic Filariasis *437*
- Human African Trypanosomiasis *444*
- Zoonoses *445*
- Plague *448*
- Leptospirosis *452*
- Leishmaniases *454*
- Arboviral Diseases *456*
- Yellow Fever *456*
- Dengue Fever *458*
- Japanese Encephalitis *462*
- Chikungunya Fever (Epidemic Polyarthritis) *466*
- Kyasanur Forest Disease *467*
- West Nile Fever *470*
- Rickettsial Diseases *471*
- Viral Hemorrhagic Fever Diseases *473*
- Slow Viral Diseases *476*

Sexually Transmitted Infections 477
- Extent of the Problem *478*
- Agent Factors *478*
- Host Factors *478*
- Social Factors *478*
- Demographic Factors *479*
- International Travel *479*
- Changes in the Lifestyle *479*
- High-risk Groups *479*
- Prevention and Control *479*
- Syndromic Case Management of RTI/STI (Syndromic Approach) *480*
- Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome *483*
- Antiretroviral Treatment in Adults and Adolescents (National Guidelines) *492*
- Occupational Post Exposure Prophylaxis for HIV *499*
- Parent-to-child Transmission of HIV *499*
- Prophylaxis for Opportunistic Infections *503*

Surface Infections 504
- Rabies *504*
- National Action Plan for Dog Mediated Rabies Elimination from India by 2030 *512*
- Leprosy *514*
- Yaws *530*

Chapter 22: Epidemiology of Noncommunicable Diseases..... 533
- Noncommunicable Diseases *533*
- Future Plan to Prevent and Control Noncommunicable Diseases *535*
- Global Monitoring Framework for Noncommunicable Diseases *536*
- Cardiovascular Diseases *539*
- Obesity *549*
- Obesity Among School Children *551*
- Cancer *553*
- Breast Cancer *565*
- Screening for Breast Cancer *566*
- Global Breast Cancer Initiative *568*
- Diabetes Mellitus *568*
- Mass Screening for Diabetes Mellitus *572*
- Accidents *574*
- Burns *576*
- Blindness *578*
- Stress *584*
- Suicide *585*
- Fluorosis *588*

SECTION VI: Health-related Disciplines

Chapter 23: Maternal and Child Health Services (Preventive Obstetrics and Pediatrics) 595
- Mother and Child—One Unit *595*
- Services to the Mothers *596*
- Services to Children *606*
- Wash in Schools *635*
- Swasthya *636*
- Indicators of Maternal and Child Health Care *636*
- Home-based Newborn Care *644*
- Rights of the Child *645*
- National Programmes for the Welfare of Mothers and Children *645*
- Womens' Empowerment (Surakshit Bharat for Women) *646*
- Gender Equality and Women Empowerment *647*
- Integrated Mother (ICDS-Scheme) Development Services Scheme *648*
- Street Children *651*
- Child Welfare Policies, Legislations and Programs in India *653*

Chapter 24: Demography 656
- Sources of Demographic Data *656*
- Demographic Cycle (Demographic Stages) *656*
- Size of the Population (Magnitude of the Population) *657*
- Registration of Vital Events *658*
- Epidemiological Transition *658*
- Sample Registration System *659*
- Population Trend in the World *660*
- Population Trend in India *661*
- Mid-Year Population *662*
- Composition of the Population *662*
- Population Pyramid (Age and Sex Composition) *662*
- Gender Equity and Gender Equality *666*
- Distribution of the Population *667*
- Population Dynamics *667*
- Population Explosion (Population Bomb) *669*
- Population Stabilization *670*

Chapter 25: Family Planning 672
- Scope of Family Planning Services *672*
- Contraceptive Methods (Fertility-regulating Methods; Techniques of Birth Control) *673*
- Medical Termination of Pregnancy Amendment Bill, *2020* *689*
- Male Contraception *695*
- Evaluation of Contraceptive Methods *696*
- Unmet Need for Family Planning *696*
- Social Marketing *697*
- National Family Health Survey-5 (2019–2021) *697*

Chapter 26: Biostatistics 699
- Variability *699*
- Applications of Biostatistics *699*
- Data—Definition, Types and Sources *699*
- Presentation of Statistical Data *701*
- Measures of Central Tendency *708*
- Measures of Location: Quantiles *710*
- Measures of Dispersion *711*
- Probability *713*
- Sampling *717*
- Significance *722*
- Correlation and Regression Correlation *732*
- Life Table *735*
- Meta-analysis *736*
- Research Methodology *737*
- Formulation of a Research Question *739*

Annexure I* *740
Annexure II* *741
Annexure III* *742
Annexure IV* *743

Chapter 27: Social Science 744
- Community *744*
- Culture *745*
- Acculturation *745*
- Family System *745*
- Family Cycle *745*
- Family Types *745*
- Family Functions *746*
- Role of Family in Health and Disease *747*
- Role of Social and Cultural Factors in Health and Disease *748*
- Role of Emotions in Health and Disease *749*
- Social Process *749*
- Poverty—Its Relation with Health and Disease *753*
- Below Poverty Line *754*
- Social Factors in Health and Disease (Social Pathology) *755*
- Management or Prevention of Social Pathology *759*
- Operational Research *761*
- Medicosocial Worker *761*
- Personal Hygiene *761*
- Refugee Health *762*
- Social Psychology *763*

Chapter 28: Information, Education and Communication 765
- Information *765*
- Education *765*
- Communication Cycle *765*
- Steps in IEC: Planning, Implementation, and Evaluation *766*
- Types of Communication *766*
- Barriers of Communication *766*
- Health Education *767*
- Social Mobilization *774*
- Domains of Learning Process *775*
- Soap Technique *775*
- Body Language *775*
- Anatomy of Lesson Plan *775*
- Evaluation of Health Education Program *775*
- Counseling *776*

Chapter 29: Human Genetics 779
- Multiplication *779*
- Chromosomes *779*
- Deoxyribonucleic Acid *780*
- Population Genetics *780*

- Classification of Genetic Disorders *781*
- Gene Therapy *784*

Chapter 30: Preventive Geriatrics 786
- Population of the Aged *786*
- Theories of Aging Process *786*
- Health Problems of the Aged *787*
- Geriatric Counseling *787*
- Steps Taken by Government of India for the Elderly *788*
- Prevention of Health Problems Among Aged *789*
- Elder Abuse *790*
- Old Age Social and Income Security *791*
- Helpage India *792*
- National Policy on Older Persons *793*
- Benefits Given to Senior Citizens of India *793*

Chapter 31: Mental Health 795
- Magnitude of the Problem *795*
- Causative Factors (Risk Factors) *795*
- Types of Mental Disorders *796*
- Prevention and Control of Mental Illness *797*
- Primary Mental Health Care *798*
- Community Psychiatry *799*
- Mental Health Laws in India *800*

Chapter 32: Adolescent Health 802
- Changes During Adolescence *802*
- Challenges in Adolescent Health in India *803*
- Impacts of Adolescence *803*
- Adolescent Health Problems *803*
- Services in Adolescent Health Clinics *804*
- Recommendations *804*
- Conclusion *804*
- Adolescent Reproductive and Sexual Health *804*
- Adolescent Friendly Health Clinic *805*
- "All in" Initiative *806*
- Rashtriya Kishor Swasthya Karyakram *806*
- Navya Initiative *807*

Chapter 33: Alcoholism and Drug Addiction 809
- Definition *809*
- Addiction as a Disease *809*

Section VII: Health Administration and Organization

Chapter 34: Tribal Health 817
- Recommendations *819*

Chapter 35: Health for All 823
- Meaning *823*
- Historical Perspective *823*
- Importance *823*
- Milestones *823*

Chapter 36: Millennium Development Goals 825

Chapter 37: Sustainable Development Goals 829
- Background *829*
- Health Goal and Health Targets *831*
- One Health Approach *831*

Chapter 38: Health Care 834
- Three-tier System of Health Care *834*
- Selective Primary Health Care *836*
- Universal Health Coverage *837*
- Universal Health Coverage in India *838*
- Health Services Research *839*
- Organizational Framework of Healthcare Systems *840*
- Provision of Primary Health Care in Rural India *852*
- Indian Public Health Standards for Sub-centers (Revised 2012) *853*
- Health and Wellness Centres (Ayushman Arogya Mandir) *861*
- Role of Computers in Health Care *863*
- Essential Medicines/Drugs *866*
- Counterfeit Medicines *867*
- Central Drug Standard Control Organization *868*

Chapter 39: National Health Mission 870
- National Rural Health Mission *871*
- National Urban Health Mission *876*
- Autodisable Syringes *878*
- Schemes *879*
- Maternal Child Health and Nutrition Day *892*

Chapter 40: Health Planning and Management 894
- Health Planning *894*
- Health Information Technology *898*
- Health Budgeting *900*
- Evaluation of Health Services *900*

Chapter 41: National Health Planning 902
- Bhore Committee Report, 1946 *902*
- Mudaliar Committee Report, 1962 *903*
- Chadha Committee Report, 1963 *903*
- Mukherjee Committee Report, 1965 *903*
- Mukherjee Committee Report, 1966 *903*
- Jungalwalla Committee Report, 1967 *903*
- Kartar Singh Committee Report, 1973 *903*
- Srivastav Committee Report *904*
- National Developmental Plans *904*
- Niti Aayog *907*
- Rural Development *908*
- Poverty Alleviation Programs in India *909*

Chapter 42: National Health Policy 911
- National Digital Health Mission *915*

Chapter 43: National Voluntary Health Agencies/Organizations 917
- Voluntary Health Agencies *917*
- Voluntary Health Association of India *917*
- Voluntary Health Agencies in India *918*

Chapter 44: International Health Organizations 920
- Concept of International Health *920*
- World Health Organization *920*
- United Nations International Children's Emergency Fund *923*
- United Nations Development Programme *925*
- Food and Agricultural Organization *925*
- International Labor Organization *925*
- United Nations Fund for Population Activities *926*
- United Nations Educational, Scientific and Cultural Organization *926*

- United Nations High Commission for Refugees *926*
- World Bank *927*
- United Nations Joint Programme on AIDS *927*
- World Food Programme *927*

Chapter 45: Bilateral Agencies 928
- Nongovernment (Nonunited Nations) Agencies *928*
- Other Voluntary Organizations *931*

Chapter 46: National Health Programs 932
- National Filaria Control Programme *938*
- National Kala-Azar Control Programme *940*
- National Kala-Azar Elimination Programme *940*
- National Japanese Encephalitis Control Programme *941*
- National Dengue Fever or Dengue Hemorrhagic Fever Control Programme *941*
- National Leprosy Control Programme *942*
- National Leprosy Eradication Programme *942*
- National Guinea Worm Eradication Programme *943*
- Universal Immunization Programme *943*
- Polio End Game Strategy (2019–2023) *943*
- National Tuberculosis Elimination Programme *947*
- Who 1+1 Initiative *957*
- National Acute Respiratory Infections Control Programme *958*
- National Diarrheal Diseases Control Programme *958*
- National Aids Control Programme *958*
- '3 By 5' Initiative *966*
- National Framework for Joint HIV/TB Collaborative Activities *966*
- National Strategic Plan for HIV/Aids and STI, 2017–2024 *967*
- National Action Plan Combating Viral Hepatitis in India and National Viral Hepatitis Control Program *971*
- National Yaws Eradication Programme *973*
- National Programme for the Control of Blindness and Visual Impairment *973*
- National Programme for the Control of Blindness and Visual Impairment Achievement *975*
- National Programme for Prevention and Control of Deafness *976*
- National Programme on Prevention and Control of Cancer, Diabetes, Cardiovascular Diseases and Stroke *977*
- National Mental Health Programme *982*
- District Mental Health Programme *984*
- National Iodine Deficiency Disorders Control Programme *986*
- National Programme for Control and Treatment of Occupational Diseases *987*
- National Programme on Prevention and Control of Fluorosis *988*
- National Programme for Prevention and Management of Trauma and Burn Injuries *990*
- National Vitamin 'A' Prophylaxis Programme *991*
- National Nutritional Anemia Control Programme *992*
- National Special Nutrition Programme *992*
- National Balwadi Nutrition Programme *992*
- National Mid-day School Meal Programme *992*
- National Integrated Child Development Services Scheme *992*
- Poshan Abhiyaan *992*
- National Programme for Health Care of the Elderly *994*
- National Tobacco Control Programme *996*
- National Family Welfare Programme *998*
- National Family Planning Indemnity Scheme *999*
- Reproductive and Child Health Program *1000*
- All India Hospital Postpartum Programme *1001*
- National Water Supply and Sanitation Programme *1002*
- Minimum Needs Programme *1002*
- Twenty Point Programme *1003*
- National Oral Health Programme *1004*
- National Clean Air Programme *1005*

Chapter 47: Community Development Programs of India 1009
- Characteristics *1009*
- Importance of CDP *1009*
- Aim of CDP *1009*
- Objective of CDP *1009*
- Organization *1010*
- Rural Development Schemes *1010*

SECTION VIII: Allied Subjects

Chapter 48: Emerging and Re-emerging Infectious Diseases 1015
- Emerging Infectious Diseases *1016*
- Re-emerging Infectious Diseases *1034*

Chapter 49: Disaster and Disaster Management 1036
- Disaster *1036*
- Disaster Management *1038*
- National Disaster Management Authority *1040*
- Floods and Health *1040*

Chapter 50: Integrated Disease Surveillance Project: 2004–2009 1043
- Surveillance *1043*

Chapter 51: Bioterrorism 1046
- Historical Background *1046*
- Evolution of Chemical and Biological Weapons *1046*
- Combating Biological Weapons Incident/At-Risk Group *1048*

Chapter 52: Global Warming 1049
- History of Global Warming *1049*
- Hazards of Global Warming *1049*
- Who is the Worst Polluter? *1050*
- Kyoto Protocol *1051*
- Action Needs to be Taken Now *1051*

Chapter 53: Integrated Management of Neonatal and Childhood Illness .. 1052
- Strategy *1052*
- Basis of IMNC Guidelines *1052*
- Integrated Case Management Process *1052*

Chapter 54: Telemedicine in Public Health 1054
- Definition *1054*
- Aim of Telemedicine *1054*
- Utility of Telemedicine *1054*
- Types of Technology *1055*
- Infrastructures of Telemedicine Centers *1055*
- Telecommunication Technologies *1055*
- An Epidemiological Surveillance *1056*
- Interactive Health Communication and Disease Prevention *1056*
- Telemedicine in India *1056*

- Current Efforts in India *1057*
- Ten Commandments for Success of Telemedicine *1057*
- Artificial Intelligence in Health Care *1058*

Chapter 55: Tobacco and Health 1062
- History *1062*
- Tobacco Prevalence *1062*
- Prevalence of Smoking *1062*
- Public Health Importance *1062*
- Types of Tobacco *1063*
- Composition of Smoke *1063*
- Health Hazards of Tobacco *1064*
- Health Benefits of Smoking *1065*
- Prevention and Control of Health Hazards of Smoking *1066*
- Legislative Measures *1066*
- National Tobacco Control Programme *1067*

Chapter 56: Internet and Health 1068
- Positive Effects (Merits) *1068*
- Negative Effects (Health Hazards) *1068*
- Features of Internet Addiction *1069*
- Other Effects *1069*
- Measurement *1069*
- Treatment *1069*
- Therapies *1069*

Chapter 57: Laws Pertaining to Practice of Medicine....1070
- The Clinical Establishments (Registration and Regulation) Act, 2010 *1072*
- The Consumer Protection Act, 2019 *1073*
- Registration of Births and Deaths Act, 1969 *1073*
- Preconception and Prenatal Diagnostic Techniques Act, 1994 (PC-PNDT Act, 1994) *1074*
- Protection of Children from Sexual Offences Act, 2012; Amendment 2019 *1075*
- Transplantation of Human Organs and Tissues Act, 2014 *1076*

Annexure 1079
- Stalwarts of Public Health/Medicine and their Honors *1079*
- Some Stalwarts and their Contribution to Community Medicine *1079*
- Edward Jenner (1749–1823) *1080*
- Sir Edwin Chadwick (1800–1890) *1080*
- Louis Pasteur (1822–1895) *1080*
- Joseph Lister (1827–1912) *1080*
- Robert Koch (1843–1910) *1080*
- Ronald Ross (1857–1932) *1081*
- Calendar of Important Dates and Events *1081*
- List of World Health Day Themes *1083*
- Milestones in Vaccination *1083*

Index ***1085***

CHAPTER

26 Biostatistics

JAYPEE BROTHERS

INTRODUCTION

Statistics is a methodology of collection, compilation, analysis, and interpretation of numerical facts.

Study of statistics in relation to biological science (such as biological, social, and environmental factors) is known as "biostatistics" (medical statistics).

Study of statistics in relation to health and diseases of human population and the different factors related to them, is known as "health statistics."

Study of statistics in relation to the vital events of life, such as births, deaths, marriages, and divorces is known as "vital statistics," which in turn is a branch of "demography," which deals with the study of human populations.

In medicine the diagnosis and treatment depends upon collection of data in numerical figures, such as recording blood pressure (BP), serum cholesterol level, hemoglobin (Hb) level, etc. Thus medical statistics or biostatistics is often called "quantitative medicine."

VARIABILITY

An inherent feature of all biological observations is their variability, e.g., every individual varies with one another. Similarly, each group of individuals is different from other groups. For example, the pulse rate, Hb level, and the number of white cells varies from person to person. So, a range is always expressed as a standard or normal value. Again this varies from one group to another. For example, pulse rate among infants varies from that of older age-group.

The measurable variable, such as height, weight, BP, glucose, or cholesterol level in blood, etc. is called "quantitative variable" and the nonmeasurable variable, such as sex, occupation, socioeconomic status, etc. is called "qualitative variable."

APPLICATIONS OF BIOSTATISTICS

- Helps in effective comparisons between two groups or two countries.
- Helps in measurement of health status of a community in terms of rates, ratios, proportions, etc., which in turn helps in comparison with other countries and helps to study the influencing factors.

  For example, the prevalence of typhoid fever is higher among people of poor socioeconomic status, living in unhygienic areas with unsafe water supply, not protected through immunizations, and so on. Thus, by a systematic analysis of the factors related to the disease, a health worker or a health administrator can define the problems in terms of contrasts.
- Helps in estimating the magnitude of a health problem.
- Helps in analyzing the causes of the public health problems, including epidemics, to the public health personnel.
- Helps in monitoring and evaluation of the control measures and also in introducing midcourse correction measures, wherever necessary.
- Helps in health planning and management.
- Helps in research purposes.

Thus biostatistics, if properly recorded, constitutes "eyes and ears" of a health worker/administrator, otherwise it would be like "sailing in a ship without compass."

DATA—DEFINITION, TYPES AND SOURCES

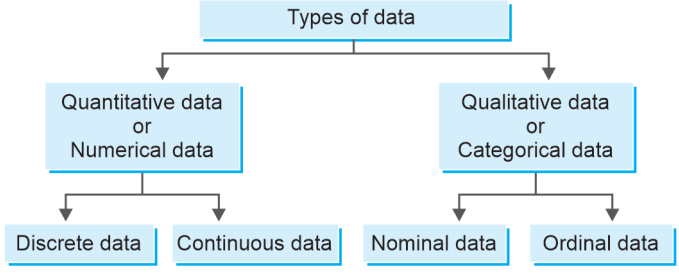
Definition

Data is defined as facts or figures or information or observation that is collected, analyzed interpreted and presented, upon which conclusion can be drawn, in research activities. The most critical objective of data collection is ensuring that the information collected is rich and reliable.

Types

Mainly there are two types of data—quantitative and qualitative; in each type, there are two subtypes, as shown in **Flowchart 26.1**.

Flowchart 26.1: Types of data.



Quantitative (Or Numerical Data)

This is a measure of numerical values, that can be counted or measured. This refers to numerical data, e.g., height, weight, length. This helps to understand how many, how much and how often about the information (e.g., How many attended the webinar? How much revenue did the company make? How often the customers use online banking? etc.).

There are two subtypes.

1. **Discrete data:** This is only fixed discrete values, containing only a finite number, which cannot be sub-divided meaningfully. Data is countable, e.g., number of students in a room, 32. It is fixed and finite; number of speakers in a phone, number of questions answered correctly, etc.
2. **Continuous data:** This has an infinite number of probable values that can be selected within a given specific range. The data is measurable, e.g., RBC count, WBC count. It can take any value between two numbers. Other examples are height of children, weight of a truck, speed of cars, amount of rain in inches/mm.

Qualitative Data

This is descriptive. That means it describes the data that fits into categories. It refers to the things that can be observed but not measured, e.g., a person's gender, color of a phone, postbox color is red, etc. There are two subtypes—nominal data and ordinal data.

1. **Nominal data:** This is used to categorize certain variables, e.g., blue, brown or black to describe the eye color; different nationalities, different languages spoken, etc. These are not measured but categorized.
2. **Ordinal data (or variable data):** This data follows a natural order. This is also observed but not measured. This data is also categorized, e.g., size of the clothes as small, medium and large, opinion as agree disagree or neutral, grading of students as A,B,C and D, customer satisfaction as poor, satisfactory or good, etc.

Sources of data: There are mainly two sources—primary and secondary.

1. **Primary source:** It is a fresh data, collected for the first time directly from the main source and thus happens to be original information, e.g., data from experiments interviews, surveys, etc.
2. **Secondary data:** This is the data collected by someone else earlier, e.g., report of previous surveys, information from books, journals, records, etc. Secondary data is derived from the primary data.

Factors to be considered before collecting secondary data are reliability, suitability and adequacy of the data.

The differences between primary and secondary data is shown in **Table 26.1**.

Differences between data and information is shown in **Table 26.2**.

TABLE 26.1: Differences between primary and secondary data.

Primary data	*Secondary data*
1. It is a first hand information	It is a second hand information
2. It is original, accurate and reliable	It is not original, not accurate and not reliable
3. Data is not easily accessible	Data is easily accessible
4. Data is collected directly by the investigator from interviews, surveys or experiments	Data is collected indirectly from the records, journals, books, etc.
5. Data is homogeneous	Data is heterogeneous
6. Not undergone statistical treatment	Data is already undergone statistical treatment
7. Data collection is time consuming and expensive	It consumes less time and is less expensive
8. Personal bias is possible	Personal bias is minimized

TABLE 26.2: Differences between data and information.

Sl. No.		*Data*	*Information*
1.	Description	It is unorganized, unrefined, randomly collected raw facts	It is the result of processing and analyzing the raw data
2.	Meaning	Data does not carry any meaning	Information carries a logical meaning
3.	Usefulness	It is used as input	It is the output of processed data
4.	Representation	It is represented as numbers, figures, symbols or graphs	It is represented as language
5.	Measuring units	Measured as bits and bytes in the context of computer storage	Measured in meaningful units
6.	Dependency	No dependency	Depends on data
7.	Peculiarity	It is vague	It is specific
8.	Significance	Data alone is not significant	It is significant by itself
9.	Decision making	Does not help in decision making	Helps in decision making
10.	Context	Data has no context	Information has context
11.	Example	Recording temperature in a location over several years	Analysis of the data helps to determine seasonal temperature pattern

Sources of Health Statistics

The different health statistics and their sources are shown in **Table 26.3.**

The drawbacks of hospital records are as follows:

- The hospital records represent only the tip of iceberg, because all cases do not come to hospital. Thus the mild and subclinical cases are missed. So it cannot be a representative sample of the population.

JAYPEE BROTHERS

TABLE 26.3: Sources of health statistics.

Health statistics	*Sources*
Population	• Census reports; population estimates
Births	• Register of births • Baptism records at churches
Deaths	• Registers of deaths • Registers at the burial and cremation grounds • Postmortem records
Diseases	• Hospital and dispensary records • General practitioners records. Records of health and welfare centers • Records of educational institutions • Records of recruitment and sickness at armed forces • Industrial absenteeism records. Records of social security schemes (such as ESI, LIC, contributory health service schemes, etc.) • Records of notifiable diseases. Reports of routine and special sickness surveys

- Hospital statistics tend to be selective, because admission policy varies from hospital to hospital.
- As there is no specific catchment area and no geographical boundary, hospital statistics provide only numerator and no denominator. Therefore, it is a poor guide to estimate disease frequency.

Record Linkage

Record linkage is a process of linking or consolidating the fragmented data of the same individual or family or an event, collected from different records, to get an accurate and comprehensive profile about health and disease. It is used mainly in epidemiology, demography and medical research.

Objectives

- To compare the records,
- To determine the likelihood that a pair of records refers to the same individual or entity,
- To classify the records as matched pair or non-matched pair.

Types

- **Deterministic linkage:** In this type, the records are linked, collected from unique records, such as Aadhar card number, hospital records, insurance data, census data, vital statistical data (such as birth rate, death rate), etc.
- **Probabilistic linkage:** In this type, the records are linked, collected using name, age, gender, date of birth, address, etc., when unique identifiers are not available.

Uses

- Linking data from various sources provides a complete health profile.
- Helps to study long-term outcomes of diseases (e.g., survival rate in cancer, etc.) and of drug exposure.
- Avoids duplication in health records.
- Helps to identify the risk factors/causes of death.
- Helps in better tracking of diseases and immunization coverage.
- Provides accurate epidemiological data and helps in research of cohort studies, and clinical trials.
- Helps with the evaluation of the effectiveness of health programs.

Limitations

- Errors may be present, making matching difficult.
- Confidentiality issues of the sensitive data.
- Requires high quality of record keeping system.
- Requirement of efficient algorithm to handle massive data sets.

Examples

- **Immunization tracking:** A child's vaccination data is collected from PHC register, School health records, national immunization database. Linking these data ensures that no dose is missed or duplicated. This helps in improving coverage monitoring.
- **Maternal and child health:** Linking hospital records, Ante-natal care card, census and vital registration, helps in getting a complete health profile of MCH outcomes like maternal mortality rate, infant survival, etc.
- **Cancer registration:** Details of cancer patients collected from hospital admissions, discharge card, laboratory records and death certificates. Record linkage brings these data together completes cancer registry and helps in studying the survival rate, effectiveness of treatment and causes of death.

PRESENTATION OF STATISTICAL DATA

The raw data, collected from different sources, is rearranged (or processed) in a meaningful and understandable way. The data can be presented in several methods—as tables, charts, diagrams, pictures, and special curves (or graphs).

Tabulation

This means presentation of the data in the form of tables.

This is the most common and preliminary way of presentation of data.

There are different types of tables depending upon the nature of data and purpose of tabulation. The table can be qualitative or quantitative. Whether qualitative or quantitative table, the basic principles of tabulation of data are:

- The information should be in a simple, unambiguous, and orderly manner.
- The table should have a title, which must be brief and comprehensive reflecting the nature and contents of the table. Title should be at the top or bottom of the table.
- Rows and columns must have their own captions (titles).

JAYPEE BROTHERS

- The titles of the rows must be entered on the left side of the table, constituting the stub. The rest of the table constituting the body may contain numerical information in actual numbers, in percentages or in both forms.
- Standard codes or symbols are to be used, wherever necessary, and are explained as footnote.
- When the data tabulated is not original, the source of the data is to be mentioned at the bottom of the table.
- Number of class intervals should depend upon the aims of presentation.
- The class intervals are usually taken at equal intervals, e.g., 1–5, 6–10, 11–15, 16–20, etc.
- Units of the measurements of the data are to be specified.

With the above principles, the table becomes, as self-explanatory, as possible.

Qualitative Tables

There are three types of qualitative tables, namely:

1. One-way table, which depicts the distribution of data in relation to a single attribute (quality, characteristic; variable) **(Table 26.4)**.
2. Two-way table, which depicts the distribution of data in relation to two attributes **(Table 26.5)**.
3. Higher order table, which depicts the data in relation to more than two attributes **(Table 26.6)**.

TABLE 26.4: Distribution of leprosy patients according to blood group (one-way table).

Blood group	*Leprosy patients*	*Percentage*
A	26	21.7
B	20	16.7
AB	16	13.3
O	58	48.3
Total	120	100.0

TABLE 26.5: Distribution of leprosy patients by type and their blood group (two-way table).

Blood group	*Lepromatous leprosy no. (%)*	*Nonlepromatous leprosy no. (%)*	*Total no. (%)*
A	06 (13.9)	20 (25.9)	26 (21.7)
B	08 (18.6)	12 (15.6)	20 (16.7)
AB	07 (16.3)	09 (11.7)	16 (13.3)
O	22 (51.2)	36 (46.8)	58 (48.3)
Total	43 (100)	77 (100)	120 (100)

TABLE 26.6: Distribution of leprosy patients by type and gender (higher order table).

	Lepromatous leprosy patients		*Nonlepromatous leprosy patients*		
Blood group	*M*	*F*	*M*	*F*	*Total*
A	4	2	12	8	26
B	5	3	8	4	20
AB	4	3	6	3	16
O	15	7	20	16	58
Total	28	15	46	31	120

Quantitative Tables

When a large data is collected by measurement, it is presented in the form of a table showing the characteristics as well as the number of observations (i.e., frequency), both of which are variable. The characteristic may be discrete or continuous (indiscrete). Such tables are called frequency tables or frequency distribution tables **(Table 26.7)**.

In the above table the characteristic forms are whole group and is not split into subgroups because there is no range of variability. Illness means illness and no fractions or parts are there. This is an example of discrete frequency distribution table (i.e., no of observations are made against discrete variables).

If the data on the other hand were to be a continuous data, such as Hb% level, height, and weight, etc. having a range, from lowest to the highest. This range is subdivided into subranges and class frequency is noted opposite each group.

The interval in each subrange is called "class interval" and is equal (uniform) throughout. This is also called indiscrete frequency distribution table. For example, frequency **(Table 26.8)** for heights of 80 boys, varying from 160 cm to 170 cm, with a class interval of 2 cm as follows. The number of boys (class frequency) with the heights falling in a particular group is recorded opposite. Percentage of each frequency is calculated and mentioned in the last column.

Cumulative frequency table: In this table, there is an additional column, giving the total frequencies up to each of the class interval, from the lowest value. It is obtained by adding all the frequencies up to and above that particular class interval **(Table 26.9)**.

The above frequency distribution table is converted into cumulative frequency distribution table as follows.

TABLE 26.7: Frequency distribution of students in a hostel according to the number of illness suffered in a year.

No. of illnesses (discrete frequency)	*No. of students*	*Percentage*
0	24	4.7
1	76	14.9
2	114	22.4
3	115	22.5
4	86	16.9
5	51	10.0
6	26	5.1
7	18	3.5
Total	510	100.0

TABLE 26.8: Frequency distribution of heights of boys.

Height (cm) with class interval	*Frequency of each group*	*Percentage*
160–162	10	12.5
162–164	15	18.8
164–166	16	20.0
166–168	19	23.7
168–170	20	25.0
Total	80	100.0

JAYPEE BROTHERS

TABLE 26.9: Cumulative frequency distribution of heights of boys.

Heights (cm)	*Frequency*	*Cumulative frequency*
160–162	10	10
162–164	15	25
164–166	16	41
166–168	19	60
168–170	20	80
Total	80	

Diagrammatic Presentation

Diagrammatic presentation consists of projecting the statistical material in geometric figures, pictures, maps, which are simple, attractive, and easily understandable even to common man, thereby facilitates the communication of statistical information by a quick grasp.

Limitations of diagrammatic presentation are:

- Great care must be taken in designing the diagram
- Every aspect of the data cannot be presented in the diagram
- They fail to reflect small differences and emphasize only major characteristics
- They fail to provide basis for application of tests.

Let the diagrammatic representation plays an important role in making intricate and perplexing statistical facts intelligible to common people.

Points to be Remembered in the Construction of a Diagram or Graph

- Diagram should be simple and consistent with the type of data
- Diagram should be self-explanatory, including title, scale, index
- Diagram should not look clumsy by too many lines
- Usually the rulings of the graph paper should be light and the lines of the diagram should be heavier
- Ordinarily the values of the variable are represented on the horizontal *X*-axis and the values of the frequency on the vertical *Y*-axis (respectively called abscissa line and ordinate line)
- The lines drawn on the graph should never be extrapolated beyond the range of the values of the variable to which the graph is drawn.

The various types of diagrammatic presentation are:

- Bar diagrams/charts
- Histogram
- Frequency polygon
- Line diagram
- Frequency curve
- Cumulative frequency curve
- Scatter diagram
- Pie diagram
- Pictogram
- Map diagram.

Bar diagrams/charts: It is a simple diagram or a chart, popularly used to compare the magnitude of the qualitative data. Bars are the rectangles drawn along the graph sheet. Height of each bar or rectangle corresponds to frequencies in a data, while the breadth or base corresponds to the length of the class interval of the variable in quantitative data and to various groups in qualitative data. The bars are arranged either vertically or horizontally in ascending or descending order. Bars should stand on the same line and possess the same width and the spacing between the two consecutive bars should be equal to the half of the width of the bar. The bars can be placed either vertically or horizontally.

By comparing the height of the rectangles, comparison of different groups can be made.

A suitable scale must be chosen to present the length of the bars.

There are three types of bar diagrams—(1) simple, (2) multiple, and (3) component (proportional).

1. **Simple bar diagram:** It consists of single, rectangular bars having uniform color or shade **(Fig. 26.1)**.
2. **Multiple bar diagram:** It is also called compound bar diagram/chart. In this type, two or more bars can be grouped together **(Fig. 26.2)**. Such diagrams are useful to make comparisons between various groups. Multiple bars may have uniform or different shades.
3. **Proportional bar diagram:** This is also called "component bar diagram." In this type the bars may be divided into two or more parts, depending upon the number of subgroups to be compared in proportions or percentages. Each bar representing 100% as their height, each component representing the magnitude of an attribute or variable. A component bar diagram resembles a simple bar externally and multiple bars internally. Since the diagram represents the percentage, they are often referred to as "percentage bars."

In **Figure 26.3** the percentage of admission in each ward for a particular year is depicted within each bar.

Histogram: In Greek language "Histos" means web. Just like the toes of frog are webbed, so also in histogram the bars are joined together. It is a graphic representation of a frequency distribution table in which the vertical axis represents the frequency and the horizontal axis the class interval.

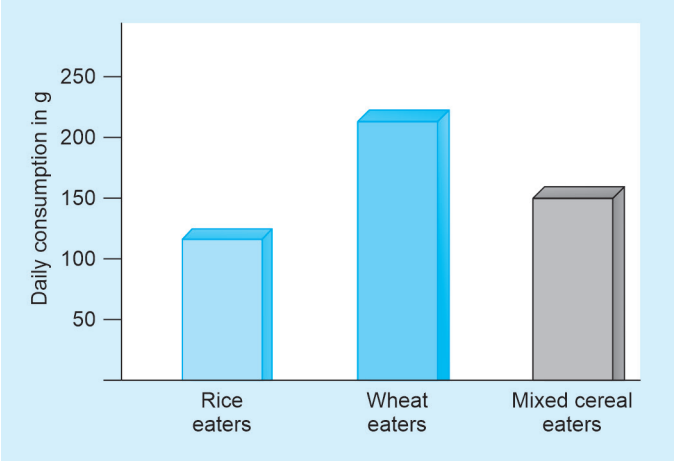


Fig. 26.1: Bar diagram showing type and quantity of cereals consumed by 500 hostelers.

JAYPEE BROTHERS



Fig. 26.2: Multiple bar diagram showing cereal consumption by quantity and type in inmates of two hostels.

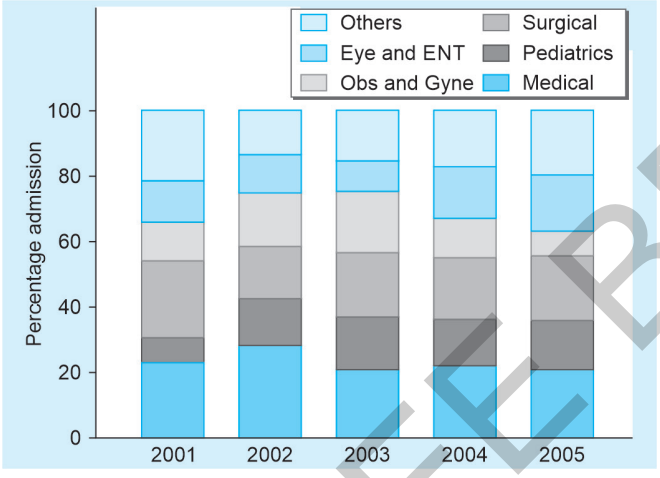


Fig. 26.3: Proportional bar diagram showing the number of admissions in a teaching hospital during 2001–2005 according to specialties.

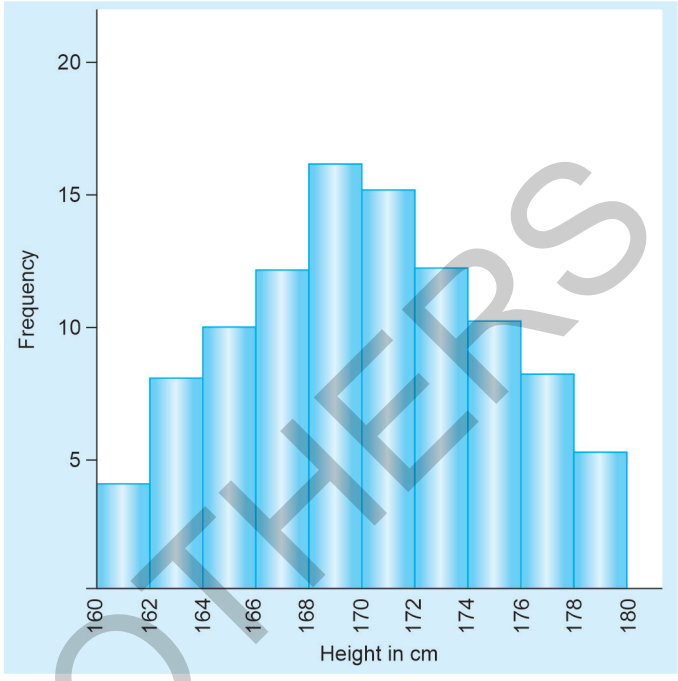


Fig. 26.4: Histogram showing the heights amongst students in a class.

It consists of series of bars adjoining to each other, the length of each bar is being proportional to the frequency and the width to the class interval **(Fig. 26.4)**.

The concept in this graph is that the area under each rectangle represents the proportion of the total area of all the rectangles, considering as one.

Histograms are ideally suited to represent the distribution of anthropometric values, such as height, weight, midarm circumference, etc. They can also represent other types of continuous data series, such as BP, pulse rate, and Hb level.

Histograms provide a better understanding of quantitative data of continuous type than frequency distribution tables.

Age and sex pyramid (population pyramid): this is a double histogram, wherein the bars are drawn on either side of a vertical axis, which forms the base, and the bars are adjoining to each other with a common scale. The left side represents the age-wise distribution of the female and the right-side, age-wise distribution of the male population. Thus, the percentage distribution or the number of persons in any age-group is taken on the horizontal axis, and the variable, i.e., age-group, is taken on the vertical axis in the center **(Fig. 26.5)**.

The age and sex pyramid reflects several features. The height of the pyramid represents the maximum age span of the population. The shape of the pyramid reflects the fertility behavior of the population (it has a broad base, concave outer borders, and acute angle of apex for the developing countries and a narrow base, convex outer border, and obtuse apical angle for the developed countries). Asymmetry of the pyramid indicates the presence of gender bias.

Frequency polygon: It is a line diagram that represents a frequency distribution table. It can be obtained by joining the midpoints (dots) of the heads (height) of the histogram, each dot represents two characteristics: (1) class interval as indicated on the horizontal axis and (2) the class frequency as indicated on the vertical axis.

Joining the dots gives a curve with many angles. Hence, the name "frequency polygon" **(Fig. 26.6)**.

This type of diagram is useful, especially when it is necessary to compare two or more frequency distributions. The curves for different distributions should be drawn with different types of lines on the same graph paper for easy comparison, which is not possible through histograms because of overlapping of rectangles resulting in confusion.

Frequency curve: When the number of observations is very large and the group interval is reduced, the frequency polygon tends to lose its angulations, resulting in a smooth curve, known as "frequency curve" **(Fig. 26.7)**.

JAYPEE BROTHERS

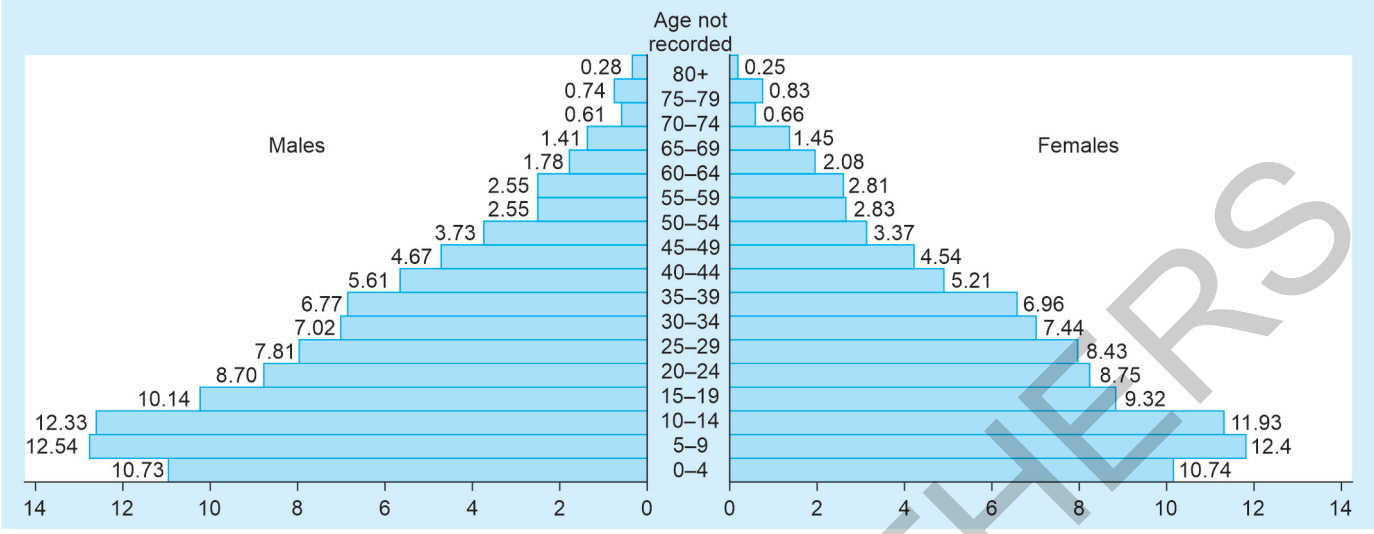


Fig. 26.5: Age and sex pyramid of India 2001 (histogram showing the population pyramid).
Source: Sunderlal, Adarsh, Pankaj. Textbook of Community Medicine, 1st edition, 2007.

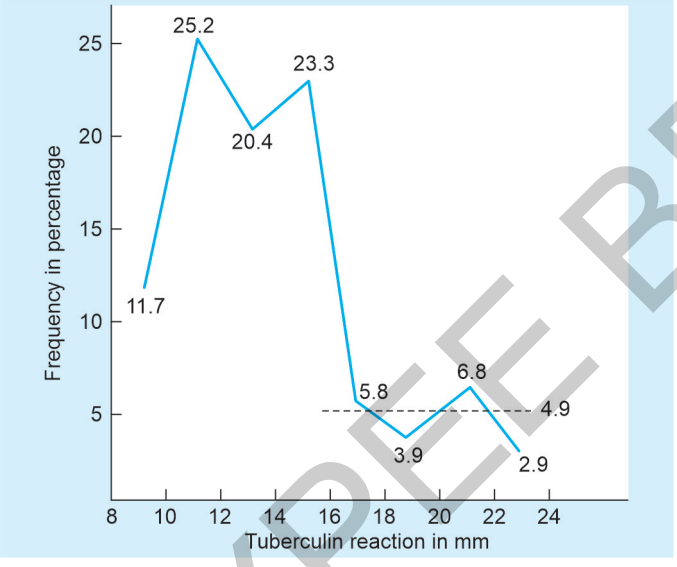


Fig. 26.6: Frequency polygon showing tuberculin reaction in 206 persons, never vaccinated.
Source: Mahajan BK. Methods in Biostatistics, 3rd edition, 1981.

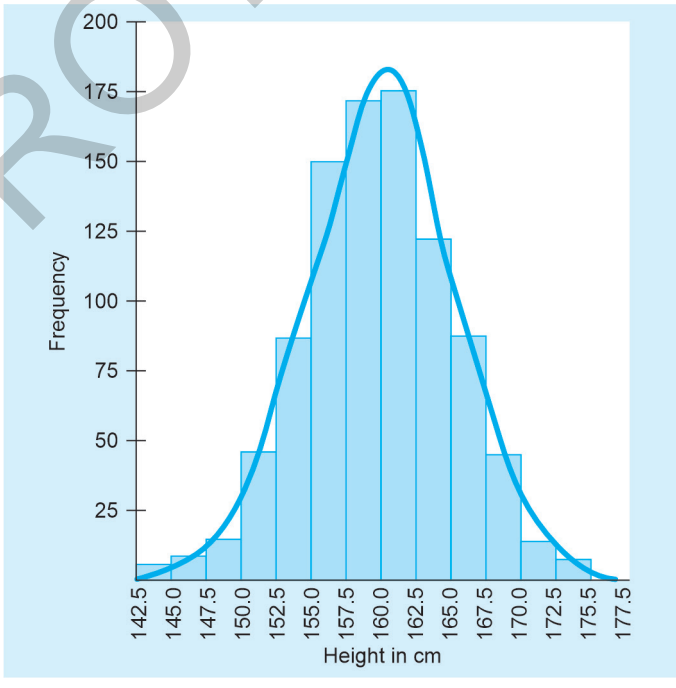


Fig. 26.7: Frequency curve of heights superimposed over histogram.
Source: Mahajan BK. Methods in Biostatistics, 3rd edition, 1981.

Line diagram: In this diagram the vertical axis represents the magnitude and the horizontal axis represents time. Thus, this diagram provides a simple, easily understandable, and highly effective means of understanding the trend or behavior of an event over a period of time, e.g., rising or falling or fluctuations, such as birth rate, death rate, population trend, etc.; the class interval may be 1 month, 1 year, or one decade **(Fig. 26.8)**.

Since line diagrams do not occupy any space, several lines may be projected on the same graph for comparing the trends of interrelated events.

Multiple line diagrams can coexist only if they share the scales given on the two axes of the graph.

Cumulative frequency curve or "Ogive": It is a line diagram, representing the cumulative frequency distribution of the quantitative data. To draw this, an ordinary frequency table in a quantitative data has first to be converted into a cumulative frequency table. Cumulative frequency is the total number of persons in each particular range from the lowest value of the characteristic up to the any higher group value **(*see* Table 26.8)**.

The curve is plotted by taking the variable on the *X*-axis and the cumulative frequency on the *Y*-axis **(Fig. 26.9)**.

JAYPEE BROTHERS

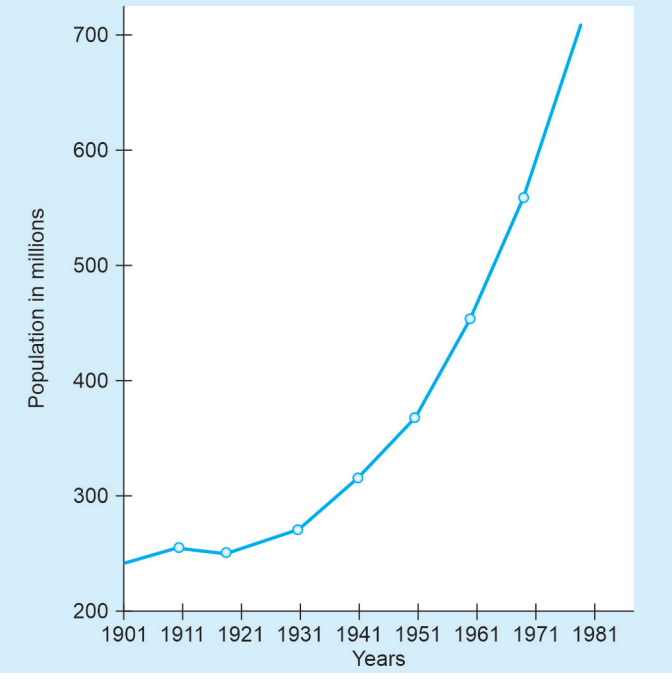


Fig. 26.8: Line chart showing the population trend in India.
Source: Mahajan BK. Methods in Biostatistics, 3rd edition, 1981.

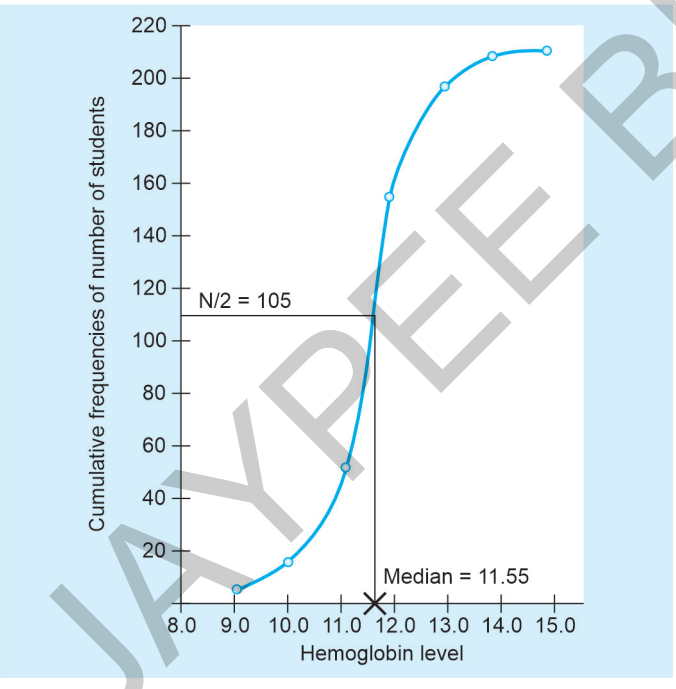


Fig. 26.9: Ogive for the distribution of hemoglobin levels of students in a class.
Source: Rao NSN. Elements of Health Statistics, 1st edition. Reprint 1985.

From the ogive, median value of the characteristic (variable) can also be calculated. The calculation and the application will be discussed under "central tendency."

Scatter diagram or dot diagram: All the diagrams described so far are useful only for frequency distributions with single variables. When observations for two variables (e.g., weight and midarm circumference or weight and height) are made in each of the individuals in a group, it helps to study the relationship between the two variables. One variable is represented on X-axis and another variable on Y-axis. Perpendiculars drawn from these readings meet to give one scatter point. There will be as many points as there are individuals in the observation. When all the points are plotted, the diagram gives the picture of a scatter. Hence, the name "scatter diagram" (dot diagram). The direction of scatter helps to determine presence or absence of the association. There are 3 types of scatter diagrams. In the first type, high values of x_1 (e.g., heights) are associated with high values of x_2 (e.g., weight) and low values of x_1 with low values of x_2. This is "positive correlation" **(Fig. 26.10A)**. In the second type, high value of x_1 (e.g., income) are associated with low values of x_2 (e.g., tuberculosis) and low values of x_1 are associated with high values of x_2. This is "Negative correlation" **(Fig. 26.10B)**. In the third type, there is no association between the variables x_1 (e.g., height) and x_2 (e.g., pulse rate). The point are highly scattered. This is "No Correlation" **(Fig. 26.10C)**.

Pie diagram: It is an improvement over a bar diagram. It is an another way of depicting qualitative data, such as blood groups, age-group, sex groups, morbidity, mortality rates, etc. in a population. It is a circular diagram in which the frequencies of observations are shown as sectors or wedges in a circle, the size of each sector being proportional to the frequency. Degrees of angle denote the frequency and the area of the sector gives comparative difference at a glance. ("Pie" means a piece or a sector).

Before constructing a pie diagram, angles are distributed among the various frequencies in proportion to their individual magnitudes, treating the total magnitude of the attributes (frequencies) as 360°; 360° is representing the sum total of all angles at the center of the circle.

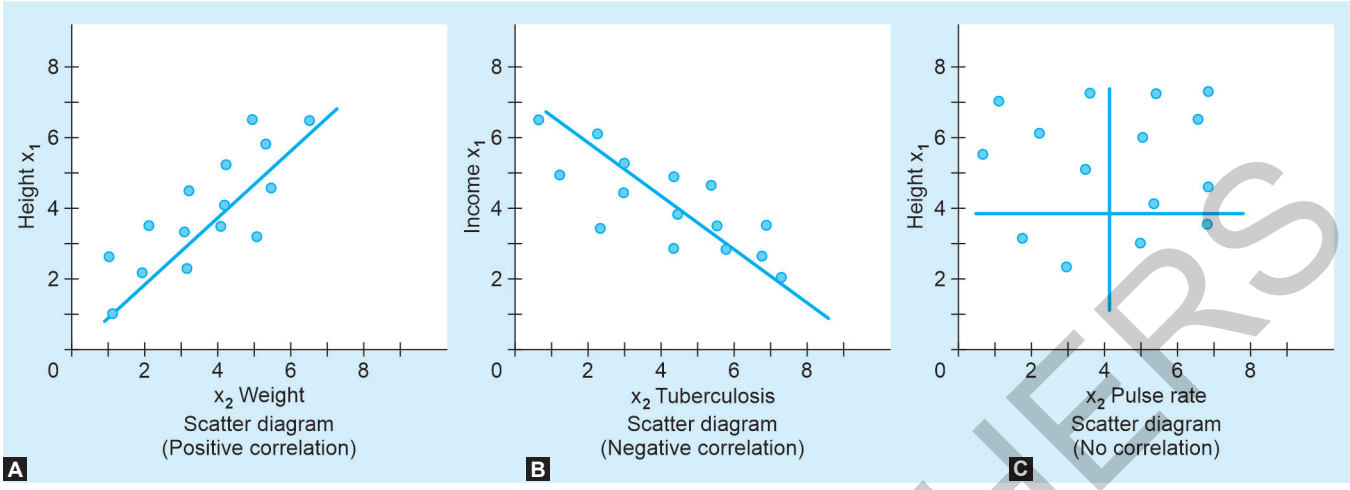
Sectors are then outlined by hatching or coloring or shading. Thus a pie diagram is more attractive and gives a better visual display of its component parts.

To draw a pie diagram, first a circle is drawn. The radius is marked. A second radius clockwise is drawn at an angle with the first radius, depending upon the angle for the sector, which can be calculated by the following formula:

$$\text{Angle for any sector} = \frac{\text{No. of observations (frequency) in a specific group}}{\text{Total number of observations in all the groups}} \times 360°$$

Example: The results of a study of domestic accidents were as follows. Out of total 180 accidents in the house, 60 occurred in the kitchen, 50 in the bedroom, 40 under the staircase, and 30 in the drawing room. The findings are portrayed in the form of a pie diagram **(Fig. 26.11)**.

JAYPEE BROTHERS



Figs. 26.10A to C: Scatter diagram showing different types of correlations.

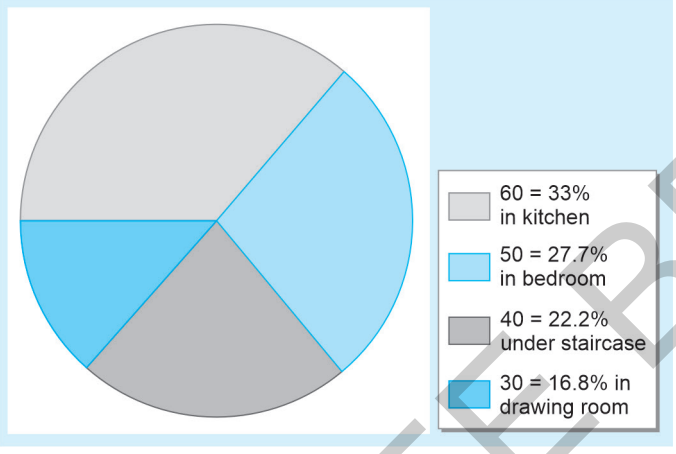


Fig. 26.11: Pie or sector diagram showing distribution of accidents.

Degree

Angle for accidents in kitchen = $\frac{60}{\cancel{180}} \times \overset{2}{\cancel{360}} = 120°$

Angle for accidents in bedroom = $\frac{50}{\cancel{180}} \times \overset{2}{\cancel{360}} = 100°$

Angle for accidents under staircase = $\frac{40}{\cancel{180}} \times \overset{2}{\cancel{360}} = 80°$

Angle for accidents in the drawing room = $\frac{30}{\cancel{180}} \times \overset{2}{\cancel{360}} = 60°$

Percentage of Accidents

(considering 360 = 100%)

$$\frac{\overset{1}{\cancel{60}}}{\underset{3}{\cancel{180}}} \times 100 = \frac{100}{3} = 33.3\%$$

$$\frac{50}{18\cancel{0}} \times 10\cancel{0} = \frac{500}{18} = 27.7\%$$

$$\frac{40}{18\cancel{0}} \times 10\cancel{0} = \frac{400}{18} = 22.2\%$$

$$\frac{30}{18\cancel{0}} \times 10\cancel{0} = \frac{300}{18} = 16.8\%$$

The sectors should be arranged clockwise either in ascending or descending order of magnitude. It is often necessary to indicate the percentage for easy comparison.

This type of representation is effective only when the number of observations is not many and the number of observations in each group can be added to get the total number of observations of all the values of the variable.

The pie diagram can be made more attractive by giving a three dimensional effect to it.

Each sector can be sliced out from the main diagram to highlight the fact.

Pictogram: It is also called picture diagram or pictorial diagram. It consists of a series of small pictures or symbols or silhouettes, each representing a group of data. This method is used to impress a lay person, who cannot understand the orthodox charts **(Fig. 26.12)**.

The pictures are drawn in horizontal lines, each picture indicating a unit of 10, 20, 30, etc. happenings. The number of pictures in each row gives an idea of frequency of the attribute. In essence, pictograms are a form of bar charts.

Map Diagram or Statistical Map: This is commonly used to represent the geographic distribution of disease or deaths of public health importance.

The statistical maps are of two types—(1) spot maps, and (2)shaded maps.

Spot map (geographic spot map): In this type the distribution of the disease frequency is represented in the form of dots or spots, each dot representing a unit number of 10, 20, 30, etc. in the area map prepared. Number of dots will indicate the frequency in units. Such maps show at a glance areas of high

JAYPEE BROTHERS

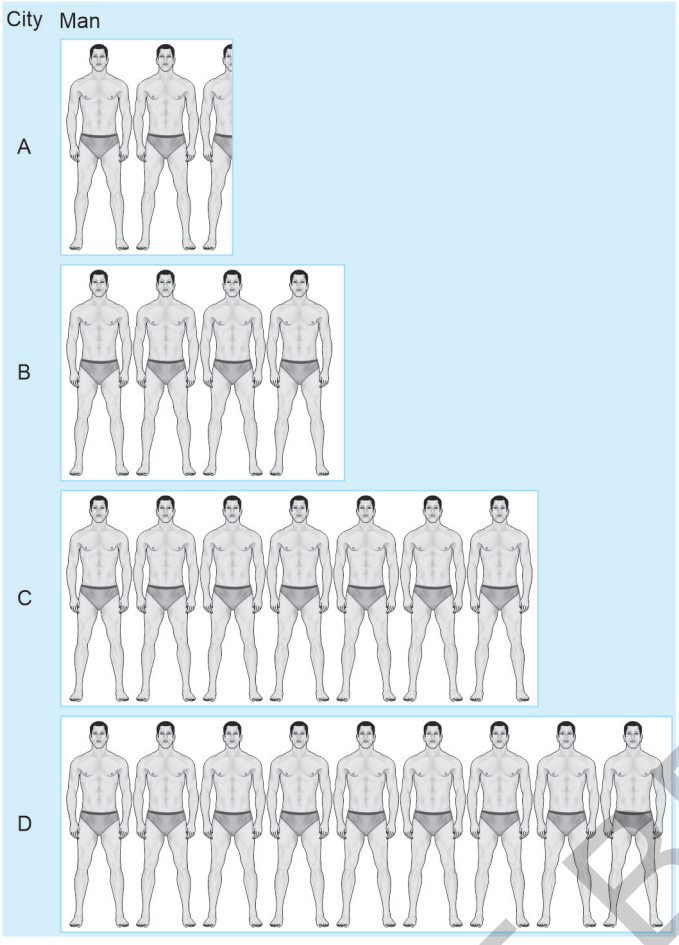


Fig. 26.12: Pictogram of cholera in four cities, each man's figure indicates 10 cases.

frequency (clustering of spots) or low frequency. Clustering of spots may indicate a common source of infection or a common risk factor shared by all the cases. This is how John Snow of England, in 1854, was able to find out that the particular water pump in Broad Street of London was the source of infection of cholera **(Fig. 26.13)**. Thus, he was able to hypothesize that cholera was a waterborne disease, much before the bacilli were isolated by Robert Koch. It was with the spot map, Maxcy (1922–1925) hypothesized rodent as the reservoir of typhus fever. Thus, spot map helps the epidemiologist to study the place distribution, source/reservoir of infection, and behavior of a disease.

Two different colored dots may be marked on the map to show attacks and deaths, in the area.

Maps prepared on weekly or monthly basis help in monitoring changes in the magnitude of epidemics over a period of time and also the direction of their spread.

Shaded map: These are used to indicate the variability in the incidence and prevalence of diseases in different parts of the world/country or from time to time.

Shaded maps indicating state-wise prevalence of diseases are commonly used for formulating appropriate strategies for the control of diseases of national importance. Shaded maps also help in evaluating the progress achieved in reducing the burden of diseases over a period of time.

MEASURES OF CENTRAL TENDENCY

When a series of observations of continuous series are made, it is found that a large number of them concentrate at the center of the series and a small number of them lie at the periphery. This tendency of the values to aggregate in the center of the distribution series is called "central tendency." In other words, it is the measure of the central value of the distribution. The central value is also called "statistical average." Some observations are above the central value, while others lie below this value.

There are three measures of central tendency (statistical averages)—(1) mean, (2) median, and (3) mode.

Mean

Mean is the arithmetic mean unless otherwise specified. It is like the "center of gravity" of a series of data. Other means are geometric mean and harmonic mean. These are uncommon ones.

The arithmetic mean is calculated by summing up all the observations and dividing the total by number of observations. This central value has got the highest utility in statistics. It represents the most appropriate measure of central tendency.

Formula:

$$\text{Mean} = \frac{\text{Sum of all observation made}}{\text{Number of observations made}}$$

Mean is symbolically represented as $\bar{x} = \frac{\Sigma x_1}{n}$

where x_1 is the value of each observation in the data, $\bar{x}$ is the mean value (read as x-bar), Σ means sum of, and n is the number of observations in the data.

Example: The systolic BP in mm Hg of 10 students are as follows:

116, 118, 122, 120, 120, 124, 122, 116, 118, and 124

Mean, $\bar{X} = \frac{1,200}{10} = 120$ mm Hg

The limitations of mean is that it is greatly affected by extreme values.

Median

It is the central value of a series of observations or variables, when arranged in a definite order, either ascending or descending, so that one half of the value lies above it and the other half below it. It is easy to locate when there are odd number of values. When there are even number of values, the median is taken as the average of the two central values of the distribution series.

JAYPEE BROTHERS



Fig. 26.13: Spot map Asiatic cholera in London.
Source: Park K. Park's Textbook of Preventive and Social Medicals, 18th edition, 2005.

The advantage with the median value is that it is easily detected and much less disturbed by extreme values. The disadvantage is that it is not suitable for mathematical treatment.

Example: In a hospital ward, the following are the number of days of stay of patients.

13, 42, 8, 9, 7, 3, 6, 52, 8, 2, 11, 11, 10, 9

For calculation of the median, all the numbers are first arranged in the ascending order, as

2, 3, 6, 7, 8, 8, 9, 9, 10, 11, 11, 13, 42, 52

As there are 14 patients, the average of the periods of stay corresponding to the seventh and eighth patients is calculated as median:

i.e., Median = $\frac{9+9}{2}$ = 9 days

The application of median in pharmacology is to denote the median lethal dose of a drug. That means, it is that dose or concentration of the drug, which is toxic to 50% of the laboratory animals or kills 50% of the animals.

Graphic Method for Locating the Median

The median can also be calculated from the curve "Ogive" obtained by plotting the cumulative frequencies against the end-points of the corresponding class intervals as shown in **Figure 26.9**. Curve is prepared by taking the variable on X-axis and the cumulative frequency on Y-axis.

Median can be located by the following procedure. On the Y-axis, N/2nd frequency is located (e.g., 210/2 = 105). From this point a line parallel to X-axis is drawn to meet the curve. From the point of intersection a line parallel to Y-axis is drawn to meet the X-axis. The point of intersection on X-axis is the median value of the observation.

As seen in the **Figure 26.9**, the median value of the series of observations is 11.55. It corresponds to 50th percentile value (details later).

JAYPEE BROTHERS

Median is a better indicator of an attribute than mean when the lowest and the highest observations are wide apart or not so evenly distributed.

Mode

This is the most frequently occurring value in a series of values or observations. It is least influenced by the extremes of values. It is easy to calculate. But the disadvantage is that it may not exist in a small group of values and it cannot be subjected to mathematical treatment.

Example 1: When studying the age of onset of a disease, it is advisable to know the age at which maximum number of persons are affected rather than the mean age of onset or median age of onset.

Example 2: Following is the ages of 10 medical students.

18, 18, 19, 19, 20, 20, 20, 21, 22, 23

Mode is 20 years of age.

Mode is the only average that can be applied to qualitative data.

In a normal frequency, distribution mode corresponds to mean and median. Mode can also be calculated from the relationship

Mode = Mean – 3 (Mean – Median).

It is also to be noted that in a frequency polygon, the mode can be located from the point, where the curve takes a turn from increase to decrease. There can be more than one mode for a series of data.

MEASURES OF LOCATION: QUANTILES

Quantiles are the values of a variable in an ordered series. They divide the total frequencies in equal parts in order of magnitude. There are five common quantiles. They are median, quartiles, quantiles, deciles, and percentiles **(Fig. 26.14 and Table 26.10)**.

1. **Median:** It divides the total frequencies into two equal parts, each part will have 50% of the total observations on each side of the median (Q_2).
2. **Quartiles:** They are three (Q_1, Q_2, and Q_3), dividing the series in four equal parts, each part having 25% of the total observations.
3. **Quantiles:** They are four, dividing the total frequencies into five equal parts, each part having 20% of the total observations.
4. **Deciles:** They are 9, dividing the total frequencies into 10 equal parts, each part having 10% of the total observations.
5. **Percentile:** They are 99, dividing the total frequencies into 100 equal parts, each part having 1% of the total observations. Thus it reflects the position of an individual, in a grouped series of 100, when arranged in a definite order, either ascending or descending.

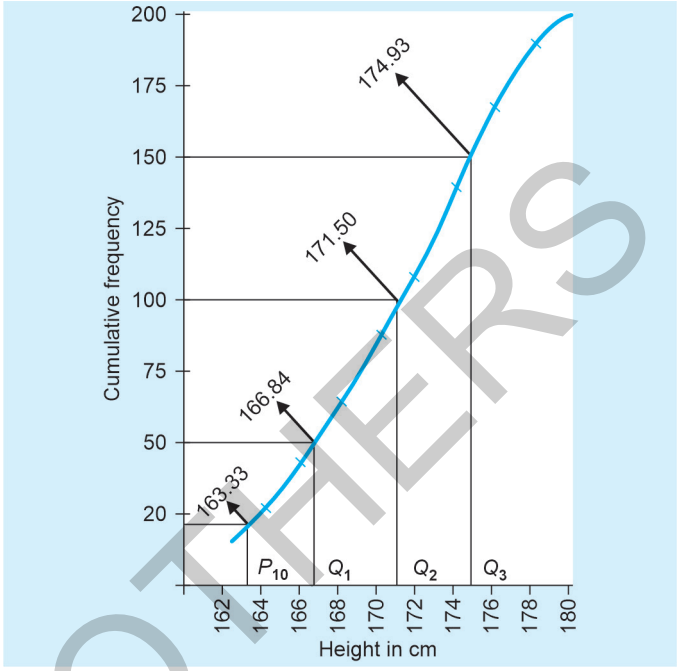


Fig. 26.14: Cumulative frequency diagram showing height values of median (Q_2), first or lower quartile (Q_1), third or upper quartile (Q_3), and 10th percentile (P_{10}).

Source: Mahajan BK. Methods in Biostatistics, 3rd edition, 1981.

TABLE 26.10: Cumulative frequency.

Height of groups (in cm)	*Frequency*	*Cumulative frequency*
160–162	10	10
162–164	15	25
164–166	17	42
166–168	19	61
168–170	20	81
170–172	26	107
172–174	29	136
174–176	30	166
176–178	22	188
178–180	12	200
Total	**200**	

Thus median is the second quartile (Q_2), 5th decile, or 50th percentile, having 50% of the total observations on either side.

First quartile will be Q_1 or P_{25} dividing the total observations into two parts, the first part having 25% and the second part 75% of the observations.

Second quartile will be Q_2 or P_{50} or 5th decile or median, as described above.

Third quartile will be Q_3 or P_{75}, dividing the total observations into two parts, one part having 75% of the observations and the other part 25%.

Accordingly, third quantile divides the series into two parts, the first part having 60% of observations on one side and the remaining 40% on the other side.

JAYPEE BROTHERS

7th decile will have 70% values in the first part and 30% values on the other side, represents 70th percentile. 10th percentile will have 10% in the first part and 90% in the other part and so on.

MEASURES OF DISPERSION

In the previous chapter, it was understood that measurement of central tendency was useful for defining a distribution in a concise manner. But by knowing only the mean, median, and mode, it is not possible to fix the distribution completely. For example, there are two groups of cricket teams, having their diastolic pressures (in mm Hg) as

Team A: 92, 90, 88, 88, 88, 86, 84, 84, 84, 82, 80

Team B: 100, 98, 96, 94, 90, 86, 82, 78, 76, 74, 72

It is seen that both the groups have their mean as 86 mm Hg. At the same, it is also seen that the range as well as the diastolic pressures of the two groups are different. Hence, two sets of data with a common mean need not be same with regard to various individual values of the observation.

So to know how far these observations are scattered from each other or from the mean, certain parameters are employed to measure the dispersion or scatter. These are range, mean deviation (MD), standard deviation (SD), and coefficient of variation (CV).

Range

It is the interval between the lowest and the highest of the values. Thus, range gives the values of the extremes but does not give any information about the values in between the extreme values.

In the above example the range in team A is 12 and in team B is 28. The values are scattered or dispersed around the mean 86 in both the groups.

Range usually defines the limits of normalcy. However, it will be misleading when the extreme values are of unusual occurrence.

Sometimes, instead of knowing the range between all the observations, it may be necessary to know the range within which a certain percentage of observations lie. Such a range is called "percentile range." For example, the range between Q_1 and Q_3 (*see* **Fig. 26.14**) includes middle half of observations in an ordered series. It is also called "interquartile range." This gives little more information about the distribution of the individual observations than range. It will indicate whether the concentration is more at the extremes or in the middle or the observations are evenly distributed.

Mean Deviation

It is the mean or average of the deviations. The deviation (new value) is obtained by deducting the arithmetic mean from each observation. All the deviations are summed up and then the average of all the deviations is called as "MD." It is calculated by the following procedure.

- The "mean" of the observations is calculated.
- Then the mean is subtracted from each of the observations to calculate the deviation.
- The mean (or average) of these deviations is then calculated by totaling the differences from the mean and divide by the number of observations without considering the sign of the deviation, which gives mean deviation.

Mean deviation is given by the formula,

$$MD = \frac{\Sigma |x_i - \bar{x}|}{n}$$

where MD is the mean deviation, Σ is the summation, | | is the vertical bracket (modulus) refers to absolute value, ignoring + or – sign (because if the signs are considered, the total value becomes zero), x_i is the individual value of observation, $\bar{x}$ is the mean of observations, and n is the number of observations.

Though calculation of MD is simple and easy, it is not used in statistical analysis being of less mathematical value, particularly in drawing inferences.

Example: Calculate the MD of systolic BP in mm Hg of 10 students, which is as follows:

115, 117, 121, 120, 118, 122, 123, 116, 118, 120

(Individual observation) x_i	*(Individual observation) - mean ($x_i - \bar{x}$)*		*Deviation*
115	115 – 119	=	–4
117	117 – 119	=	–2
121	121 – 119	=	+2
120	120 – 119	=	+1
118	118 – 119	=	–1
122	122 – 119	=	+3
123	123 – 119	=	+4
116	116 – 119	=	–3
118	118 – 119	=	–1
120	120 – 119	=	+1

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1190}{10} = 119 \qquad \Sigma x_1 \bar{x} = 22$$

$$\therefore \quad MD = \frac{\Sigma |x_i - \bar{x}|}{n} = \frac{22}{10} = 2.2$$

Standard Deviation

Standard deviation is an improvement of the MD. In the calculation of the MD the signs of the deviation (+ or –) of the observations from the mean were not taken into consideration. In order to avoid this discrepancy, instead of the actual values of the deviations, the squares of the deviations are considered for calculation and then the average of the squares is taken, which is known as "variance." As this variance is a square, square root of the variance is considered a measure of variation of the observations. The square root of the variance is known as "SD."

JAYPEE BROTHERS

In other words, SD is the square root of the mean of the squared deviations of the individual observations from the mean. It is also called "root-mean squared deviation."

Of all the measures of dispersion, SD is the most important one and is the best measure of dispersion of observations. It is most widely used in statistical methodologies. It is one of the important parameters of the standard distributions, such as normal distribution and skewed (i.e., asymmetrical) distributions, which are described later.

Further, SD is useful in testing of significance or in measurement of correlation between two sets of data and other statistical methodologies (explained below).

Standard deviation is computed by the following steps:

- Mean of the observations is calculated.
- Mean is subtracted from each of the observation to calculate the deviation (i.e., difference of observation).
- Square the differences of observations (deviations) from the mean, which removes the problem of signs + or –.
- Total (adding) of the squared values to get "sum of squares."
- This sum is divided by the number of observations minus one (n – 1) to get "mean-squared deviation," called "variance."
- Finally, square root of this variance is obtained to get "root-mean squared deviation," called SD (having squared the original, reverse step of taking square root must finally be taken).

Formula:

$$SD = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

where SD is the standard deviation, $\sqrt{}$ is the square root of, Σ is the summation of, x_i is the value of each observation, $\bar{x}$ is the mean of observation, $x_i - \bar{x}$ is the deviation from the mean, n is the number of observations, n – 1 = when "n" is less than 30.

Example: The pulse rate per minute of 10 students in a class are as follows:

80, 90, 96, 80, 94, 72, 84, 92, 82, 90

Mean of the observations = 86

The SD of the pulse rate for the group is calculated as follows:

x_i	$(x_i - \bar{x})$		Deviation	$(x_i - \bar{x})^2$
80	80 – 86	=	–06	36
90	90 – 86	=	+04	16
96	96 – 86	=	+10	100
80	80 – 86	=	–06	36
94	94 – 86	=	+08	64
72	72 – 86	=	–14	196
84	84 – 86	=	–02	04
92	92 – 86	=	+06	36
82	82 – 86	=	–04	16
90	90 – 86	=	+04	16

$$\text{Mean } \bar{x} = \frac{\Sigma x_i}{n} = \frac{860}{10} = 86 \qquad \Sigma(x_i - \bar{x})^2 = 520, n = 10$$

$$\therefore S = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{520}{9}}$$

$$= \sqrt{57.78}$$

$$= 7.60 \text{ per minute}$$

Applications of Standard Deviation

- A standard deviation is the universally accepted unit of dispersion of values from the mean value ($\bar{x}$).
  In a normal distribution series a confidence interval of $\bar{x} \pm 1$ SD encloses 68.27% values, an interval of $\bar{x} \pm 2$ SD encloses 95.45% values, and an interval of $\bar{x} \pm 3$ SD encloses 99.73% values, for purpose of simplicity, a confidence limit of X + 2 SD is treated as including 95% values. In other words, 6 SDs, three on either side of the mean cover almost the entire range of a quantitative series (explained under "normal curve").
- Standard deviation summarizes the variation of a large distribution in one figure and defines the normal limits of variation.
- Standard deviation measures the position or distance of observation from the mean (i.e., to measure individual variability).
- Standard deviation indicates whether the variation of difference of an individual from the mean is by chance (natural or real) due to some special reasons.
- Standard deviation is used to calculate standard error (SE) of mean (SEM) and SE of difference between two means.
- Standard deviation helps in finding the size of the sample.
- Standard deviation is used for calculation of "relative deviate" or "Z score." It is the difference of a specific observation from the mean in terms of the SD. Its formula is

$$Z = \frac{x - \text{mean}}{SD}$$

where Z is the relative deviate, *x* is the observation in question.

Example: The mean height of college students is 150 cm with SD of 10 cm. A particular student's height is 165 cm. Z-score (relative variate) of his height is calculated as follows:

$$Z = \frac{165 - 150}{10} = \frac{15}{10} = 1.5 \text{ cm}$$

SD is used in the calculation of "CV"

$$CV = \frac{SD}{\text{Mean}} \times 100$$

JAYPEE BROTHERS

Coefficient of Variation

The SD of any two quantitative groups (series) in the same group cannot be compared if the attributes are different (such as SD of height and SD of weight) and even if the values pertain to the same attribute (such as height only), comparison cannot be made if the units of measurement are different in the two groups, for example cm and inch. This limitation of SD is satisfactorily removed by converting SD into "CV."

The CV is the SD expressed as the "percentage of the mean." Thus while SD is expressed in different units of measurement, CV is a unitless number. Therefore, CV is well suited for all types of dissimilar measurements, such as height and weight or Hb and weight or pulse rate and midarm circumference.

Formula:

$$CV = \frac{SD \times 100}{Mean}$$

As an example, the variation of Hb level of a group of persons and the variation of their body weight, which are expressed in different units, such as gram percent and kilogram, respectively, will have different means and different SD. CV helps to compare the deviations in two sets of dissimilar measurements in the same group or one character in two groups.

Example 1: Example for two characters in the same group. The mean and SD of Hb level of a group is 12.6 g% and 1.5 g%, respectively while the mean and SD of the body weight of the same group is 50 and 2.2 kg, respectively.

To compare the deviations of these two sets of observations CV is calculated for each data.

$$\text{CV of Hb level} = \frac{1.5 \times 100}{12.6} = 11.9\%$$

$$\text{CV of body weight} = \frac{22 \times 100}{50} = 4.4\%$$

It is seen that the variation is greater for Hb level than for body weight of the same group.

Example 2: Example for one character in two groups. In two series of boys and girls of the same age of 20 years following values were obtained for the height. Find which sex shows greater variation.

Sex	*Mean height (cm)*	*SD (cm)*
Boys	163.25	150.35
Girls	6.25	5.25

$$\text{CV of boys} = \frac{6.25}{163.25} \times 100 = 3.83\%$$

$$\text{CV of girls} = \frac{5.25}{150.35} \times 100 = 3.49\%$$

Thus, it is seen that the height in boys shows slightly greater variation than in girls being in the ratio of 3.83:3.49 = 1.1:1.0.

PROBABILITY

Probability means a chance factor for the occurrence of a specific event, e.g., chances of winning a lottery, chances of being selected, chances of getting a male child in the first pregnancy, etc. This chance factor is associated with uncertainty, because information in the happenings is not available. This uncertainty or mathematical quantity, which depends upon the occurrence of the favorable or unfavorable event, is numerically expressed as "probability."

The probability of a particular event can be defined as the ratio of the number of favorable cases for the particular event to the total number of cases both favorable and unfavorable to the particular event.

For example, if the probability of a patient leaving the hospital against medical advice during a year's time is to be computed, then this will be the ratio of the number of patients (n) who left the hospital against medical advice during a year's time to the total number of patients (N) admitted to the hospital during that year.

$$\text{Formula} = p = \frac{n}{N} = \frac{\text{No. of favorable cases}}{\text{Total number of both favorable and unfavorable cases}}$$

Example: Suppose there are 100 students in a class, 90 boys and 10 girls, then the probability of a student chosen at random to be a girl is

$$= \frac{\text{No. of girls in the class (n)}}{\text{Total number of students (both boys and girls)N}} = \frac{10}{90+10}$$

$$= \frac{10}{100} = 0.1$$

The probability scale varies from the lowest value "0" (zero) to the highest value "1" (one). In other words, if none of the cases are favorable, i.e., when the numerator is zero, then the probability of the event will be zero. On the contrary, if all the cases are favorable for the event, then both the numerator (n) and the denominator (N) will be equal, then the probability of that event will be 1. As in the above example, if there are no girl students and all are boys, then the probability of a student to be a girl is zero and the probability of a student to be a boy is one. Thus, zero indicates no occurrence and one implies 100% certainty.

When there are only two complimentary events of occurrence, then the probability of one event will be equal to (1 - probability of the other event).

As seen in the above example, the probability of selecting a boy at random.

$$= \frac{90}{100} = 0.9$$

$$= (1 - \text{probability of selecting a girl})$$

$$= (1 - 0.1)$$

JAYPEE BROTHERS

The probability of an event becomes more and more consistent when the number of observations is increased.

Probability Distributions

The common distributions which help in calculation of probability are normal, binomial, and Poisson distributions.

Normal Distribution and Normal Curve

Normal distribution is a frequency distribution, in which a large number of observations of any variable biological characteristic, such as Hb, height, weight, pulse, and BP, etc. are made with a small class interval. Such a distribution will have the following characteristics:

- Half the measurements lie above the mean and half below
- Most of the measurements (observations) are concentrated around the mean
- All the observations are symmetrically distributed on each side of the mean, i.e., if they are arranged in an order, the highest frequencies will be seen in the middle around the mean and the lowest at the extremes, decreasing smoothly on both sides. This is called "central tendency" or concentration of observations around the central value or mean.

Such distribution is called "normal distribution" or "Gaussian distribution" and when plotted graphically it is called normal curve or Gaussian curve, named after a famous mathematician and astronomer Karl Gauss **(Fig. 26.15)**. The term "normal" is not used in medical sense. By "normal," it is meant usual and not "not abnormal."

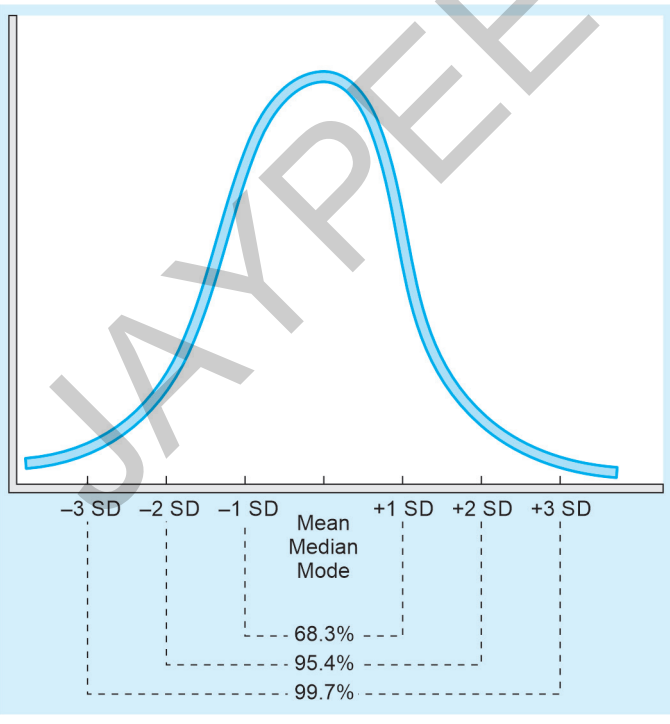


Fig. 26.15: Normal distribution curve.

The features of a normal curve are as follows:

- It is a smooth curve, resembling the shape of a bell.
- It is symmetrical about the middle point.
- The rim of the bell does not rest on the abscissa but is separated from it by a gap. That means the curve stretches from infinity to infinity.
- All the three measures of central tendency, i.e., mean, median, and mode coincide, i.e., a perpendicular drawn from the peak of the curve to the abscissa and that point on the abscissa is the mean, median, and the mode. The mean divides the area under the curve into two equal halves.
- Maximum number of observations are at the value of the variable corresponding to the mean and the number of observations on both sides of this value gradually decrease and there are few observations at the extreme points.
- The area under the curve is proportional to the frequency (i.e., 1).
- **The empirical rule is often referred to as 68–95–99.7 rule, i.e.**, 68% of the data falls within the 1st standard deviation (SD) from the mean. That means there is 68.26% probability of all observations (data) lie within the range of -1SD to +1SD (i.e., 34.13% lies on either side of the mean).
  95.45% of the data fall within the range of –2SD to +2SD from the mean and 99.73% of the data fall within the range of –3SD to +3SD from the mean.
  This relationship of the data values falling within the range of ± 2SD and ±3SD is useful in fixing the confidence interval of the variates. The confidence interval is the limit up to ±3SD on either side of the mean (i.e., the limits up to ±3SD on either side of the mean are called "confidence limits"). The distance of the observed value X from the mean $(\overline{X})$ of the curve in terms of SD is called "relative variate" or "standard normal variate". It is expressed as 'Z-score'.
- **Z-score:** A Z-score, also called as "standard score", is a statistical measure that describes, how many standard deviations an observed value is above or below the population mean value. Z-score is measured in terms of standard deviation away from the population mean.
  Z-score is calculated by subtracting the mean from the observed value and divided by standard deviation.
  Formula:

$$Z = \frac{X - \overline{X}}{SD}$$

  Where Z = Z-score
  X = observed value
  $(\overline{X})$ = mean of the sample population
  SD = Standard deviation of the sample population.
  Interpretation: Z-score can be zero, positive or negative.
  If Z = 0 → The observed value is identical to the mean
  Z = > 0 → The value is above (higher than) the mean
  Z = < 0 → The value is below (lower than) the mean
  Z = + 1 → The observed value is 1 standard deviation above the mean
  Z = +2 → The observed value is 2 standard deviation above the mean **(Fig. 26.16)**

JAYPEE BROTHERS

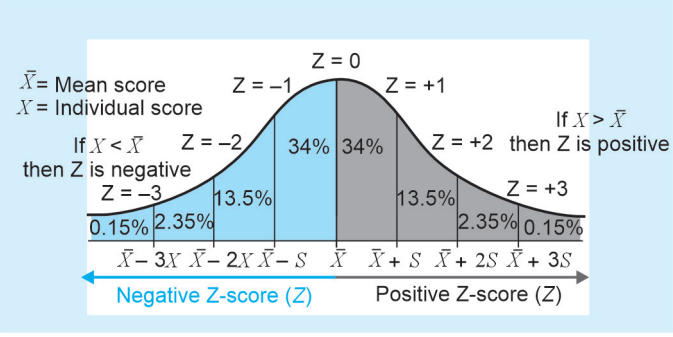


Fig. 26.16: Z-score.

Note: To calculate Z score, one should first determine the mean and the standard deviation of the data.

Example:

Suppose in an internal assessment test in a class, mean score ($\bar{X}$) = 50, standard deviation (SD) = 10, and a student's score (observed value X) is 70, calculate Z-score.

$$Z = \frac{X - \bar{X}}{SD}$$

$$= \frac{70 - 50}{10} = \frac{20}{10} = 2$$

That means, the student's score is +2 standard deviations more than the mean. If the ratio value is more than two, it is significant and if more than three, it is highly significant.

Uses:

- A Z-score helps us to calculate the probability of a score occurring within a standard normal distribution. (It is based on the assumption that sample values are normally distributed).
- It also enables us to compare two scores that are from different normal distribution.
- In case of small samples, instead of Z, '*t*' ratio is used.
- A Z-score in pediatrics is used to find how a certain birth-weight of a newborn deviates from the mean weight of all babies.

It is also used to assess the nutritional status of child (in growth monitoring), by comparing the weight for age, height for age, or weight for height to WHO reference means. Children with measurements weight for age below –2 Z-scores are considered under nourished, while those below –3 Z-scores are classified as severely malnourished. Above +2 indicates over-weight and +3 indicates obesity (refer growth monitoring).

Height for age Z-score. If Z-score = <-2 the child is stunted (chronic malnutrition). Weight for height Z-score. If Z-score = –2, the child is having wasting.

- The properties of a normal distribution and a normal curve form the basis of various tests of significance.
- Values larger and smaller than mean ± 3 SD will be rare (less than 1%) in nature and those larger and smaller than mean ± 2 SD will occur less than 5%. In other words, suppose we say that the confidence limit is 99%, that means 99% of the values are distributed within the range of x ± 3 SD and the probability of occurrence of any value falling outside this range is only 1% (p = 0.01). Similarly, suppose we say that the confidence limit is 95%, that means 95% of the values are distributed within the range of x ± 2 SD and the probability of occurrence of any value falling outside or beyond this range is only 5% (p = 0.05).

The practical application of the above procedure is illustrated in the following example:

The mean height of 500 students is 160 cm and the SD is 5 cm.

a. What are the chances of heights above 175 cm being normal if height follows normal distribution?
b. What percentage of boys will have height above 168 cm?
c. What number of boys will have height between 168 cm and 175 cm?

i. First, the SNV is calculated for the given value of the variable.

$$\text{i.e. } Z = \frac{x - \bar{x}}{SD} = \frac{175 - 160}{5} = \frac{15}{5} = 3$$

To this calculated value 3, the probability level is noted by referring to the table of "Unit Normal Distribution" (Annexure I) of this chapter. This corresponds to 0.0013. When this probability value is multiplied by 100 gives the percentage population beyond the value of the variable. Thus only 0.0013 × 100 = 0.13% of students have chances of being taller than 175 cm.

ii. $z = \frac{168 - 160}{5} = \frac{8}{5} = 1.6$

This corresponds to 0.0548 as per the table of "Unit Normal Distribution." Thus the percentage of boys having height above 168 cm will be = 0.0548 × 100 = 5.48%.

iii. The number boys having height above 168 and below 175 cm will be 0.0548 – 0.0013 = 0.0535 out of 1. Thus only 5.35% of students will have height within the range of 168–175 cm.

Binomial Distribution

When the population under observation can be divided into two distinct groups, one with a certain characteristic of observation and the rest without this characteristic, the distribution of the occurrence of the characteristic in the population is called as "binomial distribution."

For example, suppose it is required to define the distribution of prevalence of certain disease, then the entire population can be divided into two groups, one with the disease and the other without the disease. If the probability of a person having the disease is "p" and not having the disease is "q" (i.e., 1 – p), then the distribution of various probabilities, the probability of finding name, one, two, three, four ... Persons with the

JAYPEE BROTHERS

disease in a group of "n" persons is given by the successive terms of the binomial distribution $(q + p)^n$. (If the sample is divided into multiple classes, such as blood groups A, B, AB, and O, that sample is called "multinomial".)

Mean features of the binomial distribution

- Mean number of frequencies with the characteristic is given by "np," where "n" is the number of observations and "p" is the probability of occurrence of the characteristics
- Standard deviation is npq, where q = 1 – p
- If "p" and "q" are equal, i.e., each equal to 0.5, the curve will be symmetrical
- If "p" and "q" are not equal, then the curve will be skewed (asymmetrical)
- For any fixed value of "p" and "q," if "n" is increased to a sufficiently large quantity, binomial distribution becomes more symmetrical and tends to be normal distribution.

Poisson Distribution

Poisson distribution is the probability of the distribution of an event, occurring in a given fixed period of time or space, when the frequency of occurrence of that event is known. For example, if the average number of diners for 7 days, in a hotel, is 500, we can predict the probability of a certain day having more customers. This is used to hire extra workers temporarily or to make more purchases to accommodate the guests coming. Similarly, Poisson distribution is used by businessmen to make forecasts about the sales or number of customers on certain days or seasons of the year. Other examples are number of meteorites striking earth in a year, number of phone calls received by a call center per hour, number of patients arriving in an emergency ward between 1 AM and 2 AM.

Poisson distribution is a discrete distribution. Each occurrence is independent of other occurrences. Thus Poisson distribution is a tool that helps predict the probability of certain event happening in a fixed interval when we know how often the event has occurred. Poisson distribution is named after French Mathematician, Simeon Denis Poisson (1838).

Assumption: The Poisson distribution is an appropriate model if the following assumptions are true.

- *k* is the number of times an event occurs in an interval (a random variable)
- Events occurring independently
- The rate at which the events occur is constant
- Two events cannot occur at exactly the same instant.

If these conditions are true, the "k" is a Poisson random variable and the distribution of "k" is a Poisson distribution.

The probability of observing k events in an interval is given by the equation (formula).

$$P = \frac{(e^{-\lambda})(\lambda^k)}{k^!} \quad \text{or} \quad P(x:\mu) = \frac{(e^{-\mu})(\mu^x)}{x^!}$$

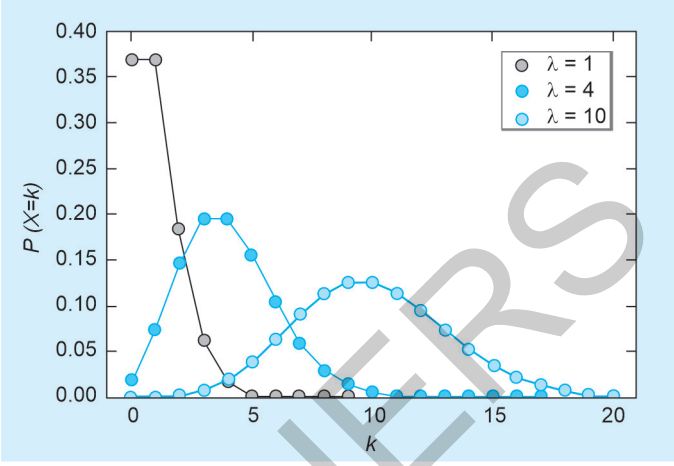


Fig. 26.17: Poisson distribution, probability mass function.

where P is the probability of occurrence of the event; e is the constant number 2.71828 (Euler's number); λ (or μ) is the average number of events per interval (event rate or rate parameter); k (or x) is the number of events 0,1,2,...; ! = this number is a factorial; k! = k × (k – 1) × (k – 2) × ··· × 2 × 1.

This equation is the probability mass function (PMF) for a Poisson distribution. The PMF is shown in **Figure 26.17**.

The horizontal axis is the index *k*, the number of occurrences. λ or μ is the expected number of occurrences. The vertical axis is the probability of *k* occurrences given. The function is defined only at integer values of *k*. The connecting lines are only guides for the eye.

Example: The average number storms in a city is 2 per year. What is the probability that exactly three storms will hit the city in the next year?

Data. $\mu = 2$ (Av. No. of storms per year)
$x = 3$ (The No. of storms may hit next year)
$e = 2.71828$ (Euler's mathematical number, constant).

Formula:

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

$$= \frac{(2.71828^{-2})(2^3)}{3!} \qquad = \frac{(0.13534)(8)}{6}$$

$$= 0.180 \text{ or } 18\%$$

The usual way to calculate Poisson distribution in real life is with the software, such as International Business Machines (IBM), Statistical Package for the Social Sciences (SPSS).

Note: If the equation has the exact probability and we have to find the probability of happening in a certain number of times out of *x* (i.e., 10 times out of 100 or 99 times out of 1,000), we have to use binomial distribution. Both Poisson distribution and binomial distribution measure the number of events within a certain frame. The difference is that the Poisson distribution is used when an average number of events are given and binomial distribution is used when the number of events is known exactly.

JAYPEE BROTHERS

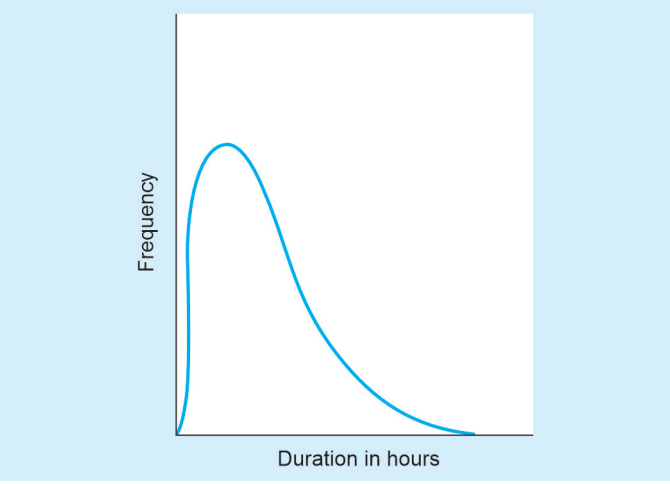


Fig. 26.18: Asymmetrical distribution of duration of labor (first child) showing skewing to the right.

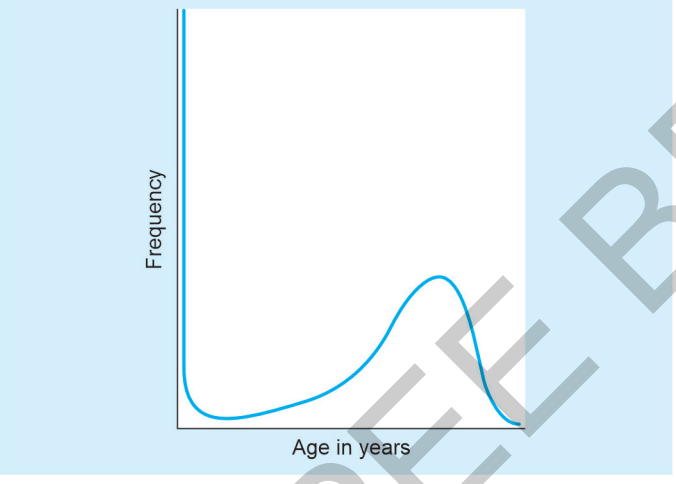


Fig. 26.19: Asymmetrical distribution of deaths according to age showing two modes or peaks in infancy and old age.

Skewed Distribution

When the frequency distribution or a frequency curve is not symmetrical about the peak, it is said to be "skewed" (asymmetrical). In other words, one tail of the curve will be longer than the other. This skewness can be either to the right or to the left of the peak **(Figs. 26.18 and 26.19)**.

A relative measure of the skewness as given by Karl Pearson is

$$\text{Skewness} = \frac{3(\bar{x} - \text{med})}{s}$$

Where $\bar{x}$ = mean

Med = median

and s = SD of the distribution

When this calculated quantity is positive, it indicates that the skewness is to the right and when negative, it indicates skewness to the left.

SAMPLING

The common terms used are population, sample. The word "population" or "universe" means an aggregate of all "elementary units," each unit may be animate or inanimate, about which an information is required.

Universe or whole population may be "finite" (e.g., 100 kg of rice in a sack or all the inhabitants of a city) or it may be "infinite" (e.g., stars in the sky, etc.). Universe may be "homogeneous" i.e., made up of uniform class (e.g., polished rice in a sack; all Hindu women of reproductive age in the city, etc.) or it may be "heterogeneous," made of dissimilar classes of persons or animals or objects.

It is not possible for any scientific study to cover the whole population because of the cost, time, and practicability. So a representative portion of the universe is taken for the study. It is called a "Sample," e.g., a handful of rice from a sack of 100 kg rice, a group of people attending a public meeting, etc. The observations made out of a representative sample are then applied to the universe at large, but generalization is valid, only if the sample is sufficiently large and representative.

Sampling is a procedure, adopted for selection of a representative unit of the universe units. The sampling procedure can be adopted with the help of "sampling frame," which consists of a list of "enumeration units."

When the investigation is carried out for the entire population it is called the "census enumeration" and when it is carried out for the sample, it is called "Sample enumeration."

Suppose a sample is selected by selecting only those units of the population, so that it suits a specific purpose as per the desire of the investigator, it is called "purposive sampling." This method serves only a limited purpose. Suppose a sample is selected in such a way that the characteristic of the population is represented, it is called "representative sampling."

Characteristics of a Representative Sample

- It is selected by a sampling technique from the universe or population it represents.
- Usually, it does not differ from the universe in composition. If it differs, it is solely by chance.
- Each member of the universe from which it is taken has an equal opportunity of being selected.
- Bias has been ruled out and the sample will give an estimate of the attribute under study, almost equal to the population value called "true value."
- It is chosen according to a rule that is independent of the observations to be made on the sample.
- Sample should be sufficiently large in size to represent population from which it is drawn.

JAYPEE BROTHERS

Census Enumeration versus Sampling Enumeration

Census enumeration	Sampling enumeration
Investigation is carried out for the entire population (all units)	Investigation is carried out only for a sample
Useful when detailed information is required	Useful when overall information is required
Organization is costly and consumes more time	Organization is cheap and consumes less time
Greater attention cannot be paid for each unit because of the vastness of the population	Greater attention can be paid for each unit because of relative smallness of the sample
Difficult to achieve a complete and accurate study	Easy to achieve a complete and accurate study
Requires more personnel for study	Requires less personnel for study

Sampling Designs (Sampling Methods and Sampling Techniques)

Types of Sampling Techniques

The methods (techniques) of sampling may be classified into two categories, namely: (1) probability, and (2) nonprobability sampling methods, based on the concept of random selection or nonrandom selection (deliberate selection).

1. Probability sampling methods (random sampling methods)
   i. Simple random methods (unrestricted random sampling)
   ii. Systematic random sampling
   iii. Stratified random sampling
   iv. Multistage random sampling
   v. Cluster random sampling
   vi. Multiphase random sampling
2. Nonprobability sampling methods
   i. Accidental or incidental sampling
   ii. Judgment sampling or purposive sampling or deliberate sampling
   iii. Quota sampling
   iv. Convenience sampling
   v. Sequential sampling

Probability Sampling

Simple random sampling: The principle in this method is that every unit of population has an equal chance of being selected. Random sampling can be easily done by using "Tippets random number table" (Annexure II) of this chapter. This method is also called "unrestricted random sampling."

This method is fairly applicable when the population is small, homogeneous, and readily available, such as a village, a household, or an individual in a community or patients coming to hospital or lying in the wards.

Example: To select a random sampling of 25 students (n) from a class of 75 students (N).

Procedure

- Unit of selection is decided, i.e., a student. All the 75 students are arranged in an order say in the alphabetical order of their names and they are numbered from 1 to 75. This is known as the "sampling frame."
- Twenty units from the serial numbers are obtained as detailed below, by referring Tippets random number table, which has 50 columns of single digit numbers, in each table. Out of these, a row and a column are selected at random. Say the eighth row and sixth column core are selected, the number corresponding to the crossing of this row and column is 6.
- The total population size is two digits, i.e., 75; and hence, two adjacent columns are to be clubbed together, i.e., sixth and seventh columns are clubbed together and is read as two digit numbers, which becomes the starting number. In the table, it corresponds to 61 (if total units are three digits, three columns are clubbed together).
- Starting from this number 61, the numbers in the list are read downward. All the numbers less than or equal to 75 are noted down.
- When the last row of these columns is reached, then the corresponding number of next two columns, i.e., 8 and 9 are taken and the procedure is continued till 25 unduplicated numbers (the desired sample) are obtained (all the duplicated numbers are deleted).
- The numbers so read are

  61, 62, 2, 31, 51, 11, 56, 64, 21, 1, 16, 39

  6, 38, 26, 34, 8, 65, 22, 52, 7, 29, 30, 14, and 18.

  These units constitute the sample.

Another way of drawing the sample is by noting the serial numbers of students on 75 cards and shuffled them well. One card is drawn out and the number is noted. The card is replaced, reshuffled, and second card is drawn. The process is repeated till 25 numbers are drawn. The card drawn for the second time is rejected. The 25 students bearing the card number constitute the sample. Similar procedure is followed to select the control group in case–control study.

If the table of random numbers is not available, currency notes can be used. The numbers printed on them are considered. Depending upon the number of digits of the universe or population (two digits in the above example 75), first two or middle two or last two digit numbers are recorded. The subjects (units) having the corresponding numbers are included in the sample.

Grid system: This system is employed for selecting samples of an area. For that, first an area map is prepared. Then a screen with squares is placed upon the map and the areas falling within the selected squares are taken up as samples.

Systematic sampling (Systematic random sampling): This method is preferred when the population (universe) is large,

JAYPEE BROTHERS

scattered, and not homogeneous, such as number of houses in a village or town. No prenumbering is necessary as in random sampling.

As in the above example of drawing 25 (n) out of 75 (N), "N" is divided by "n" to get a quotient "r" (75 ÷ 25 = 3). Then one unit of N is selected at random and the other units are subsequently selected by the addition of this quotient "r" to the previous selected number. Since "r" being 3, any one number (unit) is drawn between 1 and 3, say 2, then the sample will be made up of students with numbers 2, 2 + 3 (5), 5 + 3 (8), 8 + 3 (11), and so on. This quotient "r" is known as "sampling interval."

Suppose there are 210 villages in a community development block and 40 villages are desired to be selected, then N/n = 210/40 = quotient is 5, remainder is 10. Then a random number is selected out of 10, say 6, which becomes the first number. Then the quotient 5 is added to 6 and so 11 will be the second number. Again 5 is added to 11, so 16 becomes the third number and so on.

Hence, the serial numbers of the villages to be selected will be 6, 11, 16, 21, 26, 31, ... so on up to 40 numbers.

Stratified random sampling: In this method, the entire population (universe) is divided into certain homogeneous subgroups (or strata) depending upon the characteristics to be studied (the basis for status stratification being age-group, sex group, area wise, socioeconomic status, etc.) and simple random sampling is drawn independently from each subgroup or strata. This technique gives more representative sampling than a simple random sampling in a given large population.

This type of sampling is used when the population is heterogeneous with regard to the characteristic under study or when the characteristic is influenced by the different sections of the universe. As an example, if it is known that the prevalence of a particular disease is different in different age-group, the population (universe) is stratified into different subgroups as children, adults, and old persons and samples are drawn from each strata; the number of people in each age-group may not be equal in the population.

By this type of sampling the precision of the estimate of the characteristic under study is increased and also due representation of the population is maintained. Another advantage is that the estimate of the characteristics under study can be made for each strata separately. It also ensures that all strata are adequately represented.

Multistage sampling: As the name implies this method consists of sampling procedure carried out in several stages, using random sampling techniques.

This is convenient when the population of entire district (or state or country) is to be studied, within limited resources. To bring down the cost involvement the size of the sample is reduced progressively in stages, till a conveniently representative sample is obtained.

First random numbers of districts are chosen from the states. Then random numbers of *talukas* are chosen, followed successively by villages and houses.

Example for hookworm survey in a district, 10% of *talukas* are chosen, followed by 10% of villages. Then all persons in 10th house are subjected for stool examination.

Advantages

- This method introduces flexibility in sampling, a feature lacking in other techniques.
- It enables the use of existing divisions and subdivisions and thus saves the extra labor involved in independent enumeration or census.
- It permits available resources to be concentrated on a limited number of units of the frame, which results in a lower cost per unit of the enquiry and in preparing a complete sampling frame.

Disadvantages

- Sampling error is usually increased,
- Sampling units will be of unequal size at various stages, resulting in analytical difficulties.

Cluster sampling

in this case the enumeration (sampling) units are not individuals but clusters, such as families in a village, villages in a district, schools and wards of a city, etc. A sample of clusters proportionate to their size is randomly drawn. Either everyone in the sample is studied or only a certain number of subjects with specified age or age-group are examined.

Cluster sampling is employed for carrying out evaluation survey of immunization coverage. In addition, it is used when the list of sampling units is not available.

Point or line sampling: On the map or aerial photograph of the area to be studied random horizontal and vertical lines are drawn. The points where these lines intersect are included in the sample. People at these points are examined.

Biased sampling: Following are the convenient methods but not the correct method of sampling:

- **Alphabetic sampling:** This consists of a sample of persons whose name begins with a particular alphabet.
- **Chunk (opportunistic) sampling:** This consists of a group of persons attending a cinema show or public hall or found at the market place, etc.
- **Volunteers:** It consists of those volunteers who have come for examination following an advertisement in a paper.

Multiphase sampling: In this method, study is carried out in several phases. Suppose a cross-sectional study on nutrition has to be carried out, all the families in the original sample is covered for KAP study (Knowledge, Attitude and Practice) in the first phase. A subsample of the families is then surveyed for dietary intake in the second phase. Then a subsample of the family members covered in the second phase are subjected to anthropometric examination in the third phase. A further subsample of the members covered in

JAYPEE BROTHERS

the third phase are subjected to biochemical examinations the nutrients in the fourth phase. Thus in multiphasic sampling the number of subjects/units gets reduced in every succeeding phase, thereby reducing the magnitude of the complicated and costly procedure reserved for the last phase. Thus multiphasic sampling procedure makes the studies less expensive, less time consuming, less laborious, and more purposeful.

Nonprobability sampling

Accidental or incidental sampling: This consists of selecting a group of people among those who have assembled in one place with a common interest. For example, suppose diabetic survey has to be done among the people above 40 years of age by doing urine test, we have to go to a place, where people having common interest have assembled, to see football or cricket match. This is called "chunk or opportunistic" sampling.

Judgment sampling (purposive or deliberate sampling)
In this method the sample selection depends upon the judgment of the person, who is entrusted with the job, which in turn depends upon his knowledge and attitude.

Quota sampling
This is a stratified random sampling minus randomization. For example, suppose, in a town, there are 50% farmers, 25% small businessmen, and 25% workers and a sample of this has to be drawn, that sample should have the same percentages of these groups. But randomization is not done.

Convenient sampling
In this method the sample is obtained from an available source, such as that of telephone directory, automobile registers, stock exchange directory, etc. conveniently.

Sequential sampling
In this method, number of sample lots are drawn one after another from a universe depending upon the results of earlier samples. So this is used as quality control.

Size of the Sample

Always an optimum size of the sample has to be considered, keeping in view of the time, cost, and the feasibility of the study.

The estimation of the sample size involves the following factors:

- The approximate idea of the estimate of the characteristic under observation is required and that is obtained either from previous studies or from pilot study.
- The maximum permissible error that can be allowed should be decided in advance. If the error is large, then a small sample will serve the purpose and vice versa.
- The probability level, with which the desired precision of the estimate is maintained, is also taken into consideration for fixing the sample size. Higher this probability, bigger should be the sample size.
- Availability of the resources, such as men, money, and material also determine the size of the sample.

Calculation of Sample Size

For Qualitative Data

In a field survey to estimate the prevalence of a particular disease the sample size is calculated by the formula,

$$n = \frac{4pq}{L^2}$$

where n is the required sample size, p is the approximate prevalence rate of the disease obtained from previous studies or from pilot study, q = 1 – p, L is the permissible error in the estimate of "p."

The above formula has been worked out for a probability level of p = 0.05 (i.e., the prevalence rate will have 5% error or 95% correct value) in the sample size.

Example: To estimate the prevalence rate of ascariasis in a community, where it is approximately known to be 40 %, then the required sample size to estimate the morbidity (ascariasis) with 5% error with a probability of 0.05, is calculated as follows:

$$n = \frac{4\,pq}{L^2}$$

where

$$p = 40\%$$

$$q = 1 - p = 100 - 40 = 60\%$$

$$L = 5\% \text{ of } 40 = \frac{5}{100} \times 40 = 2$$

$$= \frac{4 \times 40 \times 60}{(2)^2} = 2400$$

Ans: 2400 persons are to be examined to estimate the prevalence rate of ascariasis with 5% error.

Suppose the error of percentages is increased to 10%, then the sample size would be

$$n = \frac{4 \times 40 \times 60}{4 \times 4} = 600$$

$$L = \frac{10}{100} \times 40 = 4$$

For Quantitative Data

The sample size is calculated by the formula

$$n = \frac{t^2\alpha \times s^2}{e^2}$$

where n is the desired sample size, s is the standard deviation of observation, *e* is the permissible error in the estimation of mean, and tα is the value of "t" at 5% level from "t" tables.

JAYPEE BROTHERS

Example: (a) In a community survey to estimate the Hb level, from the data already available, if it is known that the mean Hb% level is about 12 g% with a SD of 1.5 g% then the sample size required to estimate the Hb level with a permissible error of 0.5 g% on either side is obtained as follows:

s = 1.5 g

e = 0.5 g

$t_{0.05}$ can be taken as 2 (for 1.96; Annexure III of this chapter) as it is conventional to use 5% level of significance.

$$n = \frac{2^2 \times (1.5)^2}{(0.5)} = \frac{4 \times 2.25}{0.25} = 36 \text{ persons}$$

In clinical trials, usually there will be two groups, one experimental and the other control group. In order to estimate the size of the sample for each group, the difference in the response rates of the two groups is to be taken into consideration and the sample size is estimated from the following formula.

$$n = \frac{2t^2\alpha \times s^2}{d^2}$$

where n is the required sample size for each group, *s* is the pooled SD of the observation of the two groups, d is the anticipated smallest difference in the estimates for the two groups, and tα is the usually taken as "t" at 5% level.

Example: An investigator wants to estimate the increase in Hb% level in anemia cases by administration of a particular drug compared against a known drug. The minimum number of cases in each group to be investigated is calculated as follows:

Suppose d = the smallest anticipated difference in the rise of Hb% level between the two groups = 2%

s = the pooled SD = 3.0 g

t at 5% level is taken as 2.

$$n = \frac{2 \times (2)^2 \times (3.0)^2}{(2)^2} = \frac{8 \times 9}{4} = 18 \text{ persons}$$

i.e., 18 persons are to be included in each group.

Errors in Sampling

They are of two types— (1) sampling errors, and (2) nonsampling errors.

Sampling Errors

These are due to:

- Faulty sampling method
- Small size of the sample.

These errors can be minimized through proper sampling method.

Nonsampling Errors

These are as follows:

- **Coverage error:** This occurs when all the units in the sample are not covered either due to noncooperation or due to lost to follow-up. This can be reduced by an intensive effort to get complete coverage.
- **Observational (or experimental) error:** This is due to interviewer bias or due to lack of training. This can be reduced by setting up standards of interview or proper training of the workers.
- **Processing errors:** This is due to clerical mistake or computational error, which can be reduced by administrative control.

Sampling Variation

If two or more samples are drawn from the same universe (population), their means (m_1, m_2, m_3, ...) may not be equal but will show variation, even though they are from the same universe. Such a difference between the means of the different samples is known as "sampling variation."

The mean of the sample means (i.e., grand mean) however will be approximately equal to the mean (M) of the universe.

The variation between the sample mean and the universal mean (m_1 and M; m_2 and M; and m_3 and M) is known as "inherent sampling variation."

The sampling variation, from one sample mean to another, may be by chance, when it is called "natural or biological variability," or due to play of certain factors, when it is called "real variability," (e.g., effect of nutrition, vaccine, smoking, etc.; heights and weights in England are more than that in India because of good nutrition; attacks among vaccinated are lesser than that among unvaccinated; incidence of cancer is more among smokers than among nonsmokers, etc.).

Standard Error of the Mean

The means of the samples (m_1, m_2, m_3, etc.) show dispersion around the population mean (M) (or universe mean) symmetrically as in normal distribution with a central tendency and with a definite SD. The variation in the sample means is measured in terms of a parameter called "SE," which in fact is not an error but is a SD of sample means (or proportions) from that of the universe or population mean (M). This SD of the sample mean with that of the population mean is called the "SEM," denoted as SE $\bar{x}$ or simply the SE.

Standard error defines the limits of variation that occur mainly due to chance. Variation beyond these limits is significant and is probably due to play of some external factors.

Similarly, the SE of a proportion gives the SD of proportions of several samples of a qualitative data taken from the same population and is denoted as SE of proportion (SEP).

The SE can be reduced by increasing the size of the sample. To illustrate, if the average number of attacks of diarrhea per child per year is four in a particular area, we will not necessarily get an average of four per child in a sample of children from a hygienic area, where it may be only two per child or from an unhygienic area, where it may be six per

JAYPEE BROTHERS

child. These variations in samples are inherent. If we take a very large sample of children from the three areas, included almost equally, the average attacks per child will be more or less four per year. Thus by increasing the size of the sample, the variations can be reduced and vice versa.

On the other hand, if the variations in means are large, such as 10 attacks in one example and 30 and 40 in others, the SE will be increased. Thus the SE depends upon two factors.

1. Standard deviation of the sample, which measures the variability of the observations (i.e., SD of the means).
2. Size of the sample (n).

Thus SE can be calculated by the formula:

$$SE = \frac{SD}{\sqrt{n}}$$

Thus, it is clear that SE varies inversely with the root of the sample size (n) and directly proportional to the SD of the means.

Since the distribution of the means follows the pattern of normal distribution **(*see* Fig. 26.15)**, it is not difficult to visualize that 95% of the sample means will lie within the limits of 2 SE [Population mean $M \pm 2$ SE or $M \pm 2\frac{SD}{\sqrt{n}}$ on either side of the population mean (M).

Therefore, the chance that the population mean (M) lies between the limits defined by sample mean ± 2 SE is also 95%. This is referred to as 95% "confidence limits" (or confidence interval). The confidence of the limits is increased to 99% by increasing the number of SEs to ± 3 SE.

In other words, 95% confidence limits or interval which covers the range, mean ± 2 SE, includes 95% of sample values in normal distribution. Any value lying outside this range will be rare. Its probability of occurrence (p value) by chance will be only 5%, i.e., 0.05 or 1 in 20.

Similarly, 99% confidence limits or interval cover the range, mean ± 3 SE, includes 99% of sample values. The probability (chance) of any value falling outside the range is 1% or 0.01.

Standard Deviation V/S Standard Error

Sl. No.	*Standard deviation*	*Standard error*
1.	It is a measure of square root of the variance*	It is not an error but a deviation of the sample mean (m) from the population mean (M).
2.	Formula: $SD = \sqrt{\frac{\Sigma(x_1 - \bar{x})^2}{n-1}}$	$SE = \frac{SD}{\sqrt{n}}$ (where n = sample size)
3.	It is stable. Does not change much with repeated samples from the same population.	It is variable. Changes with different sample sizes from the same population.
4.	Not affected by sample size	Error decreases as the sample size increases.
5.	It is more sensitive	It is less sensitive
6.	It is a descriptive statistic. Not used in statistical tests.	It is an inferential statistic. Essential for t-tests, Z-tests, etc.

* Variance is the average of the squares of the deviation from the mean.

SIGNIFICANCE

"Significance" is opposite of "chance." If the (m) mean of one sample differs from that of another sample or of the universe (*M*) by more than two times the SE (2 SE), i.e., lying outside the 95% confidence limits, the difference is said to be "statistically significant" at 5% level (probably due to play of some external factors and not due to chance) (i.e., p value lesser than 0.05) **(Fig. 26.20)**.

A value lying outside 3 SE or 99% confidence limits is very rare and the difference is considered to be "highly significant" (P value lesser than 0.01) at 1% level.

This level of significance is expressed as "p" value (i.e., probability value). If the p value is less than 0.05 (or 5%), it is said to be significant and if less than 0.01 (1%), it is highly significant. 5% level of significance is also called "critical level of significance."

Thus the "p" value denotes whether the difference in the sample estimates (sampling variation) is due to chance or due to an external (reality) factor.

To illustrate, suppose on weighing 100 mangoes individually and found that 95% of them weigh between 435 g and 565 g, we can say that 101st mango has 5% chance of weighing either lesser than 435 g or more than 565 g; in either case it means that the mango has lost or achieved a "significance weight," because that weight is lying beyond 95% confidence limits (beyond 2 SE). If it is beyond 99% confidence limits (beyond 3 SE) it is considered "highly significant."

Null Hypothesis (Denoted as H_0)

The techniques to know how far the differences between the estimates of different samples are due to sampling variation are known as the "testing of hypothesis." This can be worked out by several methods.

The first step in testing of hypothesis (e.g., smoking, resulting in cancer; drug, resulting in cure of a disease; or a vaccine, resulting in protection, etc.) is "null hypothesis" (H_0)

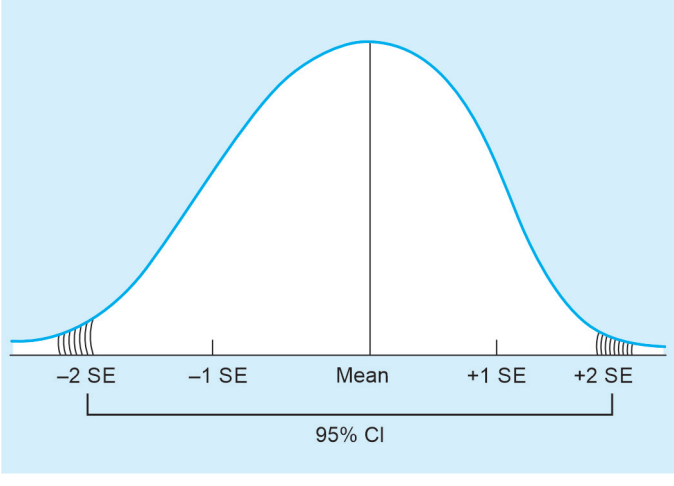


Fig. 26.20: Normal curve showing 95% confidence limits. The shaded areas indicate values that fall beyond these limits on either side.

JAYPEE BROTHERS

i.e., to study the effect of an external factor (smoking, drug, vaccine) on man in health and disease.

Assumption is made that the external factor plays no role. Such an assumption is called "Null hypothesis." Smoking has no relation with lung cancer. Similarly, there is no difference between two drugs or vaccines. Then one proceeded to test the hypothesis (e.g., smoking results in lung cancer) in quantitative terms. In such studies, two groups or samples called experimental group (smokers) and control group (nonsmokers, taken for comparison) are taken. After the period of observation, or at the end of the experiment, the values (proportions or means) of the two samples or groups are noted. If the difference between the samples is more than twice the SE so that the probability of chance occurrence is less than five times in 100 experiments, the null hypothesis (of no difference between two drugs or no relation between smoking and lung cancer) is rejected, and the hypothesis that "smoking results in lung cancer" is accepted. That means if the "*p* value" is less than 0.05, that means the difference is real (and not due to chance), and the factor under trial is considered to have definitive action on the experimental group. If the probability is more than 5% (p value more than 0.05), the null hypothesis is accepted. That means the difference is by chance. Usually the level of significance (p value) is stated while accepting or rejecting the hypothesis.

In certain conditions the null hypothesis (H_0) is rejected though it should be accepted, when there is no significant difference, i.e., when the observation lies within 95% confidence limits. This is stated as "type 1 error."

Similarly, in certain other conditions, the null hypothesis is accepted, though it should be rejected, when there is significant difference, i.e., when the observation lies beyond 95% confidence limits. This is stated as "type 2 error."

By increasing the limit or level of significance to 99%, type 1 error is increased.

Tests of Significance

These are the statistical tests or mathematical methods employed to measure the probability or chance of occurrence of biological variation, in samples by SE. It defines the limits of variation by chance and determines the probability of the factor playing the role.

For example, suppose the mean weight gain in 100 children receiving a nutritional supplement was 2 kg after 9 months (increase in mean weight of 15–17 kg) while the gain in the control group was 1 kg (increase of mean weight of 15–16 kg). Whether this difference of 1 kg in weight gain is a chance occurrence of no significance or whether it is attributable to supplementary feeding has to be decided to establish the value of latter.

The test of significance will measure the probability (p value) of increase in weight by chance. If the p value is less than five times out of 100 ($p < 0.05$), the difference of weight gain of 1 kg is significant statistically. That means the difference is due to nutritional supplement in more than 95% of such experiments. It is said to be significant at 5% level of significance, i.e., the difference occurs by chance in only one in twenty such trials.

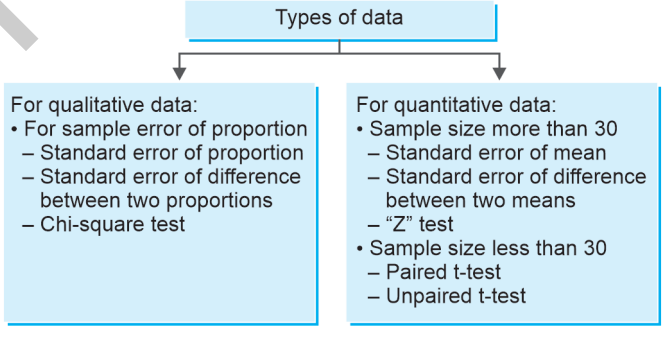
The difference is considered highly significant, if the *p* value is less than 0.01.

General Procedures for Testing a Hypothesis

- A "null hypothesis" (H_0) suitable to the problem is set up
- An alternate hypothesis is defined if necessary
- A suitable statistical test, using a relevant formula, is calculated
- The degree of freedom (DF) is determined
- Then the probability level (*p* value) is found out, corresponding to the calculated value of the test and its DF. This can be found out from the tables
- If the p value is less than 0.05, the null hypothesis is rejected and if the p value is more than 0.05, null hypothesis is accepted.

The different tests employed are as follows:

- Standard error of mean
- Standard error of the difference between two means
- Standard error of proportion
- Standard error of difference between two proportions (SEDP)
- "t" test
- Chi-square test.



Standard Error of Mean (SE $\bar{x}$)

Simply denoted as SE, it is a parameter used to measure the variation among the sample means (m_1, m_2, m_3, etc.) (i.e., sampling variation). SE is not an error, but it is the SD of the sample means from the population mean (M) (explained already). It helps to measure the chance variation or biological variation. [It helps to measure the chance difference between the sample mean (m) and the population mean (M).]

Uses of SE $\bar{x}$

- To find the confidence limits of population mean, if SD of the sample is known. Population mean (M) = Sample mean (m) ± 2 SEM (for 95% confidence limits).
- To tell whether a particular sample is drawn from the known population or not, if the population mean (M) is known. Here the characteristics of the normal curve are used. If the sample mean (m) ($\bar{x}$) lies within the range of *M* ± 2 SE $\bar{x}$. It is considered belonging to the known population

JAYPEE BROTHERS

because 95% of the sample means lies within M ± 2 SE $\bar{x}$. If the sample mean is outside this range (beyond 95% confidence limits) then the sample is considered to be from some other population or it is under the influence of some other factor.

- To find the SE of difference between two means to know if the observed difference between the means of two samples is real and statistically significant or it is apparent and insignificant due to chance (i.e., to test the significant difference between two sample means).
- To calculate the size of the sample in order to have desired confidence limits, SD of the population is known, by the formula $SE\,\bar{x} = \frac{SD}{\sqrt{n}}$

Calculation of Standard Error of Mean

First the SD of means of samples is calculated, by squaring the summation of the difference between sample mean and population mean, by using the formula,

$$SD = \sqrt{\frac{\Sigma(m-M)^2}{n}}$$

But in practice, the SD of means of all samples is not known because only one sample is drawn of which SD and mean are calculated, by the formula $\frac{SD}{\sqrt{n}}$, i.e., SD is divided by the square root of the number of observations.

Example: Mean and SD for the height of 50 boys were 150 and 7 cm, respectively. Find SEM and 95% confidence limits of heights in nature. Could this sample be from the universe with a population mean (M) of 154 cm?

Mean height = 150 cm

SD = 7 cm

n = 50 boys

$$SE\,\bar{x} \frac{SD}{\sqrt{n}} = \frac{7}{\sqrt{50}} = \frac{7}{7.07} = 0.98$$

95% confidence limits (or 2 SE range) in nature will be

Mean height ± 2 SE

150 ± 2 × 0.98 = 150 + 1.96 = 151.96

and 150 – 1.96 = 148.04

This sample of 50 boys with a mean height of 150 cm is not drawn from the universe with a population mean of 154 cm, because 154 is much more than 95% of confidence limit of 151.96 cm.

$$\text{Relative deviate, } Z = \frac{\text{observation} - \text{mean}}{SE\bar{x}}$$

$$= \frac{154-150}{0.98} = \frac{4}{0.98} = 4.08$$

4.08 is more than four times the SE.

Since this ratio is more than 3 (i.e., beyond 3 SD), it is considered to be highly significant, i.e., beyond 99% confidence limits. So the probability of this sample (with a mean height of 150 cm) being drawn from the specified universe (with a mean height of 154 cm) is less than 1% ($p < 0.01$).

Standard Error of the Difference Between Two Means

Denoted as SE ($\bar{x}_1 - \bar{x}_2$). Unlike the SE $\bar{x}$, which measures the chance difference between the sample mean and the population mean (m – M), standard error of the difference between two means [SE of difference in means (SEDM)] measures the chance difference between the means of the two samples (a pair of samples) drawn from the same population.

Two samples (s_1 and s_2) are simultaneously drawn from difference in these two means is noted ($\bar{x}_1 - \bar{x}_2$). It will be plus or minus few centimeters.

The samples are returned to the universe and two fresh samples (s_3 and s_4) are obtained, their means are calculated, the difference in their means will be ($\bar{x}_3 - \bar{x}_4$) plus or minus few centimeters. If the procedure is repeated again and again, the mean of the differences in the paired samples would be zero or nearly so.

Moreover, the differences in means of the different sets of twin samples will follow normal frequency distribution. The SD of these differences therefore will define the range within which 95% of differences in the means would lie. Such a SD is actually called "SEDM," (unlike $SE\bar{x}$ which is the SD of sample means).

In practice, it is not possible to find differences of large number of samples and then find SE of these differences. The test is applied to one pair directly if SDs of two means are known.

If the observed difference between the two means is more than twice the SE of difference, it is significant at 95% confidence limits.

If the observed differences between the two means are greater than three times the SE of difference, it is significant at 99% confidence limits. It is real variability in more than 99% cases and the chance factor is less than 1%.

Apply relative deviate (Z) test for large samples and "t" test, for small samples, for correct estimation of chance limits, because WS Gosset observed that the normal distribution gives biased results in case of small samples.

Standard error of difference between two means is denoted as SE ($\bar{x}_1 - \bar{x}_2$). Mathematically it is the root of the sum of the squares of the SE of two samples means.

It is already seen that the SE of a sample mean is equal to $\frac{SD}{\sqrt{n}}$

$\therefore$ SEDM denoted as $S\bar{d} = \sqrt{(SE_1)^2 + (SE_2)^2}$

$$\text{Substituting } SE = \frac{SD}{\sqrt{n}} = \frac{SD^2}{n}, S\bar{d} = \sqrt{\frac{(SD_1)^2}{n_1} + \frac{(SD_2)^2}{n_2}}$$

Procedure

- Calculate the two means $\bar{x}$ and $\bar{x}_2$, corresponding to the two samples with samples sizes n_1 and n_2, respectively.

JAYPEE BROTHERS

- Set up the null hypothesis that the two samples are from the same population and that the difference between the two sample estimates is due to sampling variation.
- Calculate the SD of the two samples and their SEs (SE_1 and SE_2), respectively.
- Calculate the SEDM as $\sqrt{(SE_1)^2 + (SE_2)^2}$
- Calculate the quantity Z =

$$\frac{\text{Difference between sample estimates}}{\text{SE of difference}}$$

- Refer to normal distribution table and corresponding to this calculated value of Z, find the value of probability "p."
- If the *p* value is less than 0.05, (or *Z* value is >2 SE) reject the null hypothesis and conclude that the difference between two sample estimates as significant. If the Z value is >3 SE, ($p < 0.01$) it is highly significant.

If the p value is greater than 0.05, accept the null hypothesis and conclude that the difference between the two sample estimates as insignificant.

Example 1: In a study on growth of children, one group of 100 children had a mean height of 60 cm and SD of 2.5 cm, while another group of 150 children had a mean height of 62 cm and SD of 3 cm. Is the difference between the means statistically significant?

$$\text{SEDM} = \sqrt{(SE_1)^2 + (SE_2)^2}$$

$$= \sqrt{\frac{(SD_1)^2}{n_1} + \frac{(SD_2)^2}{n_2}}$$

$$= \sqrt{\frac{(2.5)^2}{100} + \frac{(3)^2}{150}}$$

$$= \sqrt{\frac{6.25}{100} + \frac{9}{150}}$$

$$= \sqrt{0.06 + 0.06}$$

$$= \sqrt{0.12}$$

$$= 0.35$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{60 - 62}{0.35} = \frac{2}{0.35} = 5.71$$

Since the observed difference is more than three times the SE, it is highly significant ($p < 0.01$). The growth is more in the second group than in the first.

Example 2: Pregnant women attending an *Anganwadi* and receiving nutritional supplementation numbering 49 were matched with 64 pregnant women not attending an *Anganwadi* and not getting the supplementation. Both groups were followed up. The mean birth weight of babies born to the former was 3.5 kg with SD 1.4 kg and that of those born to the latter, 3.0 kg with SD 1.6 kg. Is the higher birth weight in the former due to the nutrition supplementation?

First, the SEM for each of the group is calculated as follows:

$$\text{SEDM} = \sqrt{(SE_1)^2 + (SE_2)^2}$$

$$= \sqrt{\frac{(SD_1)^2}{n_1} + \frac{(SD_2)^2}{n_2}}$$

$$= \sqrt{\frac{(1.4)^2}{49} + \frac{(1.6)^2}{64}}$$

$$= \sqrt{\frac{1.96}{49} + \frac{2.56}{64}}$$

$$= \sqrt{0.04 + 0.04}$$

$$= \sqrt{0.08}$$

$$= 0.28$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{3.5 - 3.0}{0.28} = \frac{0.5}{0.28} = 1.78$$

Since the observed difference is less than two times the SE, (within 95% confidence limits), it is not significant. That means the difference in mean birth weight is due to chance and not due to nutritional supplementation.

Standard Error of Proportion

The SEM and SEDM are the tests of significance of variation in means of samples (more than 30) of quantitative data. The SEP and the SEDP are the tests of significance of variation in proportions of samples of qualitative data (more than 30).

In qualitative data the character remains the same while the frequency variations are dealt with.

If a sample consist of characters of two attributes (positive and negative) only, such as male and female, rich and poor, vaccinated and nonvaccinated, died and survived, successes and failures, etc. it is a sample of binomial classification and if the division of the sample is made into more than two classes, such as blood groups A, B, AB, and O or WBCs into polymorphs, lymphocytes, eosinophils, etc. the sample is of multinomial classification.

The proportion of individuals, having a specific character or attribute, in a binomial distribution, is expressed as "p," either as a fraction of 1 or as percentage.

$$p = \frac{\text{Number of individuals having a specific character}}{\text{Total number in the sample}}$$

The remaining proportion of individuals having the other (negative) characters is represented as "q."

Arithmetically, q = 1 – p in terms of fraction of 1 or 100 – p in percentage.

Thus "p" is the probability of occurrence of a positive attribute and "q" is the probability of occurrence of the negative attribute (or nonoccurrence of the same positive attribute).

JAYPEE BROTHERS

To illustrate, suppose there are 80 girls in a class of 200 strength, the proportion of girls and boys is estimated as follows:

Proportion of girls (p) = $\frac{80}{200}$ = 0.4 out of 1 in the sample

Proportion of boys (q) = $\frac{200-80}{200} = \frac{120}{200}$ = 0.6 out of 1

or q = 1 – p
= 1 – 0.4
= 0.6

For statistical analysis, the proportion of girls (p) and boys (q) are expressed in percentage as below:

p = $\frac{80\times100}{200}$ = 40% girls

q = 100 – p = 100 – 40 = 60% boys in the example.

Suppose from the universe (population), multiple samples are drawn, the value of the particular characteristic (otherwise called proportion) in each sample is represented as "p" (just like the sample mean, $\bar{x}$), the proportion of the samples (p_1, p_2, p_3, ..., etc.) have a tendency to concentrate around the proportion of the universe or population and represented as "P" (just like the sample means m_1, m_2, m_3, i.e., $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, ..., etc. have central tendency around population mean M).

The sample proportions are symmetrically distributed around the population proportion, from which the samples are drawn, i.e., it gains the shape of a normal curve.

This distribution of the sample values (proportions) around the population proportion (P) is expressed arithmetically in terms "SEP." In other words, SEP is the SD of the sample proportions (just like SEM is the SD of the sample means).

The binomial confidence limits are as follows:

- 68% of sample proportions will lie within the range P ± 1 SEP.
- 95% of values lie within the range of P ± 2 SEP (95% confidence limits).
- 99% of values lie within the range of P ± 3 SEP (99% confidence limits).

Standard error of proportion is defined as a measure of variation occurring by chance between the sample proportion (p) and the population proportion (P) in a qualitative data.

This test of significance is employed to find the efficacy of a drug or a vaccine or a surgical procedure, etc. To illustrate, if 35% cases died in previous epidemic and 30% in the present epidemic, whether this reduction of 5% is due to better treatment or increased immunity or decreased virulence of the pathogen or by chance is determined by calculation of SEP.

Formula:

$$SEP = \frac{p\times q}{n}$$

where *p* is the percentage of positive character (proportion), q is the percentage of negative character (100 – p), n is the number of observations (size of the sample).

The significance of difference is found by relative deviate (*Z*) test.

$$Z = \frac{\text{Observed difference}}{SEP}$$

Example 1: The proportion of blood group A among Indians is 30%. Find SEP and the 95% confidence limits. Could this sample be from a universe, in which the prevalence of blood group A is 40%?

p = 30; q = 100 – 30; n = 100

$$SEP = \sqrt{\frac{p\times q}{n}} = \sqrt{\frac{30\times70}{100}} = \sqrt{21} = 4.58$$

95% confidence limits = p ± 2 SE

= 30 + 2 × 4.58 = 30 + 9.16 = 39.16

30 – 2 × 4.58 = 30 – 9.16 = 20.84

Since the proportion 40 is outside the 95% confidence limits, we can say with 95% confidence that the sample is not drawn from the universe having 40% prevalence of blood group A.

Example 2: What size of the sample will you take for assessing diabetics in an urban population where the prevalence was given as 3% in age group above 15 years?

$$SEP = \sqrt{\frac{p\times q}{n}}$$

$$0.3 = \sqrt{\frac{3\times97}{n}}$$

For finding the suitable size of the sample, the assumption made is that SEP does not exceed 10% (or one tenth) of the positive character, in a qualitative data.

$\therefore$ 1/10 of 3 = 0.3

$\therefore$ $(0.3)^2 = \frac{3\times97}{n}$ (sq. root is removed, n = sample size)

$\therefore$ $n = \frac{291}{0.3\times0.3} = \frac{291}{0.09} = 3233$ persons

Uses of standard error of proportion

- Standard error of proportion is used to find the confidence limits of population proportion (P) if sample proportion (p) is given.
- To tell whether the sample drawn from the known population or not.
- To find the SEDP, to know if the observed difference between proportions of two samples is real and statistically significant or due to chance and insignificant.
- To find the size of the sample if SEP and the proportions are known.

JAYPEE BROTHERS

Standard Error of Difference Between Two Proportions

The difference in the pairs of proportions of each set of twin samples drawn from the sample universe follows normal distribution, so that their SD will be the yardstick to determine the range of the attribute, within which 95% of differences in proportions will lie.

Standard error of difference between two proportions measures the chance between the sample proportions of paired samples drawn from the same population or universe. If the observed difference between the two sample proportions (p_1 and p_2) is more than twice the SE of difference, it is significant at 95% confidence limits, and if more than 3, it is significant at 99% confidence limits.

The significance of difference is found by relative deviate "Z" test. SEDP is calculated by the formula,

$$SE\,(p_1 - p_2) = \sqrt{\frac{p_1q_1}{n_1} = \frac{p_2q_2}{n_2}}$$

where p_1 and p_2 are the estimates of the proportions of the two samples.

q_1 and $q_2 = (1 - p_1)$ and $(1 - p_2)$, respectively.
(i.e., $100 - p_1$ and $100 - p_2$ in terms of %).

n_1 and n_2 = numbers of observations in the two samples.

$$Z = \frac{\text{Observed difference between two sample proportions}}{\text{SE of difference between two proportions}}$$

Then p value is found by referring to normal distribution table corresponding to *Z* value.

If the p value is less than 0.05, null hypothesis is rejected and concluded that the difference between two sample estimates is significant and vice versa, i.e., if the p value is more than 0.05, the null hypothesis is accepted and concluded that the difference between the sample estimates is insignificant.

Example 1: In an epidemic of gastroenteritis in an area, the number of cases reported in two populations consuming water from two different sources were as follows:

Source of water	*No. of people consuming water from the source*	*No. of cases*
Tap water	800	35
Hand pump water	2,400	120
Total	3,200	155

Find out whether the difference in the proportion of cases in the two groups is significant.

Ans:

1. The null hypothesis in this case is that the difference is insignificant.
2. p_1 = Proportion of cases among tap water consumers = 0.044 (i.e., 35 ÷ 800 = 0.044)
   $q_1 = 1 - p_1 = 1 - 0.044 = 0.956$
   n_1 = No. of tap water consumers = 800.
3. Standard error of proportion p_1 (SE_1)
$$= \sqrt{\frac{p_1q_1}{n_1} = \frac{0.044 \times 0.956}{800}}$$
$$= \sqrt{0.0000525}$$
$$= 0.0072$$
4. p_2 = Proportion of cases amongst hand pump water consumers
$$= (120 \div 2{,}400) = 0.05$$
$$q_2 = 1 - p_1 = 1 - 0.05 = 0.95$$
n_2 = Number of hand pump water consumers = 2400.
5. Standard error of proportion p_2 (SE_2)
$$= \sqrt{\frac{p_2q_2}{n_2} = \frac{0.05 \times 0.95}{2400}}$$
$$= \sqrt{0.0000197}$$
$$= 0.0044$$
6. Difference between two proportions = $p_1 - p_2$
$$= 0.044 - 0.05$$
$$= 0.050 - 0.044$$
$$= 0.006$$
7. Standard error of difference
$$= \sqrt{(SE_1)^2 + (SE_2)^2}$$
$$= \sqrt{0.0000525 + 0.0000197}$$
$$= \sqrt{0.0000722}$$
$$= 0.0085$$
8. $$Z = \frac{\text{Difference between the proportions}}{\text{Standard error of difference}}$$
$$= \frac{0.006}{0.0085}$$
9. From the normal distribution table, "P" corresponding to this value is more than 0.48 (or the "Z" value is less than 2), the null hypothesis is not rejected, and it is concluded that the difference is insignificant.

Example 2: If typhoid mortality in one sample of 100 is 20%, and in another sample of 100, it is 30 %, find the SEDP by both the methods. Is the difference in mortality rates significant?

Formula A:

$$p_1 = 20,\ q_1 = 100 - 20 = 80,\ n_1 = 100$$
$$p_2 = 30,\ q_2 = 100 - 30 = 70,\ n_2 = 100$$

$$SE\,(p_1 - p_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

JAYPEE BROTHERS

$$= \sqrt{\frac{20\times80}{100}+\frac{30\times70}{100}}$$

$$= \sqrt{16+21}$$

$$= \sqrt{37}$$

$$= 6.08$$

Formula B:

$$P = \frac{p_1+p_2}{n_1+n_2}\times100$$

$$= \frac{20+30}{100+100}\times100$$

$$= \frac{50\times100}{200}$$

$$= 25\%$$

$$Q = \frac{q_1+q_2}{n_1+n_2}\times100$$

$$= \frac{80+70}{100+100}\times100$$

$$= \frac{150\times100}{200}$$

$$= 75\%$$

$$SE\,(p_1-p_2) = \sqrt{P\times Q\left(\frac{1}{100}+\frac{1}{100}\right)}$$

$$= \sqrt{25\times75\left(\frac{2}{100}\right)}$$

$$= \sqrt{\frac{25\times75\times2}{100}}$$

$$= \sqrt{\frac{75}{2}} = \sqrt{37.5} = 6.09$$

$$Z = \frac{\text{Observed difference}}{\text{SE of difference}} = \frac{30-20}{6.09} = 1.6$$

Conclusion

Since Z value is less than 2, it is insignificant at 95% confidence limits.

The above test is applicable in the case of large samples. For small samples of qualitative data, the Chi-square test is usually applied. However, the results of both the tests are the same. And for small samples of quantitative data, of less than 30, the significance of difference between two means or proportions is tested by "*t*" test.

Contingency Table

Contingency table, also known as Cross tabulation or Cross tab, is a statistical tool, wherein the frequency distribution (i.e., observed values) of at least two variables (ex. smoking and lung cancer, sexual relation and HIV/AIDS) or even multiple categories of variables (e.g. blood group and lepromatous and non-lepromatous leprosy) are displayed in the form of a matrix format (i.e., diagrammatic table), consisting of rows and columns, providing cells, where in the observed values are displayed.

Structure: Minimum, there must be two variables. Rows represent categories of one variable (smoking). Columns represent categories of another variable (lung cancer). Cells show the frequency distribution (i.e., observed values). Margins show row-total, column total, sub-total and grand total.

Example: A study examining the relationship between smoking and lung cancer in a population of 100, consisting of 50 smokers and 50 non-smokers, the contingency table is prepared as follows. Out of 50 smokers, 30 developed lung cancer 20 not developed and out of 50 nonsmokers, 10 developed lung-cancer, 40 did not develop.

	Lung cancer			
Smoking status ↓	*Yes (diseased)*	*No (not diseased)*	*Total*	
Yes (smokers)	30 (a)	20 (b)	50 (a + b)	Row Sub total
No (non-smokers)	10 (c)	40 (d)	50 (c + d)	
	40 (a + c)	60 (b + d)	100	Grand Total
	Column subtotal			

This table is also called 2 × 2 contingency table, because there are 2 rows and 2 columns.

The association between smoking and lung cancer is studied by calculating the exposure rate among the cases and exposure rate among the controls (i.e., non-diseased).

[Exposure rate among the cases means, how many cases gave the history of smoking. Similarly, exposure rate among controls, means how many of non-diseased person gave the history of smoking].

$$\text{Exposure rate among the cases} = \frac{a}{a+c} = \frac{30}{40} = 0.75$$

$$\text{Exposure rate among the controls} = \frac{b}{b+d} = \frac{20}{60} = 0.33$$

Since the exposure rate among the cases is more than the exposure rate among controls, there is a definite relation between smoking and lung cancer.

To know whether this relation between smoking and lung cancer is by chance or is it a fact? This is studied by applying "Chi-square" test (explained below).

Uses and applications of contingency table:

1. Helps to identify the relationship between two variables (e.g., smoking and lung cancer, oral pills and myocardial infarction, etc.)

JAYPEE BROTHERS

2. Helps to perform statistical tests like "Chi-square", "odd's ratio", "Relative risk" and "Attributable risk".

Chi-square Test (χ^2 Test)

"χ" is a Greek letter, not equivalent of English letter "X," written as "chi," pronounced as "Kye" and typed as "χ'"

This test involves the calculation of a quantity, called "χ^2" in a qualitative data. χ^2-test was developed by Karl Pearson.

Applications of χ^2 test

- It is an alternate method of testing the significance of the difference between two proportions.
- As a test of goodness of fit:
  By χ^2 test, we can find whether the observed frequency distribution fits in a hypothetical or theoretical distribution of a qualitative data. The χ^2 test determines whether the observed frequency distribution differs from the theoretical distribution by chance or the sample is drawn from a different population.
  (Theoretical distribution, such as normal, binomial, or Poisson).
  If the calculated value of χ^2 test of the sample is higher than the table value, it is significant (null hypothesis is rejected) and vice versa.

Example: The ratio of male to female births in nature is 1:1. If 52 male and 48 female children are born in a small town, could this difference be due to chance?

	Male	*Female*
Observed frequencies (*O*)	52	48
Expected frequencies (*E*)	50	50

$$\chi^2\,(a) = \frac{(O-E)^2}{E} = \frac{(52-50)^2}{50} = \frac{4}{50}$$

$$\chi^2\,(b) = \frac{(48-50)^2}{50} = \frac{4}{50}$$

$$a + b = \frac{4+4}{50} = \frac{8}{50} = 0.06$$

Degree of freedom = 1

χ^2 table value, with a DF at 5% level of significance = 3.841 (Annexure IV).

Since the calculated value is much less than the table value, it is insignificant. That means the observed sex difference is by chance.

- To find any association between two attributes (e.g., smoking and lung cancer) is real or by chance.

Similar examples are association between blood group and incidence of leprosy, age and BP, nutritional status of the mother and weight of the newborn, obesity and diabetes, etc. There are only two possibilities, either the attributes are associated or not associated. The χ^2 test measures the probability of association by chance.

According to null hypothesis, the assumption of no association between two characters is made (no relation between smoking and cancer). If the χ^2 value is higher than the table value of χ^2 (Annexure IV) against a probability of 0.05, for the particular DF, the null hypothesis is rejected, and the association is not apparent but real, at 5% level of significance.

The χ^2 test can also be applied to find the association in two classes or groups, as in multinomial samples (or a sample with multiple characters, e.g., a group with <10 smokes/day, 11–20, 21–30, >30 smokes per day and the incidence of lung cancer).

Steps:

1. First, a table is prepared out of the qualitative data. Actual observed frequencies of two sets of events are entered in a two-way table, which is also known as "contingency table" (Latin, con = together; tangere = to touch). Since this table also helps to know the association between two sets of events, the table is also called "association table." Because there are only two events and two groups or classes, it is called 2 × 2 or fourfold or four-cell contingency table.
2. Null hypothesis is set up stating there is no association between the events.
   χ^2 text can also be applied when there are more than two classes or groups, such as social classes I, II, III, and IV among smokers and nonsmokers.
3. Then expected frequency for each cell is calculated on the assumption of no association, using the formula

$$E = \frac{\text{Row total} \times \text{column total}}{\text{Grand total}}$$

4. Then the difference between the observed and the expected frequencies for each cell is found (i.e., O–E).
5. χ^2 value for each cell is calculated by using the formula

$$\chi^2 = \frac{(O-E)^2}{E}$$

6. Then the total of χ^2 for all the four cells is calculated by the formula (summation of all four cell χ^2 values)

$$\text{Total } \chi^2 = \Sigma\frac{(O-E)^2}{E}$$

$$\text{Alternate formula for } \chi^2 = \frac{(ad-bc)^2 \times G}{(a+b)(c+d)(b+d)(a+c)}$$

   Where a, b, c, and d are observed frequencies of four cells, and G is the sample total (or grand total).
7. Next the DF is calculated by using the formula, DF = (c – 1) (r – 1), where c = no. of columns, r = no. of rows.
8. Lastly, to know whether the calculated χ^2 value is significant or not (i.e., whether the relation between the

JAYPEE BROTHERS

events smoking and cancer, e.g., is by chance or a real one), we have to refer to "Fisher's χ^2 table" (Annexure IV) for a particular DF, for a probability of 0.05, 0.01, etc. If the calculated value is higher than the table value, it is concluded that it is significant, and the null hypothesis is to be rejected and accept the hypothesis (e.g., no association between smoking and lung cancer is ruled out and accept the hypothesis that there is significant association between smoking and cancer). If the calculated χ^2 value is lower than the table value, null hypothesis is accepted.

Example 1: Apply χ^2 test to find efficacy of a drug from the data given below:

Outcome (result) of treatment with drug and placebo:

	(Efficacy of the drug) Result			
Group	***Died***	***Survived***	***Total***	
A. Control (no placebo)	(O) 10 (a) (E) (5.25)	(O) 25 (b) (E) (29.75)	35	(a + b)
B. Experimental (on drug)	(O) 5 (c) (E) (9.75)	(O) 60 (d) (E) (55.25)	65	(c + d)
Total	15 (a + c)	85 (b + d)	(G) 100	

The above values are all observed values.

The null hypothesis that the drug has no effect (drug and placebo are same) (i.e., there is no difference between the sample proportions and the population/universe proportion of 100).

The expected (E) value and χ^2 value is calculated for each cell as follows:

a. Expected number and χ^2 value of "died" in the control group.

$$E_a = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

$$= \frac{35 \times 15}{100} = 5.25$$

$$\chi^2 = \frac{(O-E)^2}{E} = \frac{(10-5.25)^2}{5.25}$$

$$= \frac{(4.75)^2}{5.25} = \frac{22.5625}{5.25} = 4.29$$

b. Expected number and χ^2 value of "survived" in the control group

$$E_b = \frac{85 \times 35}{100} = 29.75$$

$$\chi^2 = \frac{(O-E)^2}{E}$$

$$= \frac{(25-29.75)^2}{29.75} = \frac{(-4.75)^2}{29.75}$$

$$= \frac{22.56}{29.75} = 0.76$$

c. E number and χ^2 value of "died" in the experimental group

$$E_c = \frac{15 \times 65}{100} = \frac{39}{4} = 9.75$$

$$\chi^2 = \frac{(O-E)^2}{E}$$

$$= \frac{(5-9.75)^2}{9.75} = \frac{(-4.75)^2}{9.75}$$

$$= \frac{22.56}{9.75} = 2.31$$

d. E number and χ^2 value of survived in the experimental group.

$$E_c = \frac{85 \times 65}{100} = 85 \times 0.65 = 55.25$$

$$\chi^2 = \frac{(O-E)^2}{E} = \frac{(60-55.25)^2}{55.25} = \frac{(5.25)^2}{55.25}$$

$$= \frac{22.56}{55.25} = 0.408$$

$$\Sigma\chi^2 = \text{Total } \chi^2 \text{ value of all 4 cells}$$

$$= 4.29 + 0.76 + 2.31 + 0.41$$

$$= 7.77$$

$$\text{DF} = (c-1)(r-1) = (2-1)(2-1) = (1 \times 1) = 1$$

where DF is the degree of freedom, *c* is the number of (vertical) columns, and *r* is the number of (horizontal) rows.

On referring to Fisher's χ^2 table, with 1 DF, the tabulated χ^2 value, corresponding to probability of 0.05 (at 95% significance level), is 3.84.

Since the calculated value (7.77) is more than the table value (3.84), the null hypothesis is rejected, accepting the alternate hypothesis, i.e., the assumption that the drug is not efficacious (no difference between drug and placebo) is ruled out and accepted that the drug is efficacious (i.e., the difference between the two death rates is significant).

Note: If the calculated value becomes lesser than χ^2 table values, null hypothesis has to be accepted.

Example 2: In a case–control study of oral contraceptive use and risk of myocardial infarction, it was found that out of 156 women with myocardial infarction, 23 were oral contraceptive users at the time of their hospital admission. Of the 3,120 control women without myocardial infarction, 304 were current oral pill users.

- Construct an appropriate 2 × 2 table
- Calculate the measure of association between the oral contraceptive use and myocardial infarction
- Interpret the results.

JAYPEE BROTHERS

Construction of the appropriate 2 × 2 contingency table

		Disease (MI)		Total
		Case (disease present)	Control (disease absent)	
Oral pill users	Yes	23 (a)	304 (b)	327 (a + b)
	No	133 (c)	2816 (d)	2949 (c + d)
Total		156 (a + c)	3120 (b + d)	(G) 3276 (a + b + c + d) = G

Null hypothesis: There is no association between oral pills and myocardial infarction.

$$\text{Expected (E) value} = \frac{\text{Column total} \times \text{Row total}}{\text{Grand total}}$$

$$E_a = \frac{156 \times 327}{3276} = 15.50$$

$$\chi^2 = \frac{(O-E)^2}{E} = \frac{(23.0-15.5)^2}{15.5} = 3.62$$

$$E_b = \frac{3120 \times 327}{3276} = 311.40$$

$$\chi^2 = \frac{(304-311.4)^2}{311.4} = 0.17$$

$$E_c = \frac{156 \times 2949}{3276} = 140.40$$

$$\chi^2 = \frac{(2816-2808.5)^2}{2808.5} = 7.50$$

$$DF = (c-1)(r-1) \qquad \Sigma\chi^2 = 11.68$$
$$= (2-1)(2-1)$$
$$= 1$$

χ^2 value (table value with 1 df at p value 0.05 is 3.841).

Results:

- Since the observed or calculated value 11.68 is more than the table value 3.841, null hypothesis (H_0) is rejected and the alternate hypothesis is accepted that there is significant association between ($p < 0.05$) oral pills and myocardial infarction.

$$\text{Odds ratio} = \frac{ad}{bc} = \frac{23 \times 2816}{304 \times 133} = 1.6$$

Oral contraceptive users are 1.6 times at a greater risk of getting myocardial infarction than nonusers.

- Exposure rate among cases and controls are compared.

$$\text{Exposure rate among cases} = \frac{a}{a+c} = \frac{23}{156} = 0.15$$

$$\text{Exposure rate among controls} = \frac{b}{b+d} = \frac{304}{3120} = 0.09$$

Since, the exposure rate among the cases is greater than that of controls, there exists a relation between the oral pills and the myocardial infarction.

Example 3: χ^2 test for tables with more than two rows and columns. The following data in the table shows the distribution of the nonleprosy, lepromatous leprosy, and nonlepromatous leprosy patients according to blood groups A, B, AB, and O. Find out whether there is any association between the blood group and the disease leprosy.

Data: Distribution of leprosy and nonleprosy groups according to blood groups.

Blood group	Nonleprosy	Lepromatous leprosy	Nonlepromatous leprosy	Total
A	30	49	52	131
B	60	49	36	145
O	47	59	48	154
AB	13	12	16	41
Total	150	169	152	471

- **Null hypothesis (H0):** There is no difference (relation between the blood groups and the disease leprosy. (i.e., blood group distributions are same in all the three groups of people).
- Expected frequencies are calculated for each cell (for different groups)

Blood group	Nonleprosy	Lepromatous leprosy	Nonlepromatous leprosy
A	*E* = (131 × 150)/471 = 41.7	= (131 × 169)/471 = 47.0	= (131 × 152)/471 = 42.3
B	E = (145 × 150)/471 = 46.2	= (145 × 169)/471 = 52.0	= (145 × 152)/471 = 46.8
C	E = (154 × 150)/471 = 49.0	= (154 × 169)/471 = 55.3	= (154 × 152)/471 = 49.7
AB	E = (41 × 150)/471 = 13.1	= (41 × 169)/471 = 14.7	= (41 × 152)/471 = 13.2

- χ^2 value is then obtained for each cell by the formula

$$\chi^2 = \frac{(O-E)^2}{E}$$

χ^2 values for each cell.

Blood groups	Nonleprosy	Lepromatous leprosy	Nonlepromatous leprosy	Total
A	3.28	0.09	2.22	5.59
B	4.12	0.17	2.49	6.78
O	0.08	0.25	0.06	0.39
AB	0.00	0.50	0.59	1.09
Total	7.48	1.01	5.36	13.85

The summation of all the χ^2 values of all the cells = 13.85

$$DF = (c-1)(r-1)$$
$$= (3-1)(4-1)$$
$$= 2 \times 3$$
$$= 6$$

The table value of χ^2 at DF 6 for the p 0.05 is 12.592.

JAYPEE BROTHERS

Result

- Since the calculated χ^2 value 13.85 is greater than the table value 12.59, null hypothesis is rejected and the alternate hypothesis is accepted. For example, the relation between the blood groups and leprosy disease is statistically significant.
- In order to understand which of the blood groups are more prone to leprosy, the contribution to the total χ^2 from each cell is to be examined. From the table, it is seen that the maximum contributions are from B and A groups.

Further test among each series will help to understand the problem more precisely.

Nonparametric Tests

When the variables are distributed normally, tests, such as *t* test and χ^2 tests, are applied to test the hypothesis. But at times, there are variables, such as rankings, which do not conform to any probability distribution. In such situation the above tests may not be applicable but the so-called nonparametric tests may be applicable. These are the Sign test, the Wilcoxon's signed rank test, and Mann-Whitney test.

Sign test: This is applied under the following circumstances, e.g., there are two diagnostic procedures of a disease, which are to be tested for their difference or significance. The two procedures are applied on a random sample of patients of the disease. The observations are categorized as follows:

- It is categorized as "one" when a person is considered a patient of that particular disease or as "zero" when not considered a patient, based on the diagnostic procedure. Thus there are two sets of ranks, in terms of "one" and "zero."
- The difference between one and zero corresponds to the two test procedures, applied on a single individual.
- The significance of difference between the two procedures can be tested using the usual χ^2 test.

Wilcoxon's signed rank test: This test is useful for testing the significance of difference in paired observations (e.g., serum fibrinogen degradation products in milligrams per milliliter among 12 patients were noted before and after prostate operation). The difference in the paired values of each unit/patient is ranked according to the size of the difference, and the signs of the original differences are also assigned to the ranks. Under the null hypothesis, it is equally likely that any difference can have + or –. The sums of all ranks with +ve sign and –ve sign are obtained separately.

The smaller sum out of the above two is referred to Wilcoxon's table, without considering the sign, against the number of pairs observed.

When the calculated smaller rank sum is less than the tabulated value, at 5% level, conclusion is made that the difference between the two groups of observations is significant at 5% level.

Suppose there are more than one pair of common values for the differences, they are "tied values." Then each of the tied difference is assigned the average of the ranks that would be assigned if there is no tie. When the number of pairs exceeds 16, which is the limit of number of pairs in the tables given by Wilcoxon, the normal approximation for getting the probability of rejection of null hypothesis is obtained by the formula.

$$Z = \frac{[(1\mu - T) - 1/2]}{\sigma}$$

where

T = smaller rank sum

$\mu = n(n+1)/4$

$$\sigma = \frac{\sqrt{(2n+1)\mu}}{4}$$

n = number of pairs of observations

P is obtained from the normal distribution table corresponding to the value of calculated Z. As a rule of thumb Z >1.96 signifies rejection of the null hypothesis at P<0.05.

Mann-Whitney test: The test is useful for testing the difference between unpaired observations (e.g., control and experimental groups). The observations of both the groups are arranged in the order of magnitude. Tag is made for the common observations in both the groups separately. Ranks are assigned to the observations. When there are common observations, average of the ranks is given to them, as in the case of Wilcoxon's test. Then the sum of ranks of each of the samples (groups) is calculated separately.

The smaller rank sum out of the above two is referred to Mann-Whitney table, which gives the maximum sum of ranks required for the rejection of null hypothesis, under different probability levels.

Whenever the calculated smaller rank sum is less than the tabulated value, the null hypothesis is rejected.

CORRELATION AND REGRESSION CORRELATION

Interpretation of Bivariate Data

Often two variables or attributes are simultaneously recorded (i.e., bivariate data) to know whether one influences the other, e.g., height and weight, age and BP, weight of pregnant mother and weight of the newborn, temperature, and pulse, etc. The purpose of such study is to know whether change in one variable is associated with the change in the other variable. If there is a change like that, then it is assumed that there is a "correlation" between the two variables and the degree of this correlation is measured in terms of "correlation coefficient."

Correlation is the linear relationship or mutuality of association of two variables; it is independent of the units of measurements (whereas χ^2 test indicates the presence or absence of relationship between the qualitative data but not the degree of association).

JAYPEE BROTHERS

Scatter Diagram

Suppose, the correlation of height and weight of 100 children has to be studied and correlation coefficient is to be calculated, the extent and type of the relation between the two variables has to be known by plotting the variables into a scatter diagram, wherein one variable (X_1) is represented on X-axis and another variables (X_2) on the Y-axis. Each pair of observation is plotted as a dot. The diagram will be a scatter of dots (each dot represents height on one axis and weight on another axis).

There are five types of scatter diagrams (*see* **Figs. 26.10A to C and 26.21**).

1. **Diagram 1:** Positive correlation or direct correlation: In this type, the variables denoted on X- and Y-axes are directly proportional. That means as the X value increases, the Y value also increases. The "regression line," which is in the center of the plots, ascends toward the right. For example, height and weight of growing children, temperature and pulse.
2. **Diagram 2:** Negative correlation or indirect correlation: In this type, the variables are inversely proportional to each other, i.e., when one *x* variable increases, the other y decreases, and vice versa. The regression line ascends toward the left. For example, socioeconomic status and incidence of tuberculosis; income and malnutrition.
3. **Diagram 3:** No correlation: In this type, the variables are different and independent. The dots are haphazardly distributed, leading to conclusion that there is no relation. The regression lines are two, inclined on one another, but they are parallel to corresponding axes. For example, Ht and pulse rate.
4. **Diagram 4 (Fig. 26.21A):** Partial positive correlation: In this type, the regression lines incline on one another and ascend toward right. For example, age of the husbands and age of the wives; infant mortality rate and overcrowding, etc.
5. **Diagram 5 (Fig. 26.21B):** Partial negative correlation: In this type, the regression lines incline on one another and ascend toward left. For example, income and infant mortality rate; age and vital capacity among adults.

Correlation Coefficient

It is an indicator to measure the degree of correlation between the two variables and is represented by the symbol "r," the value of which ranges from –1 through 0, to +1, as follows:

$$-1_____0_____+1$$

The correlation measures only the degree of association but does not tell what caused it. Moreover, it does not mean that one variable causes the change in the other. The cause of the change may be due to other factor or factors also.

The formula for the calculation of CV is given by

$$r = \frac{\text{Covariance between two variables}}{\text{SD of first variable} \times \text{SD of second variable}}$$

where "covariance" is the amount of covariation between the two variables, and the SD gives the variation (variance) of a variable within itself.

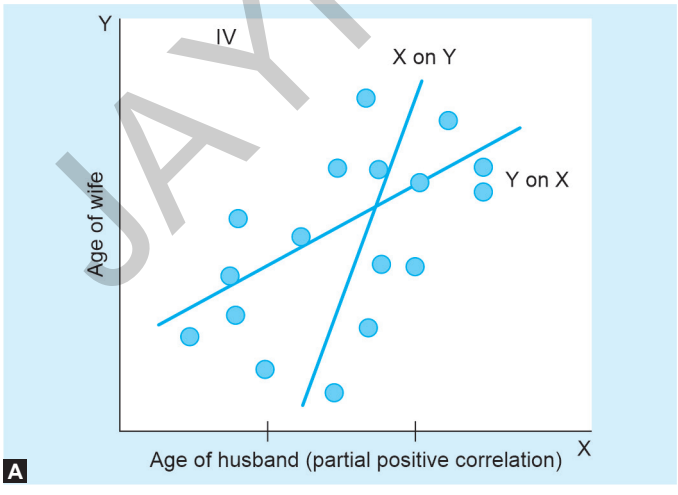
$$\text{i.e. } r = \frac{\Sigma (X - \bar{x})(Y - \bar{y})}{\sqrt{\Sigma (X - \bar{x})^2 (Y - \bar{y})^2}}$$

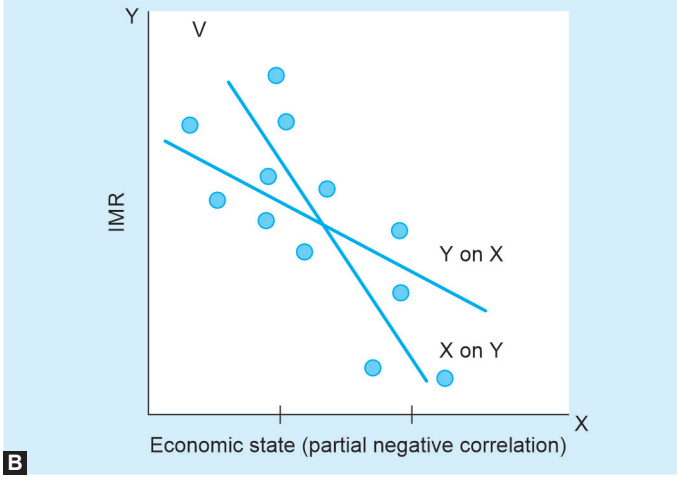
where X and Y are the two characteristics being measured (i.e., mean individual variables).

$\bar{x}$ and $\bar{y}$ = means of the two groups of variables.

The value of "r" varies from –1 to +1. Depending upon the degree and direction of correlation, there are five types of correlation.

1. If the value is +1, it shows perfect positive correlation.
2. If the value is "0," it shows absolutely no correlation.
3. If the value is –1, it shows perfect negative correlation.





Figs. 26.21A and B: Scatter diagram showing: (A) Partial positive correlation; (B) Partial negative correlation.

JAYPEE BROTHERS

4. If it is between –1 and 0, it shows partial negative correlation.
5. If it is between 0 and +1, it shows partial positive correlation.

Since the correlation coefficient is subject to sampling variation, its value has to be tested for its significance. The significance of *r* value can be evaluated by referring to "correlation table," using the correct DF.

To apply the test of significance, first the sampling variation is measured by SE of correlation coefficient (SECC)

$$SECC = \sqrt{\frac{1-r^2}{n-2}}$$

where "r" is the correlation coefficient and *n* is the number of pairs of observations.

"t" test is applied when the sample is very small (i.e., test of significance for correlation coefficient).

Formula is $t = r \times \sqrt{\frac{(n-2)}{1-r^2}}$ with n–2DF

"p" value can be found out by referring to t distribution table (correlation coefficient table).

Example for calculation of correlation coefficient

Height and weight of 18 female medical students recorded in centimeters and kilograms, respectively, and presented as height/weight was observed as follows:

160/55, 160/51, 157/65, 165/61, 165/45, 168/51, 158/49, 162/53, 165/40, 159/70, 159/65, 159/66, 158/55, 162/45, 157/55, 165/50, 155/56, and 161/51.

Height			Weight			
X (cm)	$X - \bar{x}$	$(X - \bar{x})^2$	Y (kg)	$Y - \bar{y}$	$(Y - \bar{y})^2$	$(X - \bar{x})(Y - \bar{y})$
160	–0.8	0.64	55	+0.4	0.16	–0.32
160	–0.8	0.64	51	–3.6	12.96	+2.88
157	–3.8	14.44	65	+10.4	108.16	–39.52
165	+4.2	17.64	61	+6.4	40.96	+26.88
165	+4.2	17.64	45	-9.6	92.16	-40.32
168	+7.2	51.84	51	–3.6	12.96	–25.92
158	–2.8	7.84	49	-5.6	31.36	+15.68
162	+1.2	1.44	53	–1.6	2.56	–1.92
165	+4.2	17.64	40	–14.6	213.16	-61.32
159	–1.8	3.24	70	+15.4	237.16	–27.72
159	–1.8	3.24	65	+10.4	108.16	–18.72
159	–1.8	3.24	66	+11.4	129.96	–20.52
158	–2.8	7.84	55	+0.4	0.16	–1.12
162	+1.2	1.44	45	-9.6	92.16	–11.52
157	–3.8	14.44	55	+0.4	0.16	–1.52
165	+4.2	17.64	50	-4.6	21.16	–19.32
155	-5.8	33.64	56	+1.4	1.96	-8.12
161	+0.2	0.04	51	–3.6	12.96	–0.72
2,895	$\bar{x} = 160.8$	214.52	983	$\bar{y} = 54.6$	118.28	–233.16

Calculation of correlation coefficient value

$$r = \frac{\Sigma (X - \bar{x})(Y - \bar{y})}{\sqrt{\Sigma (X - \bar{x})^2 \, \Sigma (Y - \bar{y})^2}} = \frac{-233.16}{\sqrt{214.52 \times 118.28}}$$

$$r = \frac{-233.16}{\sqrt{2,39,893.42}}$$

$$= \frac{-233.16}{489.79} = -0.48$$

Inference: Since the r value lies between –1 and 0, partial negative correlation exists between the height and weight measurements of female medical students.

Regression

It was observed that the degree of correlation between the two variables could be measured using the correlation coefficient.

In some cases, one variable (known, called independent variable) causes the change in the other variable (unknown, called dependent variable), e.g., change in height (independent) of the child causes change in weight (dependent variable) of the child.

This change of dependent variable with respect to change in independent variable is known as "regression."

Regression Coefficient

Regression means change in the measurements of a variable, on the positive or negative side, beyond the mean. This regression is measured through "regression coefficient," a constant, which gives the amount of increase in the dependent variable (e.g., weight) for a unit increase in the independent (e.g., height) variable. This quantity is represented by "b," when these variables are plotted on a graph, gives "regression line," depicting that the relationship is linear.

Regression Equation

The relationship between the two variables can be represented by a mathematical equation, known as "regression equation."

When the relationship is linear, the equation will be of the type y = a + bx

where y is the dependent variable, x is the independent variable, b is the regression coefficient, and a represents the value of y for x = 0.

The regression coefficient is calculated by using the formula

$$b_{xy} = \frac{\Sigma xy - (\Sigma x \Sigma y)/n}{\Sigma x^2 - (\Sigma x)^2/n}$$

where b_{xy} is regression coefficient of independent variable *x*, upon dependent variable *y* and usually represented as *b*.

Nonlinear Regression

Sometimes, it may happen that the dependent variable increases for each change of the independent variable up to a certain value and then it may start decreasing with the

JAYPEE BROTHERS

change in the independent variable. The reverse process may also happen. In such cases, the regression is said to be "nonlinear." For example, the prevalence rate of tuberculosis in a community rises with the age up to a certain age and then starts decreasing with the increase in the age after reaching a maximum at certain age-group. This may be due to immunity attained by the community by the time they reach a certain age.

In such a case, the relationship is said to be quadratic.

Multiple Correlation and Regression

In simple correlation, it is observed that there is one dependent variable and only one independent variable. But it may also happen that there may be more than one independent variable for the causation of the dependent variable. For example, weight of the newborn baby (dependent variable) is related to more than one independent variable, such as age of the mother, period of gestation at the time of birth, parity of delivery, health condition of the mother, and so on. If the overall correlation of all these independent factors with the dependent factor is considered, then it is known as "multiple correlation."

If an equation is fitted to estimate the dependent variable (e.g., weight of the baby at birth) in terms of other independent variables (as in above factors), then the equation is called as the "multiple regression equation."

LIFE TABLE

It is a table showing the mortality experience of a group of population of a country, at different ages, when followed every year, starting from their birth until all are dead. Thus it explains the outcome of a specified population in terms of number of deaths, number of surviving at each age, number of person years lived, and the average longevity of these persons. In other words, it helps to derive at the expectation of life at particular age, e.g., expectation of life at birth, at age of 1, 2, 5, 10, 60 years, etc. Expectation of life is defined as the average number of years a person is expected to live after attaining a specified age, exposing/experiencing the current risk of mortality. Thus life table is a "biometer" of the population.

The life tables are constructed once in 10 years, after each census separately for males and females. Usually a specified population of a standard number of 100,000 newborn babies, of the same sex, born at the same time are followed through various ages.

To construct a life table, two things are required:

1. Population living at all individual ages in a selected year.
2. Number of deaths occurring in these ages.

The group with which it is started (cohort group) is known as "radix" of the life table. When mortality is studied in that group, it is seen that there is reduction in the radix group as age advances. The reduction is high in the early years of life, gradually decreases as age advances till middle age is reached, and again increases in older ages because of increased mortality in extremes of ages.

TABLE 26.11: Life table for the females in India (2000) (abridged version).

Age X	l_x	q_x	d_x	L_x	e_x^0
<1	100,000	0.0638	6,377	94,261	62.7
1–4	93,623	0.0376	3,525	366,035	65.9
5–9	90,099	0.0125	1,128	447,675	64.5
10–14	88,971	0.0066	588	443,384	60.2
15–19	88,383	0.0103	907	439,645	55.6
20–24	87,475	0.0139	1,213	434,345	51.2
25–29	86,263	0.0145	1,247	428,194	46.9
30–34	85,015	0.0151	1,281	421,874	42.5
35–39	83,735	0.0159	1,334	415,339	38.1
40–44	82,401	0.0222	1,883	407,421	33.7
45–49	80,568	0.0320	2,577	396,396	29.4
50–54	77,991	0.0440	3,432	381,374	25.3
55–59	74,559	0.0673	5,020	360,243	21.4
60–64	69,539	0.0998	6,937	330,350	17.7
65–69	62,601	0.1514	9,475	289,318	14.4
70–74	53,126	0.2173	11,545	236,767	11.5
75–79	41,581	0.3075	12,786	175,938	9.0
80–84	28,795	0.4262	12,273	113,290	6.9
85–89	16,522	0.5698	9,441	59,080	5.2
90–94	7,110	0.6824	4,852	20,996	3.9
95–99	2,258	0.8300	1,875	5,669	2.9
100+	384	1	384	823	2.1

An extract of All India Life Table for females, obtained from WHO world mortality 2002, is depicted in **Table 26.11**.

Description

There are different rows and columns in the life table. Each row refers to particular age. Columns are described below.

Column 1 (Age X)

This gives the age-group, i.e., the exact number of completed years since birth. It is denoted by X. A detailed life table shows the age at every year (no frequency interval), and an abridged life table shows the age-group intervals (as shown in the table). The first one is <1 year, next is 1–4 years, because of different mortality rates in these two groups, and then subsequently the group intervals are of 5 years, the last one being 100 and above.

Column 2 (l_x)

The figures in this column show the number of persons surviving at age x out of the total number of persons started at age 0 (at birth). At birth, it is 100,000 newborn babies, and at 1–4 years it is 93,623, and at 5–9 years 90,099, and so on. It is represented as l_x.

JAYPEE BROTHERS

Column 3 (q_x)

This gives the probability of dying as experienced by the cohort group 1,00,000 at different ages, i.e., the age-specific death rate experienced by the cohort group, denoted as q_x. In the above table, experiencing the infant mortality rate of 63/1,000 live births (i.e., 6,377/100,000 live births), the number of deaths is 6,377 and the survivors are 100,000 – 6,377 = 93,623. These survivors experience the age-specific death rate during 1–4 years. A total of 3,524 (shown as 3,525 in the table) would die leaving behind 90,099 as survivors (l_x). This goes on till all are dead.

Column 4 (d_x)

This column shows the absolute number of deaths occurring in the different age-groups. It is arrived at by applying the current age-specific death rate to the cohort group in the respective age-groups. In the above table, it is 6,377 during infancy, 3,525 during 1–4 years, and so on. It is denoted as d_x.

Column 5 (L_x)

This gives the number of years survived by 93,623 individuals. Since all 6,377 infants have not died at one point of time, probably each one might have survived on an average of 3 months, contributing to survival period of 638 years, thus the number of years survived will be more than the actual number of survivors. 93,623 + 638 = 94,261, denoted as L_x.

Column 6 (e^0_x) (Life Expectancy)

This column shows the expectation of life at the start of a particular age-group. This is denoted by e^0_x. The expectation of life at 1 year of age is higher than the expectation of life at birth, because they have passed at the risk of infant rate. This is obtained by dividing the total number of years lived at age X by the number of starters/cohort group.

Expectation of life at birth =

$$e^0_x = \frac{\Sigma L_x}{l_x} = \frac{\text{Total number of years lived}}{\text{Total number of persons living at birth}}$$

$$= \frac{62,68,417}{1,00,000} = 62.7 \text{ years}$$

$$\text{Expectation of life at age 1} = \frac{\Sigma L_x - L_x}{l_x}$$

$$= \frac{62,68,417 - 94,261}{93,623}$$

$$= 65.9 \text{ years}$$

Uses of Life Table

- It helps to estimate the life expectancy of a country, which is a useful indicator to assess the health status which, in turn, is used for international comparison also
- Expectation of life at insured age or at retirement age helps in budgeting for social security, such as life insurance policies, pension benefits, etc.
- Probability of survivors at different ages helps to plan services, such as school health services for children, services for adolescent girls, etc.
- Modified life table method helps to know the survival rates after the treatment of chronic diseases, such as tuberculosis, cancer, diabetes, hypertension, coronary heart disease, etc.

META-ANALYSIS

It is a quantitative statistical analysis of several separate but similar independent experiments (or studies), in order to test the pooled data for statistical significance (i.e., to test the effectiveness of results). When the treatment effect (or effect size) is consistent from one study to the next, meta-analysis is used to identify the common effect. When the treatment effect varies from one study to the next, meta-analysis is used to identify the reason for variation, thus helps to resolve uncertainty when the results disagree. Pharmaceutical companies use meta-analysis to gain approval for new drugs.

The basic tenet behind meta-analysis is that there is a common truth behind all conceptually similar scientific studies, but which has been measured with a certain error within individual studies. The aim then is to use approaches from statistics to derive a pooled estimate closest to the unknown common truth based on how this error is perceived in essence, all existing methods yield a weighted average form the results of individual studies and what differs is the manner in which these weights are allocated and also the manner to which the uncertainty is computed around the point estimate thus generated. In addition to providing an estimate of the unknown common truth, meta-analysis has the capacity to convert results from different studies and identify patterns among study results, sources of disagreement among those results, or other interesting relationships that may come to light in the context of multiple studies.

Steps in a Meta-analysis

- Formulation of the problem
- Search of literature
- Selection of studies (incorporation criteria):
  - Based on quality criteria, e.g., the requirement of randomization and blinding in a clinical trial
  - Selection of specific studies on well-specified subject e.g., the treatment of breast cancer
  - Decide whether unpublished studies are included to avoid publication bias.
- Decide which dependent variables or summary measures are allowed, e.g., discrete data or continuous data
- Selection of meta-regression statistical model, e.g., simple regression, fixed effect meta-regression or random effect meta-regression.

JAYPEE BROTHERS

Advantages

- Results can be generalized to a large population
- The accuracy of the estimate can be improved as more data is used. This in turn may increase the statistical power to detect an effect
- Inconsistency of results across studies can be quantified and analyzed
- Hypothesis testing can be applied on summary estimates
- The presence of publication bias can be investigated
- Moderators can be included to explain variation between studies.

RESEARCH METHODOLOGY

Research is a scientific, organized and systematic way of search for relevant, new fact/information on specific topic. It is an art of scientific investigation. It is a movement from unknown to the known. In other words, it is a voyage of discovery. Research is carried out for two purposes; one, for the discovery of new facts and the second one, for the verification of old facts. Research is organized because it is well planned and systematic because of definite procedure, to find answers to questions.

Research methodology is a logical, systematic way of explaining how to carry out a research. Research methodology (RM) ensures reliable, valid results that addresses the aim and objectives of the research question. RM encompasses what data is going to be collected, where from collected and how it is being collected, analyzed and interpreted. It is an intellectual activity involving techniques.

Importance of Research Methodology

- RM provides scientifically sound findings.
- It helps the researcher in making the process smooth, effective and manageable.
- It allows the reader to understand the approach and the methods used to reach conclusions.

Benefits of Research Methodology

- Other researcher who wants to replicate the research, have enough information to do so.
- Researchers who receive criticism can refer to the methodology and explain their approach.
- It can help, provide researcher with a specific plan to follow throughout the research.
- The methodology design process helps researcher select the correct method for the objectives.
- RM allows researcher to document what they intend to achieve with the research, from the outset.

Types of Research Methodologies and their Applications

There are three types of RM. Best is selected based on the nature of research and practicalities. They are qualitative, quantitative and combination of both (i.e., mixed method).

1. **Qualitative research:** This involves exploring the information by collecting and analyzing the non-numerical data, such as written or spoken words, descriptions, textual data, concepts, beliefs, ideas and such other intangibles. Qualitative data (non-numerical) is collected through interviews, observation and focus groups, using a few carefully chosen participants.

   The qualitative methodology is subjective and more time consuming. The qualitative methodology is used when aims and objectives are exploratory or confirmatory in nature. So it is also called "exploratory research". Different types of qualitative researches are case study, ethnographic method, grounded theory method, phenomenological method.
2. **Quantitative research:** This method invites using numbers to collect the data. It measures relation between two variables (e.g., personality and committing crime) or test a set of hypothesis. It verifies the existing theories or hypothesis. This method is adapted when the objective of the research is to confirm something. It focuses on collecting, testing and measuring numerical data, usually from a large number of participants. The data is then analyzed using statistical analysis and comparisons. Popular methods used to gather quantitative data are surveys, questionnaires, tests, databases and organizational records.

   The quantitative research method is objective and is often quicker as researcher use software programs from analyzing the data. It confirms the potential theory or hypothesis by empirical testing. So quantitative research is also called "Confirmatory research" or "empirical research". Different types of quantitative research are descriptive research, experimental research, quasi—experimental and correlation research.
3. **Mixed method:** This method attempts to bring both qualitative and quantitative research. Qualitative research explores the situation and develops conceptual framework and then use quantitative method to test that model empirically. Thus this method provides additional perspectives, creates a richer picture and presents multiple findings. Types of sampling designs and calculation of sample size (already described):

Data Collection Methods

1. **Interview method:** In this method, the interviewer collects the data by personal interview, face to face, through phone, e-mail, etc. It is a standardized technique.

   Merits: The data is accurate, original, captures verbal and nonverbal ques and the interviewee is kept focused.

   Demerits: It is time consuming and expensive. The supervisor has to do complex work of selecting, training and supervising the staff.
2. **Focus groups method:** It is a group interview involving six to twelve persons, who share common interest.

JAYPEE BROTHERS

The facilitator guides the group on a particular topic and encourages them to share their perception. Focus groups are qualitative data collection method, means the data is descriptive and cannot be measured numerically.

Merits: It is quick and easy to set up. It provides useful information that individual data collection does not provide. It helps to gain insight into a topic, which is not obtained in other method.

Demerits: The discussion can be side tracked by someone. Analysis of the data is time consuming. Does not provide valid information at the individual level. The information is not representative of other groups.

3. **Observation methods:** It is systematic viewing, coupled with consideration of seeing phenomenon viz., observing keenly the practices and behaviors of target audience.

   Merits: No subjective bias. Researcher gets current information.

   Demerits: Expensive, time consuming and limited information.. Unforeseen factors may interfere with observation tasks.

   (***Note:*** If the research is exploratory or qualitative in nature, data is collected through interview, focus groups and observation methods).
4. **Survey method (questionnaires method):** This is quite popular method. Questionnaires are prepared first and distributed or mailed to respondents, who are expected to read and understand the questions and write down the answers in the space provided. At the end, the surveyor collects them and stores for analysis. This method is preferred for large samples.

   Merits: Low cost method, answers are free from bias, and reliable results. Questionnaires can be restructured.

   Demerits: Respondents must be educated to read and understand the questions. Questions should be clear, simple, easily understandable and the sequence of questions should be smooth. The type of questions has to be determined whether the questions should be direct, indirect or open form or closed form, dichotomous (yes/no) and multiple choice questions. It is a slow method and there is no personal contact. Incomplete and wrong information may be obtained.

   (***Note:*** If the research aims to measure specific variables or test hypothesis, large scale surveys are better).

Data Collection Tool

Questionnaire is a data collection tool. No survey can achieve success without a well designed questionnaire. Designing a questionnaire is more of an art than a science. Design of a questionnaire depends upon whether the researcher wishes to collect qualitative or quantitative information. For collection of qualitative information (for the purpose of formation of hypothesis) no formal questionnaire is needed. For collection of quantitative information (for the purpose of testing hypothesis), a formal standardized questionnaire has to be designed, so that data is analyzed statistically.

Criteria of Standard Questionnaire

- Title of the project work should be on top, so that the respondent should understand the nature of the project.
- Instructions to the respondents.
- Questions should be simple, clear, easy and brief.
- Questions should provide complete and accurate information.
- It should meet the research objective.
- Questions should be in a definite order, providing continuity of thought.
- Questions should be unbiased.
- Questions should retain the interest among the respondents throughout the interview.

Type of Questions

Two types of questions are commonly used.

1. **Open ended questions:** In this type, questions are such that there is enough space kept open for the respondents to answer.
2. **Close ended questions:** In this type, each question is followed by multiple answers, including one correct answer. The respondents only have to tick ✓ mark the correct answer, whichever he/she feels correct.

At the end of the questionnaire, there must be a thankful note.

Data presentation: The raw data collected from different sources and methods, is rearranged in a meaningful and understandable way—as tables, charts, diagrams, pictures and graphs.

Data analysis: Analysis of the data depends upon whether the data is qualitative or quantitative.

- **Qualitative data analysis:** Common methods are:
  - *Content analysis:* This is used to analyze documented information, as in interviews responses, i.e., tracking the meaning of words or phrases.
  - *Narrative analysis:* This is used to analyze content from several sources like interviews, observations or surveys.
  - *Discourse analysis:* This method analyzes spoken words or written language. This helps to study how communication works in social contexts.
  - *Grounded theory:* This is used to discover or construct a theory explaining why something happened. It uses a comparative analysis of data from similar cases in different settings to derive explanations.
  - *Thematic analysis:* Closely examining the data to identify the main themes.

JAYPEE BROTHERS

- **Quantitative data analysis:** This involves turning the numbers into meaningful data by applying rational and critical thinking. Before the data is analyzed by software, the data has to be edited, validated and coded. Then reported in graphs and tables. Commonly used methods for quantitative data analysis are
  - Descriptive analysis, wherein descriptive statistics like mean, median, mode, percentage, frequency and range is used to find the pattern.
  - Inferential analysis, wherein relationship between multiple variables is shown using correlation between variables, regression and variance analysis.

  Applications used are Excel, R or SPSS (Statistical Package in Social Sciences).

Factors to Consider While Choosing Research Methodology

- Nature of the research, aims, objectives and research questions.
- Data collection
- Sample size
- Availability of the time and cost factor.

Steps to Follow While Choosing the Best Research Methodology

Step 1: Define the goals, objectives and research questions. It is essential to understand what research has to be done and what are the variables.

Step 2: Pertinent research undertaken by other researchers from the same field is referred. The method is then evaluated based on feasibility and limitations.

Step 3: Structuring the plan and finding resources.

Step 4: Writing the research methodology in detail and reviewed. This helps to know the hurdles of the research and how to overcome it.

Example for quantitative research: A survey of 300 students of an university, is done by asking question, "On a scale of 1 to 5, how much have to be satisfied with your professors?" Statistical analysis of the data is performed. The result would be "on average, the students rated their professors 4.4".

Example for qualitative research: Only 15 students are conducted an in-depth interview with open ended questions, such as "How much have you satisfied with your students and what can be done to improve the study program?" Based on answers obtained, follow-up questions can be asked to clarify things. All interviews are transcribed using transcription software and try to find commonalities and patterns.

Mixed method approach: In this method, interview is conducted to find out how much are they satisfied with the studies with open ended questions. On analysis new insights are gained. Later, survey is done on a large scale to test these insights.

Qualitative versus Quantitative Research Methods (Table 26.12)

TABLE 26.12: Difference between qualitative and quantitative research methods.

Qualitative research	*Quantitative research*
It is descriptive study	Nondescriptive study
Involves collecting non-numerical (verbal) data, e.g., spoken words, textual data, taste, smell, color, etc	Involves collecting numerical data, e.g., numbers, age, height, salary, shoe size, etc.
Data is collected through interviews, observations, focus groups	Data is collected through surveys, experiments, records
Objectives are exploratory in nature	Objectives are confirmatory in nature
Often called "exploratory research"	Often called "confirmatory/ empirical research"
Focusses on formulating hypothesis	Focusses on testing hypothesis
Method is subjective	Method is objective
Consumes more time	Consumes less time
Analysis of data is by summarizing, categorizing and interpretation	Analysis is through mathematics and statistical analysis
Common methods are content analysis, discourse analysis and thematic analysis	Common methods are descriptive analysis and inferential analysis
Requires small sample size.	Requires large sample size.
Questionnaires are open-ended	Questionnaires are closed ended (MCQs)
Based on variables	Based on numbers
Reasoning is inductive	Reasoning is deductive

FORMULATION OF A RESEARCH QUESTION

Choosing a research question (RQ) is an essential element of both qualitative and quantitative research methods. It is essential before starting any research. It aims to explore an existing uncertainty in an area of concern and points to a need for deliberate investigation. Therefore, it is pertinent to formulate a good RQ essential to guide the research project work. It pinpoints exactly what is to be found out and gives the work a clear focus and purpose.

To develop a RQ one needs to begin by identifying the subject of interest and then do preliminary research on that subject. Then the researcher defines what needs to be known in that particular subject and assesses the implied questions, frames the RQ by narrowing the focus and scope of the research subject and then evaluates the research question. Thus formulation of a research question is a first critical step in research process. It is a systematic process and has to be performed meticulously as it leads to formulation, protect, restore and maintain health.

A good RQ forms the backbone of a good research. It identifies the problem to be studied and leads to formulation of an appropriate hypothesis.

Characteristics of a good research question: The characteristics are expressed as an acronym "FINERMAPS".

JAYPEE BROTHERS

- **F**easible—to carry out the research within the ability of the investigator and also within the time frame and practical constraints like funds.
- **I**nteresting—to motivate the researcher to keep going with project work.
- **N**ovel—means the RQ should aim at establishing new facts.
- **E**thical—means the research work should get clearance from the "Ethical Committee" of the institution before starting the work. The Committee will grant permission only after fulfilling the criteria, such as minimizing the risk of harm to the participants of research, protecting the privacy, maintaining the confidentiality, and provide the participants right to withdraw from research.
- **R**elevant—to the academic and intellectual interest.
- **M**anageable—by the researcher.
- **A**ppropriate—logically and scientifically.
- **P**otential value—means the RQ should address a topic having clear implications for resolving important dilemmas in health and healthcare decisions.
- **S**pecific—enough to answer thoroughly.

Research plays an important role in developing clinical practices and instituting new health policies.

ANNEXURE I

Table of Unit Normal Distribution

Normal distribution (single-tail). Proportion of area lying to right of ordinate through $Z = \pm(X - \bar{X})/SD$

A	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*	*0.06*	*0.07*	*0.08*	*0.09*
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	4.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
0.1	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.151	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0637
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0022	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.00047	0.0045	0.044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	2.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010

If mean height *m* is 160 cm, SD is 4 cm and height *X* is 166 cm *Z* will be $\frac{X - m}{SD} = \frac{166 - 160}{4} = 1.5$ corresponding to *Z*, 1.5 proportion of area comes to 0.0668 or 6.68% heights exceed 166 cm.

Source: Mahajan BK. Methods in Biostatistics, 3rd edition, 1981.

JAYPEE BROTHERS

ANNEXURE II

Table of Random Numbers

53 74 23 99 67	61 32 28 69 84	94 62 67 86 24	98 33 41 19 95	47 53 53 38 69
63 38 06 86 54	99 00 65 26 94	02 82 90 23 07	79 62 67 80 60	75 91 12 81 19
35 30 58 21 46	06 72 17 10 94	25 21 31 75 96	49 28 24 00 49	55 65 79 78 07
63 43 36 82 69	65 51 18 37 88	61 38 44 12 45	32 92 85 88 65	54 34 81 85 35
98 25 37 55 26	01 91 82 81 46	74 71 12 94 97	24 02 71 37 07	03 92 18 66 75
02 63 21 17 69	71 50 80 89 56	38 15 70 11 48	43 40 45 86 98	00 83 26 91 03
64 55 22 21 82	48 22 28 06 00	61 54 13 43 91	82 78 12 23 29	06 66 24 12 07
85 07 26 13 89	01 10 07 82 04	59 63 69 36 03	69 11 15 83 80	13 29 54 19 28
58 54 16 24 15	51 54 44 82 00	62 61 65 04 69	38 18 65 18 97	85 72 13 49 21
34 85 27 84 87	61 48 64 56 26	90 18 48 13 26	37 70 15 42 57	65 65 80 39 07
03 92 18 27 46	57 99 16 96 56	30 33 72 85 22	84 64 38 56 98	99 01 30 98 64
62 95 30 27 59	37 75 41 66 48	86 97 80 61 45	23 53 04 01 63	45 76 08 64 27
08 45 93 15 22	60 21 75 46 91	98 77 27 85 42	28 88 61 08 84	69 62 03 42 73
07 08 55 18 40	45 44 75 13 90	24 94 96 61 02	57 55 66 83 15	73 42 37 11 61
01 85 89 95 66	51 10 19 34 88	15 84 97 19 75	12 76 39 43 78	64 63 91 08 25
72 84 71 14 35	19 11 58 49 26	50 11 17 17 76	86 31 57 20 18	95 60 78 46 75
88 78 28 16 84	13 52 53 94 53	75 45 69 30 96	73 89 65 70 31	99 17 43 48 76
45 17 75 65 57	28 40 19 72 12	25 12 74 75 67	60 40 60 81 19	24 62 01 61 16
96 76 28 12 54	22 01 11 94 25	71 96 16 16 88	68 64 36 74 45	19 59 50 88 92
43 31 67 72 30	24 02 94 08 63	38 32 36 66 02	69 36 38 25 39	48 03 45 15 22
50 44 66 44 21	66 06 58 05 62	68 15 54 35 02	42 35 48 96 32	14 52 41 52 48
22 66 22 15 86	26 63 75 41 99	58 42 36 72 24	58 37 52 18 51	03 37 18 39 11
96 24 40 14 51	23 22 30 88 57	95 67 47 29 83	94 69 40 06 07	18 16 36 78 86
31 73 91 61 19	60 20 72 93 48	98 57 07 23 69	65 95 39 69 58	56 80 30 19 44
78 60 73 99 84	43 89 94 36 45	56 69 47 07 41	90 22 91 07 12	78 35 34 08 72
84 37 90 61 56	70 10 23 98 05	85 11 34 76 60	76 48 45 34 60	01 64 18 39 96
36 67 10 08 23	98 93 35 08 86	99 29 76 29 81	33 34 91 58 93	63 14 52 32 52
07 28 59 07 48	89 64 58 89 75	83 85 62 27 89	30 14 78 56 27	86 63 59 80 02
10 15 83 87 60	79 24 31 66 56	21 48 24 93 91	98 94 95 49 01	47 50 38 60
55 19 68 97 65	03 73 58 16 50	00 53 55 90 27	33 42 29 38 87	22 13 68 83 34
53 81 29 13 39	35 01 20 71 34	62 33 74 82 14	53 73 19 09 03	56 54 29 56 93
51 86 32 68 92	33 98 74 66 99	40 14 71 94 58	45 94 19 38 81	14 44 99 81 07
35 91 70 29 13	80 03 54 07 27	96 94 78 32 66	50 95 52 74 33	13 80 55 62 54
37 71 67 95 13	20 02 44 95 94	64 85 04 05 72	01 32 90 76 14	53 89 74 60 41
93 66 13 83 27	92 79 64 64 72	28 54 96 53 84	48 14 52 98 94	56 07 93 89 30
02 96 08 45 65	13 05 00 41 84	93 07 54 72 59	21 45 57 09 77	19 48 56 27 44
49 83 43 48 35	82 88 33 69 96	72 36 04 19 76	47 45 15 18 60	82 11 08 95 97
84 60 71 62 46	40 80 81 30 37	34 39 23 05 38	25 15 35 71 30	88 12 57 21 77
18 17 30 88 71	44 91 14 88 47	89 23 30 63 15	56 34 20 47 89	99 82 93 24 98
79 69 10 61 78	71 32 76 95 62	87 00 22 58 40	92 54 01 75 25	43 11 71 99 31
75 93 36 57 83	56 20 14 82 11	74 21 97 90 65	96 42 68 63 86	74 54 13 26 94
38 30 92 29 03	06 28 81 39 38	62 25 06 84 63	61 29 08 93 67	04 32 92 08 00
51 29 50 10 34	31 57 75 95 80	51 97 02 74 77	76 15 48 49 44	18 55 63 77 09
21 31 38 86 24	37 79 81 53 74	73 24 16 10 33	52 83 90 94 76	70 47 14 54 36
29 01 23 87 88	58 02 39 37 67	42 10 14 20 92	16 55 23 42 45	54 96 09 11 06
95 33 95 22 00	18 74 00 18 38	79 58 69 32 81	76 80 26 92 82	80 84 25 39
90 84 60 79 80	24 36 59 87 38	82 07 53 89 35	96 35 23 79 18	15 98 90 07 35
46 40 62 98 82	54 97 20 56 95	15 74 80 08 32	16 46 70 50 80	67 72 16 42 79
30 31 89 03 43	38 46 82 68 72	32 14 82 99 70	80 60 47 18 97	63 49 30 21 30
71 59 73 05 50	08 22 23 71 77	91 01 93 20 49	82 96 59 26 94	66 39 67 98 60

Source: Rao NSN. Elements of Health Statistics, 1st edition. Reprint 1985.

JAYPEE BROTHERS

ANNEXURE III

Table of χ^2 Distribution

DF	*0.99*	*0.98*	*0.95*	*0.90*	*0.80*	*0.70*	*0.50*	*0.30*	*0.20*	*0.10*	*0.05*	*0.02*	*0.01*	*0.001*
1.	0.0157	0.0628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2.	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3.	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.266
4.	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.467
5.	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.515
6.	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7.	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8.	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9.	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10.	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11.	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12.	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13.	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14.	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15.	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16.	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17.	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18.	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19.	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20.	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315
21.	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	36.932	46.797
22.	9.542	10.600	12.338	14.041	16.314	18.10	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23.	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24.	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25.	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26.	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27.	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28.	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.893
29.	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.302
30.	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703
32.	16.362	17.783	20.072	22.271	25.271	27.373	31.336	35.665	38.466	42.585	46.194	50.487	53.486	62.487
34.	17.780	19.275	21.664	23.952	26.938	29.242	33.336	37.795	40.676	44.903	48.602	52.995	56.061	65.247
36.	19.233	20.783	23.269	25.643	28.735	31.115	35.336	39.922	42.879	47.212	50.999	55.489	58.619	67.985
38.	20.691	22.304	24.884	27.343	30.537	32.992	37.335	42.045	45.076	49.513	53.384	57.969	61.162	70.703
40.	22.164	23.838	26.509	29.051	32.345	34.872	39.335	44.165	47.269	51.805	55.759	60.436	63.691	73.402
42.	23.650	25.383	28.144	30.765	34.157	36.755	41.335	46.282	49.456	54.090	58.124	62.892	66.206	76.084
44.	25.148	26.939	29.787	32.487	35.974	38.641	43.335	48.396	51.639	56.369	60.481	65.337	68.710	78.750
46.	26.657	28.504	31.439	34.215	37.795	40.529	45.335	50.507	53.818	58.641	62.830	67.771	71.201	81.400
48.	28.177	30.080	33.098	35.949	39.621	42.420	47.335	52.616	55.993	60.907	65.171	70.197	73.683	84.037
50.	29.707	31.664	34.764	37.689	41.449	44.313	49.335	54.723	58.164	63.167	67.505	72.613	76.154	86.661
52.	31.246	33.256	36.437	39.433	43.281	46.209	51.335	56.827	60.332	65.422	69.832	75.021	78.616	89.272

Contd...

JAYPEE BROTHERS

Contd...

DF	0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01	0.001
54.	32.793	34.856	38.116	41.183	45.117	48.106	53.335	58.930	62.496	67.673	72.153	77.422	81.069	91.872
56.	34.350	36.464	39.801	42.937	46.955	50.005	55.335	61.031	64.658	69.919	74.468	74.815	83.513	94.461
58.	35.913	38.078	41.492	44.696	48.797	51.906	57.335	63.129	66.816	72.160	76.778	82.201	85.950	97.039
60.	37.485	39.699	43.188	46.459	50.641	53.809	59.335	65.227	68.972	74.397	79.082	84.580	88.379	99.607
62.	39.063	41.327	44.889	48.226	52.487	55.714	61.335	67.322	71.125	76.630	81.381	86.953	90.802	102.166
64.	40.649	42.960	46.595	49.996	54.336	57.620	63.335	69.416	73.276	78.860	83.675	89.320	93.217	104.716
66.	42.240	44.599	48.305	51.770	56.188	59.188	59.527	65.335	71.508	75.424	81.965	91.681	95.626	107.258
68.	43.838	46.244	50.020	53.548	58.042	61.436	67.335	73.600	77.571	83.308	88.250	94.037	98.028	109.791
70.	45.442	47.893	51.739	55.329	59.898	63.346	69.334	75.689	79.715	85.527	90.531	96.388	100.425	112.317

For odd values of n between 30 and 70, the mean of the tabular values for $n - 1$ and $n + 1$ may be taken. For larger values of n, the expression—$2n - 1$ may be used as a normal deviate with unit variance, remembering that the probability for χ^2 corresponds with that of a single tail of the normal curve (for fuller formulae, see introduction).

Source: Rao NSN. Elements of Health Statistics, 1st edition. Reprint 1985.

ANNEXURE IV

Kendall's Table

The significance of the rank correlation coefficient is worked out as follows. For samples of size over 10 pairs, the significance level of r_{rank} can be calculated similar to that of r, but for samples of size 10 or less number of pairs, it is given by the following table given by Kendall.

Size of the sample	5% Level	1% Level
4 or less	None	None
5	1.000	None
6	0.886	1.000
7	0.750	0.893
8	0.714	0.857
9	0.683	0.883
10	0.648	0.794
11	Use tables of significance of *r*	

Source: Rao NSN. Elements of Health Statistics, 1st edition. Reprint 1985.

BIBLIOGRAPHY

1. (Brig) Sunderlal, Adarsh, Pankaj. Textbook of community medicine, 1st edition. 2007. CBS Publishers and Distributors Pvt. Ltd, New Delhi.
2. Alberti KP, Guthmann JP, Fermon F, Nargaye KD, Grais RF. Use of lot quality assurance sampling to estimate vaccination coverage helps guide future vaccination efforts. Trans R Soc Trop Med Hyg. 2008;102(3):251-4. doi:10.1016/J.trstmh.2007.10.015.
3. Data Collection Methods: Definition, Examples and Sources. @ https://www.questionpro.com > blog > data-collection.
4. Data Collection Methods for program Evaluation: Focus Groups @https://www.cdc.gov > brief.
5. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629-34.
6. Ghai OP, Gupta P. Essential preventive medicine, 1st edition. 1999. Vikas Publishing House Pvt. Ltd., New Delhi.
7. Jotform Data collection methods @https://www.jotform.com > data-collection-methods.
8. Kothari CR. Research methodology.
9. Mahajan BK. Methods in biostatistics, 3rd edition. 1981. Smt. Indu Mahajan, Pusa Road, New Delhi.
10. Park K. Park's textbook of preventive and social medicine, 18th edition. 2005. Banarasidas. Bhanot Publishers. Jabalpur.
11. Ratan SK, Anand T, Ratan J. Formulation of Research Question—Stepwise Approach. @https://www.ncbi.nim.nih.gov > articles
12. Rex G. Graphical display of estimates having differing standard errors. Technometrics. 1988;30(3):271-81.
13. What is Research Methodology? Why is it important and types. @https://www.indeed.com
14. Rao NSN. Elements of health statistics, 1st edition. 1985. Ratna Publications, Varanasi.

JAYPEE BROTHERS

for detailed information on community medicine books, visit our website www.ejaypee.com, for detailed information on community medicine books, visit our website w

COMMUNITY MEDICINE
with Recent Advances

Salient Features

- Cover page reminds the reader about some of the Stalwarts, who have contributed to the public health, in terms of lives that can be saved.
- The content has been thoroughly revised and updated as per the latest NMC curriculum and the national health programs.
- As per the title, the book encompasses the recent advances in the field of community medicine.
- The exquisite flowcharts, diagrams, pictures, and tables with lucid language make the reading very easy.
- Reading clarifies the doubts and confusions thereby makes the subject matter interesting, giving a pleasurable experience.
- Comprehensively covers the syllabus as recommended by Universities and National Medical Commission.
- It is not only student friendly but also teacher friendly.
- Useful for medical officers, health professionals, and policy makers.
- Gives the researchers a thought to work in terms of preventing the disease and promoting the health, thus serving the humanity.
- A complementary booklet is provided, containing a variety of new theory pattern-based questions. These include long/short answer, reasoning based questions, MCQs and scenario based questions to aid comprehensive preparation.

AH Suryakantha MD DHA has served in Basaveshwara Medical College and Hospital, Chitradurga, Karnataka, India. He is the first batch student (1965) of the prestigious institute JJM Medical College, Davangere, Karnataka, India and served in the Department of Community Medicine for 36 years, in the same institution where he studied, in various capacities, including 16 years as Head of Department. Later, he served for 2 years as Professor in SS Institute of Medical Sciences and Research Center, Davangere. After retirement, he has been serving the community as Family/Community Physician. He has updated this book using his four decades of teaching, training, and research experience. His scientific articles have been published in state and national journals. He has presented several papers at state and national conferences. He is deeply committed to promoting the specialty of Community Medicine.

diginerve®
A Jaypee Initiative
• Video Lectures • Dr. Wise AI Chatbot
• Question Bank • Notes
Download App

Buy from ejaypee

JAYPEE BROTHERS

Shelving Recommendation
COMMUNITY MEDICINE

ISBN 978-93-7806-415-9