Eastern Economy Edition

Fifth Edition

INTRODUCTION TO BIOSTATISTICS AND RESEARCH METHODS

P.S.S. Sundar Rao
J. Richard

PHI

Introduction to Biostatistics and Research Methods

Fifth Edition

P.S.S. Sundar Rao, Dr. PH.
Research Coordinator, The Leprosy Mission Trust India
New Delhi
Former Director, Schieffelin Leprosy Research and Training Centre
Karigiri, Tamil Nadu

J. Richard, Ph.D.
Consultant, Biostatistics and Research Methods
Former Professor and Head, Department of Biostatistics
Christian Medical College, Vellore

PHI Learning Private Limited
New Delhi-110001
2012

INTRODUCTION TO BIOSTATISTICS AND RESEARCH METHODS, Fifth Edition
P.S.S. Sundar Rao and J. Richard

© 2012 by PHI Learning Private Limited, New Delhi. All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publisher.

ISBN-978-81-203-4520-1

The export rights of this book are vested solely with the publisher.

Eighteenth Printing (Fifth Edition) ... ... January 2012

Published by Asoke K. Ghosh, PHI Learning Private Limited, M-97, Connaught Circus, New Delhi-110001 and Printed by Rajkamal Electric Press, Plot No. 2, Phase IV, HSIDC, Kundli-131028, Sonepat, Haryana.

Contents

Foreword *ix*

Preface *xi*

1. Scope of Statistical Methods in Medicine **1–6**

1.1 Introduction *1*
1.2 Estimation *2*
1.3 Tests of Hypothesis *3*
1.4 Study Design *5*
Questions *6*

2. Role of Statistics in Clinical Medicine **7–10**

2.1 Introduction *7*
2.2 Some Examples of Earlier Uses of Statistics *7*
2.3 Areas of Application of Statistics *8*
Questions *10*

3. Role of Statistics in Preventive Medicine **11–13**

3.1 Introduction *11*
3.2 Some Early Examples *11*
3.3 Areas of Application *11*
Questions *13*

4. Variables **14–19**

4.1 Introduction *14*
4.2 Definition of Variables *14*
4.3 Qualitative Variables *14*
4.4 Quantitative Variables *15*
4.5 Subjective and Objective Variables *16*
4.6 Dependent and Independent Variables *17*
4.7 Measurement of Variables *17*
4.8 Scale of Measurement *18*
4.9 Recording the Variables *18*
4.10 Role of Computers *18*
4.11 Quality of Data *18*
Questions *19*

5. Frequency Distribution **20–25**

5.1 Introduction *20*
5.2 Frequency Distributions *20*

5.3 Diagrams *21*
5.4 Characteristics of a Frequency Distribution *25*
Questions 25

6. Measures of Central Tendency and Location 26–34

6.1 Introduction *26*
6.2 Arithmetic Mean *26*
6.3 Median *28*
6.4 Mode *29*
6.5 Position of Averages *29*
6.6 Selection of the Appropriate Measure of Central Tendency *30*
6.7 Geometric Mean *31*
6.8 Harmonic Mean *32*
6.9 Percentile *33*
6.10 Use of Computer *34*
Questions 34

7. Measures of Dispersion 35–43

7.1 Introduction *35*
7.2 Range *35*
7.3 Interquartile Range *36*
7.4 Mean Deviation *37*
7.5 Variance and Standard Deviation *38*
7.6 Alternative Method of Calculating Standard Deviation *40*
7.7 Coefficient of Variation *42*
7.8 The Use of Computer *43*
Questions 43

8. Probability 44–54

8.1 Introduction *44*
8.2 The Probability Scale *44*
8.3 Measurement of Probability *45*
8.4 Laws of Probability for Independent Events *46*
8.5 Conditional Probability *46*
8.6 Bayes' Theorem *47*
8.7 Application of Bayes' Theorem in Determining Diagnostic Efficacy *49*
8.8 Other Applications of Probability *51*
Questions 54

9. Probability Distributions 55–61

9.1 Introduction *55*
9.2 Binomial Distribution *55*
9.3 Poisson Distribution *57*
9.4 Normal Distribution *58*
9.5 '*t*' Distribution *60*
9.6 Other Distributions 61
Questions 61

10. Sampling **62–70**

10.1 Introduction *62*
10.2 Definitions *62*
10.3 Types of Population *62*
10.4 Sample *63*
10.5 Sampling Variation and Bias *64*
10.6 Nonprobability Sampling Techniques *65*
10.7 Probability (Random) Sampling Techniques *65*
10.8 Listing of Population *69*
10.9 Sample Size *69*
Questions *70*

11. Basis of Statistical Inference **71–75**

11.1 Introduction *71*
11.2 Sampling Distribution *71*
11.3 Statistical Inference *72*
11.4 Type I and Type II Errors *74*
Questions *75*

12. Tests of Significance and Estimation **76–89**

12.1 Introduction *76*
12.2 Procedure *76*
12.3 For Large Samples: Examples *77*
12.4 For Small Samples: Examples *81*
12.5 Estimation: Examples for Large Samples *86*
12.6 Estimation: Examples for Small Samples *88*
12.7 Reading Computer Output *88*
Questions *89*

13. Linear Regression and Correlation **90–99**

13.1 Introduction *90*
13.2 Scatter Diagram *90*
13.3 Correlation and Regression *91*
13.4 Correlation Coefficient *92*
13.5 Regression Equation *95*
13.6 Restrictions *96*
13.7 Nonlinear Regression *97*
13.8 Multiple Regression *97*
13.9 Other Extensions *99*
Questions *99*

14. The Chi-Square Test **100–107**

14.1 Introduction *100*
14.2 The Formula for Chi-Square *100*
14.3 Distribution of Chi-Square (χ^2) and Degrees of Freedom *101*
14.4 Some Applications of Chi-Square Test *101*

14.5 Yates Correction *105*
14.6 Extension of 2 × 2 Table *105*
14.7 Precautions in Using the Chi-Square Test *106*
Questions 107

15. Analysis of Variance 108–113
15.1 Introduction *108*
15.2 The *F*-Test *108*
15.3 An Illustration of One-Way ANOVA *109*
15.4 Two-way ANOVA *111*
15.5 Assumptions of ANOVA *113*
Questions 113

16. Non-parametric or Distribution-free Statistical Tests 114–121
16.1 Introduction *114*
16.2 Advantages of Non-parametric Tests *114*
16.3 Disadvantages of Non-parametric Tests *115*
16.4 Some Non-parametric Tests *115*
16.5 Some Illustrations *116*
Questions 121

17. Vital Statistics 122–139
17.1 Introduction *122*
17.2 Uses of Vital Statistics *122*
17.3 Mechanism for Collection of Vital Statistics *123*
17.4 Basic Formulae for Calculation of Vital Statistics *127*
17.5 Mortality Rates *129*
17.6 Fertility Rates *135*
Questions 139

18. Health Statistics 140–152
18.1 Introduction *140*
18.2 Utilisation of the Basic Data *140*
18.3 Sources of Health Statistics *141*
18.4 Problems in the Collection of Sickness Data *142*
18.5 Measurement of Sickness *144*
18.6 Hospital Statistics *146*
18.7 International Classification of Diseases *147*
Questions 152

19. Standardised Death Rates and Life Tables 153–165
19.1 Introduction *153*
19.2 Adjusted or Standardised Rates *153*
19.3 Life Tables *159*
19.4 Cox Proportional Hazards Model (Cox Regression) *165*
Questions 165

20. Demography **166–175**

20.1 Introduction *166*
20.2 Population and Growth *166*
20.3 Age and Sex Composition *167*
20.4 Dependency Ratio *169*
20.5 Other Indices *169*
20.6 Fertility and Mortality *169*
20.7 Demographic Transition *171*
20.8 Population Estimation *172*
Questions 175

21. Introduction to Research Methods **176–188**

21.1 Introduction *176*
21.2 Research Question *176*
21.3 Literature Review *177*
21.4 Theoretical Framework or Model *177*
21.5 Research Protocol *177*
21.6 Formulation of Objectives and Research Hypothesis *179*
21.7 Population under Study *179*
21.8 Statistical Design *180*
21.9 Sample Size *182*
21.10 Methods of Data Collection *183*
21.11 Methods of Dealing with Non-response *185*
21.12 Data Processing *187*
21.13 The Constraints *187*
Questions 188

22. Interventional Studies **189–200**

22.1 Introduction *189*
22.2 General Principles of Experimental Designs *189*
22.3 Common Designs *192*
22.4 Determination of Sample Size *195*
22.5 Clinical Trials *195*
22.6 Therapeutic Trials *196*
22.7 Prophylactic Trials *198*
22.8 Ethical Considerations *199*
22.9 Community Trials *199*
22.10 Repeated Measurement Studies *200*
Questions 200

23. Sample Size Determination **201–207**

23.1 Introduction *201*
23.2 Specifications Needed to Calculate Sample Size *201*
23.3 Examples of Sample Size Estimation *201*
23.4 Useful Hints *207*
Questions 207

24. Observational Studies **208-218**

24.1 Introduction *208*
24.2 Cross-sectional Studies *209*
24.3 Prospective Studies *209*
24.4 Retrospective Studies *210*
24.5 Sources of Bias in Observational Studies *212*
24.6 Controlling for Bias *213*
24.7 Causality *213*
24.8 Relative Advantages and Disadvantages of the Prospective and Retrospective Methods *214*
24.9 Confidence Intervals (CI) of Odds Ratio (OR) *215*
24.10 Measures of Potential Impact *215*
24.11 Some Comments *218*
Questions 218

25. Data Processing **219–224**

25.1 Introduction *219*
25.2 Computer Hardware *219*
25.3 Computer Software *220*
25.4 Choosing a Computer System *222*
25.5 Role of Computer in Data Management *222*
25.6 Further Uses of Computers *222*
25.7 A Note of Caution *223*
25.8 Clues for using SPSS and Excel *224*
Questions 224

26. Presentation of Data **225–240**

26.1 Introduction *225*
26.2 Basic Principles *225*
26.3 Elements of a Table *226*
26.4 Classifications *226*
26.5 Presentation by Graphs and Diagrams *229*
26.6 Time Comparison *230*
26.7 Space and Attribute Comparisons *232*
26.8 Analysis of Time Series *234*
26.9 Special Diagrams for Clinical Research *237*
Questions 240

Appendices ***241–253***

Bibliography ***255–262***

Answers to Selected Questions ***263–264***

Index ***265–268***

Foreword

It is indeed my pleasure and privilege to write the Foreword for *Introduction to Biostatistics and Research Methods*.

Dr. P.S.S. Sundar Rao joined Christian Medical College (CMC), Vellore, when the Madras University introduced the study of Biostatistics as a part of the syllabus of Preventive and Social Medicine in 1957 for the M.B.B.S. students in all medical colleges of the University. During that time I was the Medical Superintendent of the CMC Hospital, and later from 1961 to 1970, I served as the Principal of the College and Head of the Department of Neurological Sciences. I had the privilege to work closely with Dr. Sundar Rao in my research studies, and in the Research Committee. He made a significant contribution to the development and understanding of Biostatistics. He and Dr. J. Richard had rendered considerable help and guided in several epidemiological studies and in the research activities in CMC departments. They had not only helped CMC, but on a national level, had been conducting Summer Courses—advanced studies for teachers and research workers. They had also started a regular M.Sc. (Biostatistics) course for the Madras University in their department.

This book contains all the essentials of Biostatistics, research methods, demography and health statistics that lay a good foundation for anyone interested in these subjects. All the examples are from the health field and are based on real experience. I am pleased to recommend this book for the medical and health personnel and for all those interested in Biostatistics.

Padma Bhushan **Dr. Jacob Chandy**
M.Sc. FRCS (C), FICS, FAMS, FIASC
Formerly, Professor and Head,
Department of Neurological Sciences, and
Principal, Christian Medical College, Vellore

Preface

Providing and applying valid evidence to improve the existing practice should be the hallmark of every professional. Statistics as a science has been responsible for accelerating progress in our knowledge of defining the correct methods of collecting, analyzing and interpreting data. In this book, we have attempted to introduce the basic principles of statistics as applied to medical and health-related sciences through simple, practical examples, avoiding as much of the statistical jargon as possible. The present edition is the outcome of nearly five decades of teaching this subject to medical and nursing students, at both undergraduate and postgraduate levels, as well as to students pursuing courses in other health and social sciences.

Biostatistics is a relatively new but rapidly expanding field, and this book introduces both basic and its newer applications. With the developments in computers, though the practice of data management and statistical analysis has become easier, it is also fraught with dangers of misuse unless guided by correct statistical advice. This book helps in this process, and devotes a few chapters to research methodology and determination of the best study design and sample sizes as well as appropriate statistical tests of significance, including non-parametric tests.

The Fifth Edition of this book is a result of the positive feedback we have received from the students, faculty and other readers about the book. While the new edition continues to retain most of the topics discussed in the previous editions; we have made additions and revisions wherever felt necessary based on the feedback received and the developments that have taken place in the intervening period. We have introduced a new chapter on *Sample Size Determination*, which will be of great benefit to researchers. Besides, we have incorporated, added several new sections, revised almost all chapters and provided new examples.

Chapters 1–7 discuss, with real-life examples, the usefulness, concepts, and computation of basic biostatistics, commonly known as descriptive statistics. Chapters 8–11 deal with inferential statistics, and provide the basis of biostatistical inference using the science of probability. A wide range of examples for testing a hypothesis and estimating confidence intervals are included in Chapter 12. Chapters 13–16 are devoted to explaining measurement of associations and several useful statistical techniques such as correlation and Chi square. In this era when computers are used to obtain these statistics, one need not do the calculations manually; however, a good investigator needs to grasp the concepts behind the arithmetic, and know how these figures are arrived at and the logic on which they are built up.

Chapters 17–20 provide the learner a background of demography, vital rates and health statistics, illustrating concepts with latest information on various rates and ratios, and on the recent Census 2011. The latest health statistical data are

given in Chapter 17. Research Methodology is explained step by step in Chapter 21, followed by Chapter 22 on the design and analyses of interventional studies, and Chapter 24, on the design and analyses of observational studies. A newly introduced Chapter 23, on Sample Size Determination will meet the needs of many researchers. It explains to the investigator the concepts involved in arriving at the sample size and also a quick solution to estimation of sample sizes. The section on determining the number needed to treat is retained in the new edition. Chapter 25 on data processing is thoroughly revised keeping in mind current needs of the students. Finally, Chapter 26 deals with a variety of ways in which data can be presented.

Several persons have helped us over the years in preparing this book, particularly the staff of the Biostatistics Department of Christian Medical College (CMC) Vellore where we worked. We are deeply grateful to them for their valuable contributions. We thank especially Mr. S. Muthurethinam, Department of Biostatistics at CMC, Vellore, for his help in revisions. We acknowledge the material cited from *Health Monitor*, a useful compendium published by the Foundation for Research in Health Systems, Ahmedabad.

Finally, we wish to express our immense gratitude to the staff of PHI Learning, particularly Mr. K.C. Devasia, Executive Editor, for providing with useful ideas for improvements, and Mr. Bijendra Pandey, Copy Editor, for copyediting the manuscript and the production team for their valuable help.

We invite constructive comments and suggestions from readers to further improve this book in future revisions. In the meanwhile we hope that the new edition will provide students with an enjoyable introduction to the world of biostatistics and that it will equip them with the tools to understand and apply the principles of statistics to medical and health-related areas.

Good luck and have fun!

P.S.S. Sundar Rao
J. Richard

CHAPTER

1 Scope of Statistical Methods in Medicine

1.1 INTRODUCTION

At the outset, it might be useful to point out two different meanings attached to the term 'statistics'. The term 'statistics' has been used to indicate facts and figures of any kind: health statistics, vital statistics or business statistics. It is also used to refer to a body of knowledge known as statistical methods developed for handling data in general, particularly in the fields of experimentation and research. The following essential features of statistics, as it is understood in the latter sense, are evident from the various definitions of statistics:

1. Principles and methods for the collection, presentation, analysis and interpretation of numerical data of different kinds —

(a) Observational data, quantitative data

(b) Data that have been obtained by a repetitive operation

(c) Data affected to a marked degree by a multiplicity of causes.

2. The science and art of dealing with variation in such a way as to obtain reliable results.

3. Controlled, objective methods whereby group trends are abstracted from observations on many separate individuals.

4. The science of experimentation which may be regarded as mathematics applied to observational data.

The term 'Biostatistics' can be understood to indicate the following:

1. Statistics arising out of biological sciences, particularly from the fields of medicine and public health.

2. The methods used in dealing with statistics in the fields of medicine, biology and public health, and in planning, conducting and analysing data which arise in investigations in these branches.

A glance at any medical journal shows the extensive use of statistical methods in the collection, evaluation and presentation of biological data. This is a reflection of a need for precise quantitative assessment of different types of phenomena which are encountered in this field. Further, these phenomena are affected by sampling and biological variations, and thus there arises the need for making valid inferences. Research workers, students and other professional workers in this field will need

to study data presented by others and evaluate the evidence found therein. Also, they may have to plan and evaluate their own work. Knowledge of the logical basis of the statistical approach and the rationale of the commonly used statistical techniques will be an essential prerequisite. Excerpts from various medical journals are given as illustrations.

1.2 ESTIMATION

1. We used multistage stratified random sampling methods to identify 96 urban and 267 rural primary sampling sites in four provinces of China; the sampling frame of 113 million individuals aged 18 years or older included 12% of the adult population in China. About 63,004 individuals, identified with simple random selection methods at the sampling sites, were screened with an expanded version of the General Health Questionnaire and 16,577 were administered a Chinese version of the Structured Clinical Interview for Diagnostic and Statistical Manual (DSM)-IV axis I disorders by a psychiatrist.

The adjusted 1-month prevalence of any mental disorder was 17.5% (95% CI 16.6–18.5). The prevalence of mood disorders was 6.1% (5.7–6.6), anxiety disorders was 5.6% (5.0–6.3), substance abuse disorders was 5.9% (5.3–6.5), and psychotic disorders was 1.0% (0.8–1.1). Mood disorders and anxiety disorders were more prevalent in women than in men, and in individuals 40 years and older than in those younger than 40 years.*

2. Patients with rheumatoid arthritis satisfying the 1987 modified American College of Rheumatology criteria were included. Besides demographic data, functional impact was assessed using the Health Assessment Questionnaire (HAQ). The psychosocial impact was measured using the medical outcomes study short form 36 (SF-36) with minor modifications. Data on direct and indirect health costs were collected by direct interview.

The mean age of 101 patients (90 women) was 43.2 years and mean duration of disease was 8 years. Their mean (SD) HAQ score was 0.97 (0.69) with 8 patients having scores >2. On the SF-36 scale (0–100) the mean (SD) score for various domains were: physical functioning 49.90 (28.55), social functioning 55.51 (20.59), role limitation due to physical problems 32.67 (41.34), role limitation due to emotional problems 47.54 (40.08), mental health 47.36 (7.99), general health perception 52.38 (8.30), energy and vitality 58.56 (6.09), and bodily pain 49.26 (18.87). The summary score for the physical component was 37.95 (9.03) and for the mental component it was 47.71 (4.81). While the physical component summary score had excellent negative correlation with the HAQ score ($r = -0.84$), correlation with the mental component summary score was poor ($r = 0.32$).**

---

* Michael R. Phillips, Jingxuan Zhang, Qichang Shi, Zhiqiang Song, Zhijie Ding, Shutao Pang, Xianyun Li, Yali Zhang, and Zhiqing Wang. Prevalence, treatment, and associated disability of mental disorders in four provinces in China during 2001–05: an epidemiological survey. The Lancet, **373**, 2041–2053, 2009.

** Aggarwal Amita, Shelly Chandran, Ramnath Misra. Physical, psychosocial and economic impact of rheumatoid arthritis: A pilot study of patients seen at a tertiary care referral centre. *Natl Med J India*, **19**, 187–191, 2006.

3. Hypokalaemic periodic paralysis constitutes a heterogeneous group of disorders that present with acute muscular weakness. From 1995 to 2001, 31 patients presented with periodic paralysis (mean age 34.5 years, range 11–68 years). Of the 31 patients, 19 were men.There were 13 patients (42%) with renal tubular acidosis, 13 with primary hyperaldosteronism (42%), 2 each with thyrotoxic periodic paralysis and sporadic periodic paralysis, and 1 with Gitelman syndrome. The patients diagnosed to have renal tubular acidosis had significantly lower serum bicarbonate (18.7[4.6]v.29.6[5.0] mEq/L; $p < 0.05$) and higher levels of chloride (107.5[6.0]v.99.5[3.4] mEq/L; $p < 0.05$) compared with those who had primary hyperaldosteronism, although the potassium values were similar (2.4[0.65]v.2.26[0.48]mEq/L; $p = 0.43$). All patients with primary hyperaldosteronism had hypertension at presentation and were proven to have adrenal adenomas.*

4. We carried out a quasi-experimental study involving comparison groups, and a time-series analysis from 1 January 2002 to 30 April 2008. The study group was composed of people injured while driving or riding a light motorcycle (engine capacity 51–125 cubic centimeters), while the comparison groups consisted of riders of heavy motorcycles (engine capacity > 125 cc), mopeds (engine capacity ≤ 50 cc) or cars who were injured in a collision within the city limits. The "intervention" was a law passed in October 2004 allowing car drivers to drive light motorcycles without taking a special driving test. To detect and quantify changes over time we used Poisson regression, with adjustments for trend and seasonality in road injuries and the existence of a driver' licence penalty point system.

The risk of injury among light motorcycle riders was greater after the law than before (relative risk, RR =1.46; 95% confidence interval, CI: 1.34–1.60). Although less markedly, after the law the risk of injury also increased among heavy motorcycle drivers (RR =1.15; 95% CI: 1.02–1.29) but remained unchanged among riders of mopeds (RR = 0.92; 95% CI: 0.83–1.01) and cars (RR = 1.06; 95% CI: 0.97–1.16).**

1.3 TESTS OF HYPOTHESIS

1. About 35 consecutive patients of RA, 18 patients suffering from OA and 18 age-matched healthy controls were included in the study. Levels of CTx-I in spot urine samples and serum IL-6 levels were measured using commercial ELISA type kits.

CTx-1 levels in RA patients were compared with that in OA ($P = 0.362$), with controls ($P < 0.001$); OA patients with controls ($P = 0.001$). IL-6 values were significantly different in RA patients compared with controls ($P < 0.001$), with OA patients ($P = 0.015$); in OA patients as compared to controls ($P = 0.008$). A positive correlation was found between CTx-1 and age ($P = 0.044$), DAS28 ($P = 0.007$), swollen joint count ($P = 0.006$) and tender joint count ($P = 0.006$). Levels did not correlate with

* Narsing Rao, John Mathew, Nihal Thomas, Simon Rajaratnam, Mandalam Subramaniam Seshadri. Aetiological, clinical and metabolic profile of hypokalaemic periodic paralysis in adults: A single-centre experience. *Natl Med J India*, **19**, 246–249, 2006.

** Katherine Perez, Marc Mari-Dell Olmo, Carme Borrell, Manel Nebot, Joan R. Villalbi, Elena Santamarina and Aurelio Tobias. Road injuries and relaxed licensing requirements for driving light motorcycles in Spain: a time-series analysis. *Bull World Health Organ*, **87**, 497–504, 2009.

ESR, CRP, PA factor and disease duration. IL-6 levels were found to have significant associations with age ($P = 0.012$), CRP ($P = 0.048$), DAS28 ($P = 0.024$), swollen joint count ($P = 0.001$) and tender joint count ($P = 0.04$). No correlations were found between IL-6 and ESR, RA factor, and disease duration. A positive correlation was also seen between CTx-I and IL-6($p = 0.03$, $r = 0.362$).*

2. BIA measurements and anthropometric measurements were carried out on 67 normal, healthy men between the ages of 18 and 45 years. True muscle mass was measured from 24 h creatinine excretion. Multiple linear regression with step-wise forward selection was used to predict total muscle mass using measurements like height^2/impedence, height and weight and using arm muscle area (AMA), thigh muscle area (TMA) and calf muscle area (CMA). The prediction equation for muscle mass (kg) using height^2/impedance and height was – 12.347 + (0.363 × height^2/impedance) + (0.122 × height) [$R^2 = 0.55$; Standard error of estimate (SEE) = 2.58 kg], while the equation using appendicular muscle area was 10.122 + (0.23 × AMA) + (0.049 × TMA) [$R^2 = 0.36$; SEE 3.07 kg].**

3. A total of 943 healthy urban and 205 rural adult subjects and 76 urban and 70 rural healthy children were studied for their dietary pattern, serum calcium, phosphorus, alkaline phosphatase, 25 (OH) D, and N-tact parathyroid hormone levels (N-tact PTH).

Dietary calcium and phosphorus were significantly lower ($P < 0.0001$) in both the rural adult and children compared to that of the urban adult and children. The dietary phytate to calcium ratio was significantly ($P < 0.0001$) higher in rural adult and children compared to that of urban adult and children. N-tech PTH levels correlated with 25 (OH)D in rural ($r = 0.24$; $P < 0.002$), in urban adult subjects ($r = 0.12$; $P < 0.0001$) and in rural and urban children ($r = –0.2$; $P < 0.05$). The 25 (OH)D levels of rural adult subjects were significantly higher ($P < 0.0001$) than that of urban adult subjects in both male and female groups. The 25(OH)D levels of both the urban and rural children were low.***

4. In three Russian industrial cities, about 60,416 residents who had died at ages 15–74 years in 1990–2001 were visited in 2001–05. Cases ($n = 43,082$) were those certified as dying from causes we judged beforehand might be substantially affected by alcohol or tobacco; controls were the other 5,475 decedents. Case versus control relative risks (RRs; calculated as odds ratios by confounder-adjusted logistic regression) were calculated.

In men, the three causes accounting for the most alcohol-associated deaths were accidents and violence (RR 5.94, 95% CI 5.35–6.59, in the highest consumption category), alcohol poisoning (21.68, 17.94–26.20), and acute ischaemic heart disease

* Shingada A.K. and L.S. Bichile, Assessment of disease activity in rheumatoid arthritis using urinary CTx-I levels as a marker of bone destruction and serum IL-6 as a marker of inflammation. *Indian J Med Res*, **127**, 6; 699, 2008.

** Rebecca Kuriyan, Tinku Thomas and A.V. Kurpad. Total body muscle mass estimation from bioelectrical impedance analysis and simple anthropometric measurements in Indian men. *Indian J Med Res*, **127**, 5, 441–446, 2008.

*** Harinarayan, C.V., T. Ramalakshmi, U.V. Prasad and D. Sudhakar. Vitamin D status in Andhra Pradesh: A population based study. *Indian J Med Res*, **127**, 3, 211–218, 2008.

other than myocardial infarction (3.04, 2.73–3.39), which includes some misclassified alcohol poisoning. There were significant excesses of upper aerodigestive tract cancer (3.48, 2.84–4.27) and liver cancer (2.11, 1.64–2.70). Another five disease groups had RRs of more than 3.00 in the highest alcohol category: tuberculosis (4.14, 3.44–4.98), pneumonia (3.29, 2.83–3.83), liver disease (6.21, 5.16–7.47), pancreatic disease (6.69, 4.98–9.00), and ill-specified conditions (7.74, 6.48–9.25). Although drinking was less common in women, the RRs associated with it were generally more extreme. After correction for reporting errors, alcohol-associated excesses accounted for 52% of all study deaths at ages 15–54 years (men 8,182[59%] of 13,968, women 1,565 [33%] of 4,751) and 18% of those at 55–74 years (men 3,944 [22%] of 17,536, women 1,493 [12%] of 12,302).*

1.4 STUDY DESIGN

1. This pilot study represents the initial phase of a multi-country global childhood unintentional injury surveillance (GCUIS) project and was based on a sequential sample of children < 11 years of age of either gender who presented to selected Emergency Departments in Bangladesh, Colombia, Egypt and Pakistan over a 3–4 months period, which varied for each site, in 2007. Of 1,559 injured children across all sites, 1,010(65%) were male; 941 (60%) were aged ≥ 5 years, 32 (2%) were < 1 year old. Injuries were especially frequent (34%) during the morning hours. They occurred in and around the home in 56% of the cases, outside while children played in 63% and during trips in 11%. The mean injury severity score was highest for near drowning, 2 from falls and 2 from road traffic injuries.**

2. In a double-blind trial in 50 centres in India, 2,053 individuals without cardiovascular disease, aged 45–80 years, and with one risk factor were randomly assigned, by a central secure website, to the polycap (n = 412) consisting of low doses of thiazide (12.5 mg), atenolol (50 mg), ramipril (5 mg), simvastatin (20 mg), and aspirin (100 mg) per day, or to eight other groups, each with about 200 individuals, of aspirin alone, simvastatin alone, hydrochlorthiazide alone, three combinations of the two blood-pressure-lowering drugs plus aspirin. The primary outcomes were LDL for the effect of lipids, blood pressure for antihypertensive drugs, heart rate for the effects of atenolol, urinary 11–dehydrothromboxane B2 for the antiplatelet effects of aspirin, and rates of discontinuation of drugs for safety. Analysis was by intention to treat.

Compared with groups not receiving blood-pressure-lowering drugs, the Polycap reduced systolic blood pressure by 7.4 mm Hg (95% CI 6.1–8.1) and diastolic blood pressure by 5.6 mm Hg (4.7–6.4), which was similar when three bold-pressure-

* David Zaridze, Paul Boreham, Alex Boroda, Rostislav Karpov, Alexander Lazarev, Irina Konobeevskaya, Vladimir Igitov, Tatiana Terechova, Paolo Boffetta, Richard Peto. Alcohol and cause-specific mortality in Russia: a retrospective case–control study of 48,557 adult deaths. The Lancet, **373**, 2201–14, 2009.

** Adnan A. Hyder, David E. Sugerman, Prasanthi Puvanachandra, Junaid Razzak. Hesham El-Sayed, Andres Isaza, Fazlur Rahman and Margie Peden. Global childhood unintentional injury surveillance in four cities in developing countries: a pilot study. *Bull World Health Organ*, **87**, 345–352, 2009.

lowering drugs were used, with or without aspirin. Reductions in blood pressure increased with the number of drugs used (2.2/1.3 mm Hg with one drug, 4.7/3.6 mm Hg with two drugs, and 6.3/4.5 mm Hg with three drugs. The reductions in heart rate with Polycap and other groups using atenolol were similar (7.0 beats per min), and both were significantly greater than that in groups without atenolol ($p < 0.0001$).*

3. We undertook a randomised, double-blind, placebo-controlled, multicentre study in 40 centres in 11 countries. We enrolled outpatients with adult-onset type 2 diabetes and mild-to-moderate non-proliferative diabetic retinopathy, and randomly allocated them via sealed envelopes to either calcium dobesilate (1500 mg per day) or placebo. The primary endpoint was development of clinically significant macular oedema (CSME) within a follow-up period of 5 years. Patients who dropped out of the study early were censored. Analysis was by intention to treat.

We enrolled 635 patients. A total of 324 were randomly allocated calcium dobesilate and 311 were assigned placebo. Accounting for censored cases, estimated cumulative 5-years CSME probability was 35% and 28%, respectively (hazard ratio 1.32, 95% CI 0.96–1.81; $p = 0.0844$). Adverse events did not differ between treatment groups (78[24%] on calcium dobesilate and 90[29%] with placebo). No relevant drug-related complications were noted. Nine patients (3%) died in the calcium dobesilate group and eight (3%) deaths were recorded on placebo.**

4. The trial included patients aged 5–65 years with scabies who attended the dermatology department at the Institut d' Hygiene Sociale in Dakar, Senegal. The randomized, open trial considered three treatments: a single application of 12.5% BB over 24 hours (BB1 group), two applications of BB, each over 24 hours (BB2 group), and oral IV, 150–200 μg/kg (IV group). The primary endpoint was the disappearance of skin lesions and itching at day 14. If necessary, treatment was repeated and patients were evaluated until cured. Results were analysed on an intention-to-treat basis. A pre-planned intermediate analysis was carried out after the BB1, BB2 and IV groups had recruited 68, 48 and 65 patients, respectively.***

QUESTIONS

1. Explain the meaning and uses of 'Biostatistics'.
2. Illustrate the concept of estimation.
3. Explain the use of test of hypothesis. Give an example.

---

* The Indian Polycap Study (TIPS), Effects of a polypill (Polycap) on risk factors in middle-aged individuals without cardiovascular disease (TIPS): a phase II, double–blind, randomized trial, The Lancet, **373**, 1341–51, 2009.

** Christos Haritoglou, Joachim Gerss, Cristina Sauerland, Anselm Kampik, Michael W Ulbig, for the CALDIRET study group, Effect of calcium dobesilate on occurrence of diabetic macular oedema (CALDIRET study): randomised, double-blind, placebo-controlled, multicentre trial, The Lancet, **373**, 1364–71, 2009.

*** Fatimata, Ly, Eric Caumes, Cheick Ahmet Tidiane Ndaw, Bassirou Ndiaye and Antoine Mahe. Ivernectin versus benzoate applied once or twice to treat human scabies in Dakar, Senegal: a randomized controlled trial. *Bull World Health Organ*, **87**, 424–430, 2009.

CHAPTER

2 Role of Statistics in Clinical Medicine

2.1 INTRODUCTION

Although, the primary object of medical science is to improve the health or cure the diseases of the 'individual', the relevant knowledge usually has to be accumulated slowly by the observation of groups of individuals. Some applications also may have a group character as in the newer fields of prophylaxis and population genetics. In treating a single patient, an appraisal of the total situation must be made and a judgement formed as to the best course of action. Frequently, a balance has to be struck between conflicting requirements, such as those of medicine, economics, and ethics, though knowledge in each of these fields may at best be only partial. In drawing a conclusion from a series of case histories, or interpreting the results of a survey or a therapeutic trial, the situation is quite different. We are now inescapably involved in studying groups of individuals. Percentages have to be compared, survival curves scrutinised, attack rates measured, and so on. No longer does one have the relatively simple numerical work entailed by the measurement of some physiological functions, or the calculation of the correct pharmacological dose. Some sort of statistical appraisal is required, though it may be only an elementary one.

Much progress has been made in the theory of statistics in recent years in improving the techniques and methods of estimating and handling of numerical data. Statistics as a science provides an investigator with powerful and sensitive tools to aid him in determining the likelihood that a particular outcome of treating a group of patients is the result of treatment, the result of other factors, or the result of chance; whether a characteristic is truly associated with the occurrence of disease or an artefact of observation. These are fundamental questions. To ignore them and to scoff at the use of statistics in medicine and medical research as 'mere statistics' is to ignore the systematic observation and measurement of phenomena which are the very basis of science.

2.2 SOME EXAMPLES OF EARLIER USES OF STATISTICS

The use of statistical methods in medicine is not new. The history of the development of modern medicine and epidemiology are replete with examples of how pioneers made startling discoveries using simple methods of designing experiments and analysing the findings.

As early as in 1772, a simple prophylactic trial was made by Captain Cook when he used antiscorbutics such as inspissated juice of wort and obtained freedom from scurvy. This was in sharp contrast to the unhappy experience of earlier travellers

on the seas and Cook seems to have regarded this of as great importance as his geographical discoveries. More important in the technical development of clinical research was the work of Goldberger who was appointed by U.S. Bureau of Health in 1913 to investigate an outbreak of pellagra in the southern states of the U.S.A. By systematic dietary investigations in the field, Goldberger demonstrated that pellagra was a deficiency disease which could be prevented by specific changes in the diet. His work was a particular triumph because the initial laboratory studies on pellagra had been confusing owing to the relative insusceptibility of some of the common experimental animals to the disease and the interplay of nicotinic acid, tryptophan and intestinal bacteria in the aetiology. Other examples include the work of Panum on measles, the work of Takaki on beri-beri, Budd on typhoid, Collis on silicosis, etc.

2.3 AREAS OF APPLICATION OF STATISTICS

During the last fifty years considerable changes have occurred in the methods of clinical research. At the beginning of this period there was still much work to be done on the recognition and description of particular diseases and syndromes. The description of previously unrecognised diseases still goes on, and in addition, new diseases appear, particularly in the fields of infectious disease and industrial medicine. The pathology of survival has to be written for diseases such as diabetes mellitus or pernicious anaemia which were formerly rapidly fatal. The advance of surgery has made it necessary to try to recognise the different forms of congenital heart disease during life. Nevertheless, this descriptive type of clinical research is subject to the law of diminishing returns and is in danger of becoming the pursuit of the esoteric.

The doctor's work in the past had often consisted of weighing a number of imponderables and coming to a decision. He was conscious of the variation between individuals and had felt it part of his skill to decide what would suit one and not the other. On this account, practising clinicians have not always taken kindly to the statistical approach to medical problems in which patients are considered as units in a more or less homogeneous whole and their individual variations as an aspect of the normal distribution curve. It is doubtful if such rapid headway would have been made had it not been for the obvious contribution of statistical methods and planned trials to therapeutics. The accelerating pace of discovery entailed the constant trial of new remedies whose value could not be left to be determined by the slow processes of time and fashion as in the past.

Few diseases offer such a simple problem as tuberculous meningitis which was almost invariably fatal until the introduction of streptomycin. More often the problem is like that of the use of corticoids in ulcerative colitis where it could have been argued *a priori* that the anti-inflammatory action of the hormones might be either helpful or harmful and where the first trials gave equivocal results; only a large-scale trial could decide the balance of advantage and disadvantage. Again, any major advance in chemotherapy is followed by variations on the theme, and a succession of sulphonamides, antibiotics or antithyroid drugs is produced. The difference between these variants is likely to be small but nonetheless important, particularly if it has to be weighed against differences in side reactions and cost. Reports of therapeutic trials have thus become a staple item in the medical journals and there

can be few doctors now who do not have at least some acquaintance with such concepts as controls and levels of significance. Much of traditional therapeutics has been put to question and the new methods have been applied to the evaluation of the older drugs.

It was inevitable that the new techniques should be applied to other aspects of disease besides its treatment. Previous descriptions of disease had often resembled snapshots, vivid and dramatic, but giving no hint of what came before and after. When surgeons began to evaluate the results of their operations, they found that little was known of the natural history of many of the conditions they were treating. Indeed, it is difficult even today to say what happens to a group of patients with peptic ulcer or ulcerative colitis after 5, 10 and 20 years. Before we can assess the effects of treatment on a disease, we must know the normal variations in duration, severity and complications, and in the more chronic illnesses we must work out a survival curve or life table. Moreover, it soon became apparent that there were different samples of patients with peptic ulcer: those who went about their business without seeking medical advice, those who were looked after by their family doctors, and those who were bad enough to be sent to a hospital or a consultant. The standard textbook accounts of some diseases such as disseminated sclerosis which had been founded on experience in hospitals gave an exaggerated account of their severity.

Therapeutic experiments also necessitated a more exact definition of diseases and the estimation of the frequency of different symptoms and signs, their correlation or independence, and the accuracy with which they could be measured. The semantic and philosophic problems in defining disease have been exaggerated, for it is no more difficult than defining other biological reactions and it is a reversion to primitive ways of thinking to say there are no diseases, there are only sick people. On the other hand, the relevance and reliability of some of the classical symptoms and signs of disease have been shown to be questionable and the difference between different observers greater than had been realised. Some measurement of the accuracy of these data is desirable and this of course requires study of groups of cases.

In summary then, statistical methods are necessary in the following broad areas of clinical medicine:

1. In the documentation of the medical history of diseases, their course of progression, and variability among patients.

2. In clinical epidemiological studies on aetiology of health events, determination of relative risks and odds ratios.

3. In the planning and conduct of therapeutic and prophylactic clinical trials to determine the superiority of specific therapies.

4. In calculation of diagnostic efficacies and likelihood ratios in prognostic studies.

5. In providing methods for the definition of "normal", and indicating at what point the measure of some bodily characteristic becomes "abnormal" or "pathological".

6. In supplying the measurements of accuracy of various laboratory and clinical procedures that are commonly used in medical practice.

QUESTIONS

1. Write a brief note on the recent changes in the methods of clinical research.
2. Explain the role of statistical methods in therapeutic experiments.
3. Illustrate the uses of statistical methods in five broad areas of clinical medicine.

CHAPTER

3 Role of Statistics in Preventive Medicine

3.1 INTRODUCTION

In public health work, we may sometimes be concerned with planning of experiments and the analysis of their results, e.g., in testing the efficacy of a vaccine as a means of preventing an attack of say, whooping cough; or in measuring the effects on mother and infant of supplementing the diet of pregnant woman in some particular way. More often, however, we have to deal in this field with statistics that come from no deliberate experiment but that arise, and are collected, from a population living and dying in an everyday course of events.

3.2 SOME EARLY EXAMPLES

John Graunt, known popularly as the Father of health statistics, was neither a physician nor a mathematician. He earned his living as a haberdasher in London but he was interested in the parish registers containing records of deaths. He thought he could make a useful study of where and from what causes people die, at what age, in which season, etc. By using simple commonsense methods for analysing the information, he formulated certain laws which are as true today as they were in the year 1662 when he published his work 'Natural and political observations made upon Bills of mortality'. He thus established the uniformity and predictability of many important phenomena taken in the mass.

More than a century ago, Sir John Snow applied statistical data and reasoning in determining that cholera is spread by polluted water. Snow arrived at this conclusion several years before Pasteur enunciated the germ theory and several decades before the cholera bacillus was identified, by tabulating the cases of cholera among the users of two London water supplies.

3.3 AREAS OF APPLICATION

One of the chief functions of statistics in the field of public health is to present numerical information that is relevant, reliable, complete, comparable and up-to-date. On this secure foundation, the administrator, instead of having to rely on vague impressions, can know exactly what is happening in the population under his care and how best to organise and utilise his limited staff and resources to build a sound structure of health. It may be that one area suffers a high mortality from malaria, another from a serious degree of debility and yet another from an insidious loss of efficiency as a result of widespread eye diseases. In a less developed area

people may be dying in great numbers from epidemics of typhus or cholera; yet if the health officer cannot inform himself promptly of what is really wrong and the extent of the damage, his preventive staff are working in the dark, and time and money are being wasted. With reliable information, on the other hand, the health officer can feel the pulse of the community he serves. It would hardly be necessary at this stage of national development to lay stress on the value of statistics but time and again their functions and limitations have been misunderstood.

The second, and of course closely associated, object in the collection of such figures is to find out the basic reasons for the contrasts observed. For, unless we can determine those reasons, the development of effective preventive measures would obviously be hampered and may be misdirected. Why, for example, do infants of the more impoverished classes die at a higher rate than those of the wealthier? To what extent, for instance, is it due to the malnourishment of the mother and child, to what extent to overcrowding in the home and a more frequent risk of specific infections, to what extent to ignorance of how to care for the infant, or even, sometimes, to deliberate neglect of it?

Is a higher incidence or an epidemic of typhoid fever in a particular area or type of area due to a defective water supply, to milkborne infections, or to some other form of transmission? Does the preschool child die more frequently under conditions of overcrowding because such conditions expose it to a greater risk of infection early in life, or does it succumb more easily to the infection it has acquired through factors associated with overcrowding, such as malnutrition or possibly, lack of skilled attention in the early stages of illness?

Here the statistical method comes into play endeavouring to disentangle the chain of causation and allowing us, sometimes, to determine the most important factors in need of correction. Since we are dealing with uncontrolled observations, often liable to errors, the task may be very difficult; the effects of the 'multiplicity of causes' often cannot be completely distinguished, e.g., the effects on health of overcrowding *per se* as apart from the features of poverty which invariably accompany it. But the original vital statistics having indicated the problem and their analysis having at least suggested a cause and effect, we may be able to progress further by a more deliberate collection of additional data or sometimes by a specifically designed experiment.

Basically then, we may recognize the uses of statistics in public health and preventive medicine in the following areas:

1. In assessing the state of health in a community, in listing the met and unmet needs and in providing a community diagnosis, and in predictions.

2. To identify the basic socioeconomic, political or demographic factors underlying the state of health, and to suggest possible solutions for control, elimination and eradication of communicable and noncommunicable diseases.

3. To assist in health programme evaluation, and health systems research, in determining the success or failure of specific health interventions, and the reasons thereof.

4. In planning and implementing community health and development programmes, and in their monitoring/surveillance.

5. To promote health legislation and health equities, creating administrative standards for health activities and promoting social justice.

Vital statistics and their analyses are considered the "eyes and ears" of the public health administrator, serving as powerful and valuable indicators.

QUESTIONS

1. Explain the contributions of John Graunt and Sir John Snow.
2. Briefly explain the role of statistical methods in identifying the cause(s) of diseases.
3. What are the broad areas of public health in which statistical methods are used?

CHAPTER

4 Variables

4.1 INTRODUCTION

In the first chapter, we defined Statistics/Biostatistics in many ways, most of which emphasize the variability of data collected. This variation is due to a number of factors, some within the control of the investigator and many beyond one's control. Medical and health data have some unique features that are responsible for the variation, and it is this aspect that makes Biostatistics an interesting and challenging discipline.

4.2 DEFINITION OF VARIABLES

A simple definition of a variable is anything that varies or changes. Example of variables are age, sex, income, health status, education, occupation, exposure, height, weight, blood pressure, etc. These values will vary from person to person, from time to time even in the same person, or vary in terms of tool used or investigator's qualification. The main objective in Biostatistics is to understand and summarize variation.

4.3 QUALITATIVE VARIABLES

Characteristics such as sex, presence of a disease in an individual, occurrence of an event like birth, cause of death, etc., are examples of qualitative variables. In the case of sex, the observed individual can be either male or female. Similarly, in the case of cause of death, the observation may be any one of the listed categories of causes of death. For any characteristic, the observation we make is any one of the possible alternative situations or categories for that characteristic. When several individuals are observed for the same characteristic,we may be able to indicate the number of individuals found in each category (or class). For instance, the type of leprosy observed for 249 leprosy patients is shown in Table 4.1.

When observation is made for more than one characteristic on each individual, the classification can be made simultaneously for all these multiple characteristics and the number in each classification can be found. If in addition to the type of lepsory, information on gender was obtained, we can obtain the number

Table 4.1 Classification of Leprosy Patients according to the Type of Leprosy

Type of leprosy	No. of patients
Tuberculoid	151
Lepromatous	68
Indeterminate	18
Borderline	12
Total	249

of individuals found in each of 8 categories: Tuberculoid males, tuberculoid females, lepromatous males, lepromatous females, etc. (Table 4.2).

Table 4.2 Classification of Leprosy Patients according to the Type of Leprosy and Gender

	No. of patients		
Type of leprosy	Male	Female	Total
Tuberculoid	77	74	151
Lepromatous	35	33	68
Indeterminate	10	8	18
Borderline	7	5	12
Total	129	120	249

4.4 QUANTITATIVE VARIABLES

In the case of a qualitative characteristic, the observer finds out the class or category in which the individual can be placed. There are other characteristics for which a measurement can be made or a quantitative representation is possible. Level of haemoglobin in the blood, the stature of a child, age, or number of attacks of diarrhoea are examples of such characteristics. The observation on each individual is represented by a number or quantity. Such observations are known as quantitative variables. Quantitative observations can be of two types based on the nature of the characteristic on which they are made namely (a) continuous and (b) discrete or discontinuous. A characteristic whose observation can take any value over a particular range is known as a continuous variable, e.g., height, age, level of protein in blood, etc. However, there are characteristics the observations on which can have only intermittent values over a range, e.g., number of attacks of diarrhoea in a child during a year, white blood cell count in a blood sample, etc. Such characteristics are known as discrete variables.

Table 4.3 shows classification according to a quantitative continuous variable.

In this table the scale of measurement is continuous because any value on the scale (of grams of haemoglobin per 100 cc of blood) is theoretically possible. There are no natural classifications as in the case of qualitative observations. Therefore, we divide the range of observations in a suitable way which would bring out the essence

of the data. The number of observations in each class-interval is called *class frequency*. For convenience, usually the class-intervals are chosen to be equal so that the class frequencies are comparable.

Table 4.3 Distribution of 215 Normal Children according to the Haemoglobin Content of Their Blood

Grams of haemoglobin per 100 cc of blood	No. of children
9.0–9.4	1
9.5–9.9	4
10.0–10.4	8
10.5–10.9	22
11.0–11.4	36
11.5–11.9	47
12.0–12.4	38
12.5–12.9	34
13.0–13.4	10
13.5–13.9	7
14.0–14.4	6
14.5–14.9	2
Total	215

An example of classification according to a discrete quantitative variable is shown in Table 4.4.

Table 4.4 Differential Counts from the Blood of a Person Classified according to the Number of Eosinophils

No. of eosinophils encountered in 100 W.B.C.	No. of smears
0	11
1	20
2	20
3	7
4	3
5	4
6	2
Total	67

The variable of measurement is number of eosinophils, and the objects being measured are the blood smears.

The last two tables (Tables 4.3 and 4.4) are also known as Frequency tables or Frequency distributions.

4.5 SUBJECTIVE AND OBJECTIVE VARIABLES

Generally, qualitative variables are likely to be subjective, and quantitative variables objective. However, one needs to be cautious in judging whether a variable is

subjective or objective as it depends not only on the phenomena but largely on the measurement tool used. For example, the use of an Anthropometer or Beam Balance will provide an objective assessment of the variable, weight. On the other hand, the judgment of the investigator or the respondent on the weight will be subjective. Likewise, variables assessed during an interview or questionnaires will be subjective. On the other hand, even when there is an instrument or an objective tool, the measuring process can make it subjective by not following the rules, or by categorization. For example, height could be objectively measured but categorized as tall or short, using arbitrary cut-off points. Sometimes, subjective variables such as Pain or Severity or Stress can be given a numerical value and made to appear objective by a validated scoring system.

In health work, many variables are subjective, and it is the responsibility of the investigator to ensure its precision and accuracy, even if there is no objective instrument other than the assessor. Often, training and standardization become very important in controlling variation due to the measurement process.

4.6 DEPENDENT AND INDEPENDENT VARIABLES

When more than one variable is assessed, it becomes a matter of interest to determine the relationship or association between one variable and another. Based on experience or expertise, one variable is then considered dependent on the other. For example, in a study of maternal age and baby's birth weight, the birth weight is considered a dependent variable and the maternal age as independent. However, association *per se* should not be considered causation, which requires more careful study. In some examples, such as height and weight among adults, it may not be possible to say which depends on which.

We will deal more with this concept later.

4.7 MEASUREMENT OF VARIABLES

Data can be collected in many ways:

1. through an instrumentation system, e.g., through an anthropometer, a thermometer, biochemical equipment, etc.

2. through observation of a qualified and trained investigator, e.g., observing cardinal signs and symptoms, observing hygienic environment, presence of swelling or lump, etc.

3. through enquiry, face-to-face or telephone/electronic communication, e.g., history of past illnesses, duration of present problem, age, etc.

4. through special psychological testing or special psychological tools, such as TAT, Rorschach Inkblot test, etc.

Whichever approach is used, all tools need standardization as well as a high degree of reliability (consistency, reproduceability, repeatability) as well as validity (accuracy), without which the data collected cannot be interpreted properly. We will discuss more on this aspect later under research methodology.

4.8 SCALE OF MEASUREMENT

The weakest level of measurement is the *nominal* scale. In this scale, the observations are placed in broad categories which may be denoted by symbols or labels or names. For example, classifying patients according to diagnostic groups or type of residence (rural, urban) constitute a nominal scale. The next higher level of measurement is the *ordinal* scale. This consists of ranking all categories of a variable according to some criterion such as grading of muscle strength or ranking of socio-economic class. In this scale each category is in a unique position in relation to other categories, but the *distances* between categories are not known. The *interval* scale has, in addition to ordinal level of measurement, an important characteristic, in that, the distance between any two numbers (values of the variable) is fixed and equal. In this scale, the zero point and the unit of measurement are arbitrary. Measurement of temperature—Centigrade and Fahrenheit—is an excellent example of this. The highest level of measurement has all the characteristics of an interval scale and in addition has a true zero point as its origin. It is called a *ratio* scale. We measure weight or physical distances in ratio level of measurement.

4.9 RECORDING THE VARIABLES

While collecting proper data is important, recording them properly becomes equally important. Thus, developing the data record or pro forma needs thought in terms of the quality of paper used, the size, the spacing and layout, the instructions on how to fill in the data with the required degree of precision, all are critical in any study. Ready-made forms are useful, but custom-made pro forma will be better. The storage and filing of the records so that they could be easily retrieved is also important.

4.10 ROLE OF COMPUTERS

The availability of desktop and laptop computers with excellent and user-friendly software to record data must be fully utilized to record, edit, and store data properly. Microsoft Excel and Access, SPSS, and Epi-Info are some of the popular software available and proved to be most useful to the investigator. Data management along with data collection, ensures that the data are ready for analyses without wasting further time for data entry and encountering incomplete or inaccurate data requiring corrections. It is always good practice to take one or two backups of the data at the end of each data entry session and store them separately, outside the computer system you are using.

4.11 QUALITY OF DATA

"Garbage in-garbage out" or "GIGO" is a popular saying. This means that the researcher must ensure a high quality of data at every stage to ensure high quality results and valid conclusions. At every stage means during data collection, during data management and analyses, and during interpretation, comparing with other studies, and discussing the results. We will discuss these concepts in more detail later.

QUESTIONS

1. Define 'variables'. Bring out the essential features of qualitative and quantitative variables giving examples.
2. Explain with examples the differences between subjective and objective variables.
3. Explain the various scales of measurement, giving examples.

CHAPTER

5 Frequency Distribution

5.1 INTRODUCTION

One of the major objectives of statistical methods is to describe appropriately the characteristics of a mass of data. In any investigation, after collecting the necessary information, the next step should be to express it in some form which will permit conclusions to be drawn directly or by means of further calculations. It is not possible to detect relationships between various factors at issue from the unsorted mass of figures. The investigator must first consider the questions which he believes the material is capable of answering and then determine the form of presentation which brings out the true answers most clearly. For instance, let us suppose we have amassed a series of after-histories of patients treated for gastric ulcer and wish to assess the value of the treatment given, using as a measure, the amount of incapacitating illness suffered in subsequent years. There will be various factors the influence of which will be of interest. Is the age or sex of the patient important to the outcome? Division of the data must be made into appropriate categories and tables constructed to show how much subsequent illness was in fact suffered by each of these categories. Is the after-history affected by the type of treatment? A further tabulation is necessary to explore this point. And so on. The initial step must be to divide the observations into a relatively small series of groups, those in each group being considered alike in that characteristic for the purpose in hand. The tabulation or preparation of frequency tables thus constitutes an important step when the number of observations is large. In order to further help the reader to see the pattern or shape of any complex situation, diagrams and graphs are drawn.

5.2 FREQUENCY DISTRIBUTIONS

Tabulation (or classification) is the process by which data of a long series of observations are systematically organised and recorded so as to enable analysis and interpretation. A set of categories is formed to classify the data. The essential requirement of any such set of categories is that these categories should be selected in such a way that each one of the observations falls into one, and only one, of these categories and that all the observations fall into any one of these categories. In setting up any table it is necessary to examine whether each classification (or category) used is pertinent to the problem. Further, the reader must be attracted to read the table while reading the text of the report.

A table must satisfy the following requirements:

1. The table must be simple and clear-cut.

2. The title of the table should explain in exact terms what the data represent.

3. The figures in the body of the table must be arranged in a logical order for the point discussed in the text.

4. When several points are to be emphasised making use of the same data, it may be preferable, whenever possible, to present the data in many small tables, one to illustrate each point.

A method of forming frequency distribution for a continuous variable (interval as well as ratio-level measurement) is illustrated for the 500 haemoglobin values given in Appendix 1. Glance through the data and pick out the lowest and the highest observations. Select a suitable width (class interval) and divide the whole range of observations into 10 to 15 classes. For each observation, make a tally mark against the class in which it falls. For example, for the first observation, 11.6, make a tally mark against the class 11.5–12.4. After marking all observations, count the tallys and write the frequencies; this work-sheet will look like Table 5.1.

Table 5.1 Work-Sheet Showing the Frequency Distribution for 500 Haemoglobin Values Given in Appendix 1

5.5–6.4	/	1
6.5–7.4	𝍸 ///	8
7.5–8.4	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 /	31
8.5–9.4	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ////	44
9.5–10.4	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	95
10.5–11.4	𝍸 𝍸	125
11.5–12.4	𝍸 ////	124
12.5–13.4	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ////	49
13.5–14.4	𝍸 ///	8
14.5–15.4	𝍸 ////	9
15.5–16.4	𝍸 /	6

5.3 DIAGRAMS

In addition to tables, graphs and diagrams are widely used in the presentation of statistical data either alone or in conjunction with the tables. They furnish a visual method of examining quantitative data. They make a more lasting impression than

detailed numbers and convey an idea forcefully. They help us to get a real grasp of the overall picture rather than the details.

In order that these graphs and diagrams present ideas truthfully and emphasise correct ideas, they must be drawn following certain basic rules which are dependent partly on convention, partly on mathematical considerations, and partly on personal preferences.

Many types of graphs and diagrams are commonly used. We shall consider some important types in relation to frequency distributions.

Histogram. A histogram is a special kind of bar diagram used to present a frequency distribution of a characteristic measured on a continuous scale. Rectangles are erected over class intervals to represent the frequencies of the class intervals (Figure 5.1). In a histogram, the area of each rectangle represents the frequency of

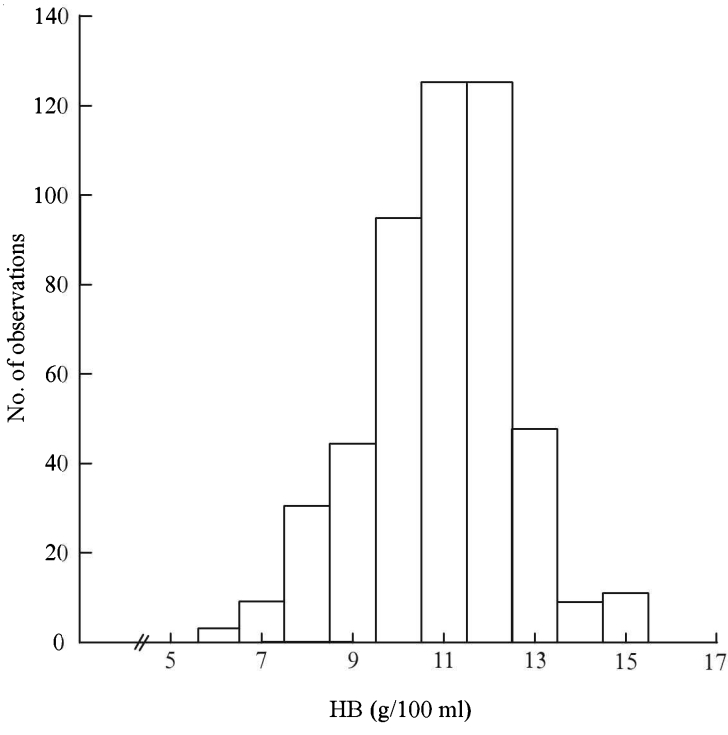


Figure 5.1 Histogram of 500 Hb Values.

the corresponding class interval. (A mistake is likely to be made in the case of a frequency distribution with unequal class intervals). The presentation in a histogram of the distribution in Table 5.2 is illustrated in Figure 5.2.

Frequency polygon. A frequency polygon is a variation of histogram. Instead of rectangles erected over the intervals, points are plotted at the mid-points of the tops of the corresponding rectangles in a histogram, and the successive points joined by

Table 5.2 Distribution of Total Serum Protein Levels (g/100 ml) in 436 Individuals (Protein Nutrition Survey in Pennathur—1956)

Total serum protein (g/100 ml)	No. of individuals
4.0–	4
5.0–	12
6.0–	7
6.2–	9
6.3–	34
6.5–	105
7.0–	237
8.0–	27
9.0–10.0	1
Total	436

Source: *Ind. Jour. Med. Sci.* **13**, 403, 1959.

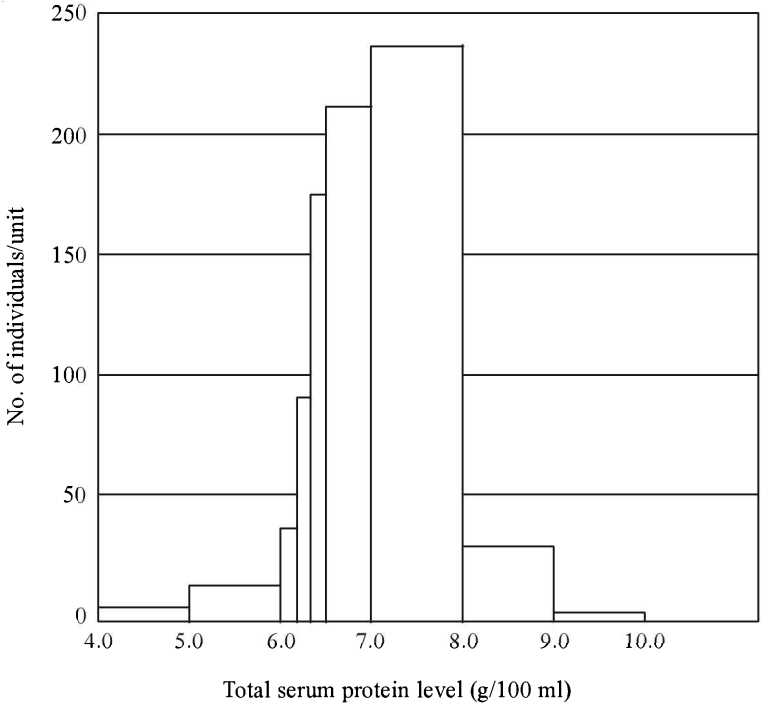


Figure 5.2 Distribution of total serum protein levels (g/100 ml) in 436 individuals (Protein nutrition survey in Pennathur—1956).

straight lines. A frequency polygon may be chosen to compare two frequency distributions as is illustrated in Figure 5.3.

Frequency curve. When the total frequency is large and when we adopt much narrower class intervals the frequency polygon will most often have a much smoother appearance. If the total frequency is increased indefinitely, the frequency polygon



Figure 5.3 Total births in Madras and Punjab classified according to age of mother.

will approach a smooth curve. This limiting condition is known as the frequency curve. In theoretical studies we come across many such curves.

Cumulative frequency polygon or Ogive. Figure 5.4 illustrates a cumulative frequency polygon which is also known as ogive. The number of observations falling

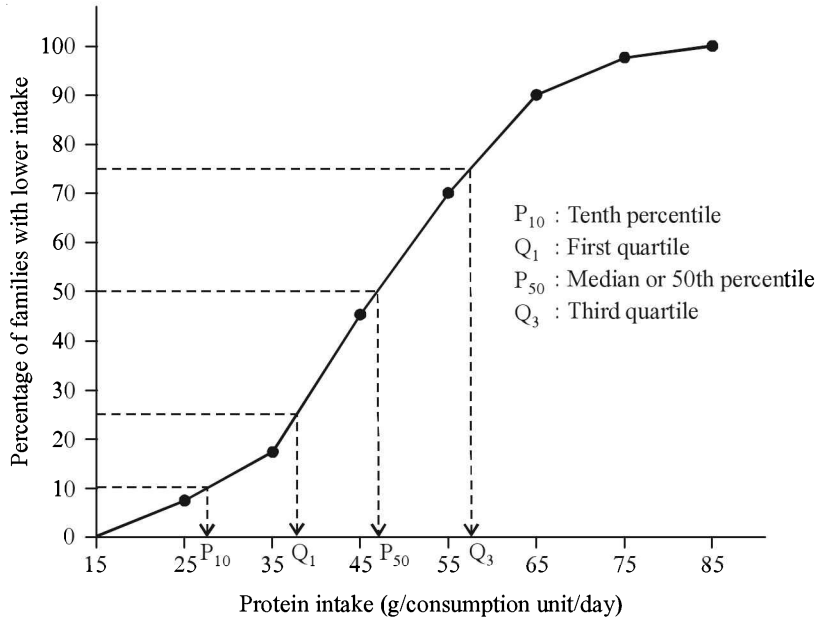


Figure 5.4 Computation of percentiles based on cumulative percentage curve (Based on Table 5.2).

below each specified value is presented. In a similar way, the number of observations falling above each specified value may also be presented. Most often the observations are presented as a percentage instead of the actual numbers.

Further methods of presentation of data as graphs and diagrams are described in Chapter 26.

5.4 CHARACTERISTICS OF A FREQUENCY DISTRIBUTION

After forming a frequency distribution, the next step is the calculation of certain values which may be used as descriptive measures of the characteristics of that distribution. These values will enable comparisons to be made between one series of observations and another.

There are four major characteristics of a frequency distribution:

1. the measure of central tendency (location, position, average);

2. the degree of scatter of the observations around the measure of central tendency (variability, dispersion, spread);

3. the extent of symmetry in the shape of the distribution (skewness, asymmetry); and

4. the flatness or peakedness of the distribution (kurtosis).

Of these 4 characteristics, the first two are most commonly employed in describing a set of data.

When more than one factor is observed, it may be necessary not only to describe each of these factors in terms of the 4 descriptive measures mentioned above but also in terms of the degree of association between various factors considered. The two most commonly used characteristics are (a) correlation, and (b) regression. In the next few chapters, we will define and discuss these concepts and the measures used.

Tips for generating frequency distribution using computer programs are given in Chapter 25.

QUESTIONS

1. Explain the method of forming a frequency distribution.
2. Mention the uses of some commonly used diagrams to represent frequency distributions.
3. What are the major characteristics of a frequency distribution?
4. Form a frequency distribution from the following Hb values.

9.0	9.0	9.7	10.3	11.5
12.6	11.6	13.0	9.5	10.0
11.0	12.3	11.6	10.2	10.5
12.8	11.8	11.0	10.5	9.7
13.0	11.6	9.7	11.2	12.5

CHAPTER

6 Measures of Central Tendency and Location

6.1 INTRODUCTION

In the case of many biological characteristics, the values of the extent of the observations are not equal, but we notice a general tendency of such observations to cluster around a particular level. In this situation it may be preferable to characterise each group of observations by such a level, which is called the Central Tendency of that group. This single value for each group of observations serves as a representative of that group. This level around which the observations tend to cluster may vary from group to group. For example, the average incubation period of one infectious illness may be 7 days and of another 11 days. Though individual values may overlap, the two distributions have different central positions and, therefore, differ in the characteristic of location. In practice, it is constantly necessary to discuss and compare such measures. A simple instance would be the observation that persons following one occupation lose, on the average, 5 days a year per person from illness while in another occupation they lose 10 days. The two distributions differ in their position and we are led to seek the reasons for such a difference and to see whether it is remediable.

Several measures of central tendency can be calculated for a group of observations. Let us consider in this chapter a few commonly used measures of central tendency, viz., the mean, median and mode. Mean could be arithmetic, geometric or harmonic. All these measures may not have the same value for a particular group of observations. Which one of these measures we have to use in a particular case depends upon the way in which the observations in the group cluster around that central tendency.

6.2 ARITHMETIC MEAN

The arithmetic mean of a group is the simple arithmetic average of the observations. This is calculated by dividing the total of all the observations by the number of observations. In the case of grouped data (frequency distribution), arithmetic mean is calculated assuming that each observation in a class interval is equal to the mid-point of that class interval.

The following formulae are used to calculate the arithmetic mean. In an ungrouped data, if x represents the character observed, and n the number of

observations, then all the observations in the data can be denoted as $x_1, x_2, \ldots, x_n$. The arithmetic mean is given by

$$m \text{ or } \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\Sigma x}{n}$$

In this formula, the Greek letter Σ denotes summation of the values (i.e., $\Sigma x = x_1 + x_2 + \cdots + x_n$). For the grouped data (frequency distribution) the arithmetic mean is given by $\bar{x} = \frac{\Sigma fx}{n}$ where f is the frequency, x the mid-point of the class interval and n the total number of observations.

Table 6.1 Calculation of Arithmetic Mean for a Series of Serum Albumin Levels (g%) of 24 Pre-School Children

2.90	3.75	3.66
3.57	3.45	3.76
3.73	3.71	3.43
3.55	3.84	3.69
3.72	3.30	3.77
3.88	3.62	3.43
2.98	3.76	3.68
3.61	3.38	3.76

The total of all these values, i.e., $\Sigma x = 85.93$.
Total number of observations (n) = 24

Therefore, the arithmetic mean, $\bar{x} = \frac{\Sigma x}{n} = \frac{85.93}{24} = 3.58$ g%

Table 6.2 Calculation of Arithmetic Mean of Protein Intake of 400 Families

Protein intake/consumption unit/day (g)	No. of families	Mid-point of class interval	Multiply f & x
Class interval	f	x	fx
15–25	30	20	600
25–35	40	30	1200
35–45	100	40	4000
45–55	110	50	5500
55–65	80	60	4800
65–75	30	70	2100
75–85	10	80	800
Total	400		19000

Arithmetic mean:

$$\frac{\Sigma fx}{n} = \frac{30 \times 20 + 40 \times 30 + \cdots + 10 \times 80}{400}$$

$$= \frac{19000}{400} = 47.50 \text{ g}$$

6.3 MEDIAN

The median is the magnitude of the observation which occupies the middle position when all the observations are arranged in order of their magnitudes. When there are an even number of observations in the group, the median is given by the average of the magnitudes of the middle pair of these observations, all the observations having been arranged in order of magnitude as before. In the case of grouped data, the median is calculated assuming that the observations in each class interval are uniformly distributed over that class interval. The formula for calculating the median for a grouped distribution is

$$\text{Median} = L + \frac{(n/2 - F)\,C}{f}$$

where L is the lower limit of the median class, n the total number of observations, F the number of observations up to the median class, f the frequency (number of observations) in the median class and C the interval of median class. The class interval that contains the median is called the median class.

Illustration. Calculation of median for the data given in Table 6.1.

Arranging all the 24 values in ascending order of magnitude, we get the following data:

2.90	3.57	3.73
2.98	3.61	3.75
3.30	3.62	3.76
3.38	3.66	3.76
3.43	3.68	3.76
3.43	3.69	3.77
3.45	3.71	3.84
3.55	3.72	3.88

The 12th value is 3.66 and 13th is 3.68; median is the average of these two.

$$\text{Median} = \frac{3.66 + 3.68}{2} = 3.67 \text{ g\%}$$

Table 6.3 Calculation of Median for the Data of Table 6.2

Protein intake/consumption unit/day (g)	No. of families (frequency)	Cumulative frequency
15–25	30	30
25–35	40	70
35–45	100	170
45–55	110	280
55–65	80	360
65–75	30	390
75–85	10	400
Total	400	

Median class is 45–55 $n = 400$

$$\text{Median} = L + \frac{(n/2 - F) \times C}{f}$$

$$= 45 + \frac{(200 - 170) \times 10}{110} = 45 + 2.73 = 47.73 \text{ g}$$

6.4 MODE

Mode is the most frequently occurring value. In other words, the mode of a group of observations is the value around which the observations tend to be most heavily concentrated. For example, for the data in Table 6.1, the observation 3.76 is most commonly occurring and hence the mode is 3.76. If we have a group of values such as 31, 33, 34, 36, 37, 39, 40 it is apparent that there is no mode. For a moderately asymmetric distribution, the mode can be calculated using the following empirical relationship:

$$\text{Mode} = 3 \text{ median} - 2 \text{ mean}$$

or using the formula

$$\text{Mode} = L_M + \frac{d_1 C}{d_1 + d_2}$$

where L_M = lower limit of modal class

d_1 = frequency in modal class *minus* frequency in the preceding class

d_2 = frequency in modal class *minus* frequency in the succeeding class

C = class interval of modal class

For the data of Table 6.2,

$$\text{Mode} = 45 + \frac{10 \times 10}{10 + 30} = 47.5$$

6.5 POSITION OF AVERAGES

In a frequency distribution, the measures of central tendency—mean, median and mode—occupy some definite relative positions.

Figure 6.1 shows a frequency distribution. The height of each point on the curve from the horizontal axis indicates the frequency of the observation represented by the point on the horizontal axis just below it. The point M on the horizontal axis indicates the arithmetic mean. It can be easily seen that at equal distances on either side of the arithmetic mean, the frequencies are equal. The frequency distribution with this property is called symmetric. It is seen that the frequencies nearer the mean are higher, giving a bell shape to its graph.

The frequency distributions of Figures 6.2 and 6.3 are not symmetrical, the highest frequency (or the mode) does not coincide with the arithmetic mean as in the previous case. Such distributions are said to be asymmetrical or skewed. Figure 6.2 presents a positively skewed distribution and Figure 6.3 a negatively skewed distribution. Note the order in which the averages occur in these two types of distributions.

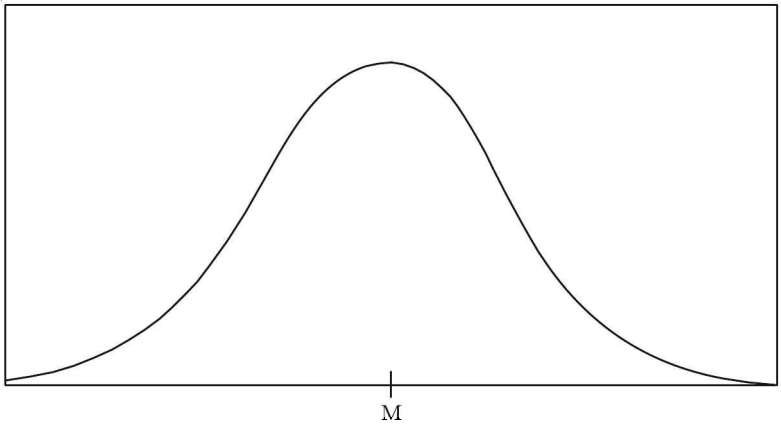


Figure 6.1 Position of averages in symmetric distribution.

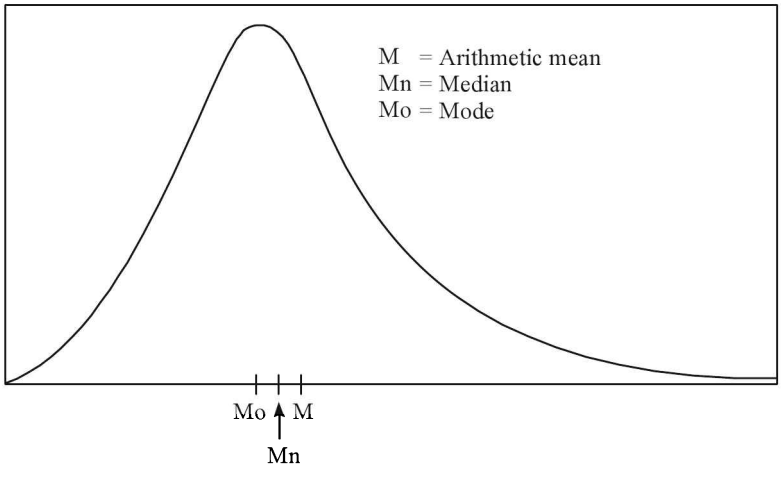


Figure 6.2 Position of averages in positively skewed distribution.

6.6 SELECTION OF THE APPROPRIATE MEASURE OF CENTRAL TENDENCY

When these measures of central tendency are considered as descriptive indices, we are faced with the problem of deciding which one of these to use to characterise a given set of data. The choice will depend upon the nature of the distribution of the observations and the concept of the central tendency which is desirable for our data. Here are a few guidelines:

1. If the data are symmetrically distributed or are approximately symmetrical, any one of these measures may be used because in a symmetrical distribution all these measures give identical values.

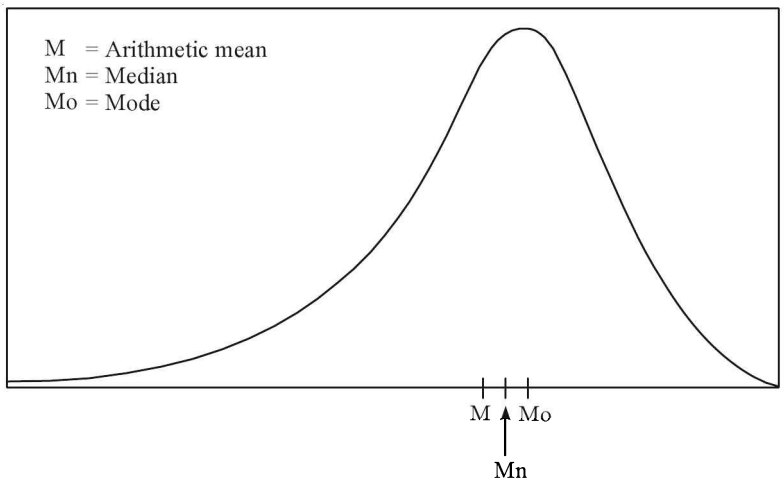


Figure 6.3 Position of averages in negatively skewed distribution.

2. When the distribution of the observations is skewed, the arithmetic mean is usually not suitable. For a positively skewed series, the mean gives a higher value than the other two measures; and for a negatively skewed series, a lower value. It may be preferable to use the median or the mode which is typical.

3. When there are some observations which relatively deviate much more than the others in the series or when heterogeneity is suspected in the series, the median may be used, instead of the mean.

4. When subsequent computations involving a measure are necessary, the arithmetic mean has certain definite advantages.

5. When the concept of 'relative' standing of the individual observations in the group is considered, the use of the median is desirable; whereas the concept of 'typical' observation necessitates the use of the mode.

Sometimes it may be advisable to use two or all three of these measures, since each measure embodies a different concept. The use of any two, mean and median, or mean and mode, will give us an idea of the amount of skewness of the distribution of the series.

There are two other measures of central tendency, geometric mean and harmonic mean for use under certain specific situations.

6.7 GEOMETRIC MEAN

The geometric mean (GM) is usually more suitable as a measure of central tendency when the values change exponentially. If there are only 2 observations, then the GM is the square root of the product of 2 observations. If there are 3 observations, then it is the cube root of the product of the 3 observations. Thus, if there are n observations, the GM will be the nth root of the product of the n observations, viz.,

$$GM = \sqrt[n]{(x_1)(x_2)\ldots(x_n)}$$

The computation is not as formidable as it appears in the above equation, especially when one takes the logarithm of both sides.

Thus, $\log GM = \Sigma (\log x)/n$, which shows that the logarithm of the geometric mean is the arithmetic mean of the logarithms of the individual observations.

Example of geometric mean (GM). The number of bacteria ($\times 10^3$) observed in an experiment at hourly intervals are as follows:

10, 25, 76, 148, 302

The geometric mean of these 5 values will be

$$\sqrt[5]{10 \times 25 \times 76 \times 148 \times 302} = \sqrt[5]{33,968,960} = 61.07$$

The same result can also be obtained by taking the arithmetic mean of the logarithm of these 5 values and then determining the antilog as follows:

Value	Log
10	1.0000
25	1.3979
76	1.8808
148	2.1703
302	2.4800

Total ÷ 5 = 8.9290 ÷ 5 = 1.7858

GM = antilog (1.7858) = 61.07

6.8 HARMONIC MEAN

The Harmonic Mean is used in situations where the reciprocals of the actual values seem more useful to determine the central tendency. For instance, it has been suggested that the sensitivity to detect clusters of observations is increased by measuring the reciprocal of distance rather than using the distance directly. The Harmonic Mean (HM) is thus defined as

$$HM = \frac{1}{(1/n)\,\Sigma\,(1/x)} \text{ or } \frac{n}{\Sigma\,(1/x)}$$

Taking reciprocals of both sides, we get

$$\frac{1}{HM} = \frac{\Sigma\,(1/x)}{n}$$

In other words, the reciprocal of the harmonic mean is the mean of the reciprocals of the individual observations.

Example of harmonic mean (HM). The distances (km) from an industrial establishment of 17 cases of chronic bronchitis are as follows:

0.8, 1.2, 3.2, 1.6, 0.7, 1.1, 2.7, 2.1, 1.3, 0.9, 1.3, 1.5, 1.1, 0.9, 1.8, 2.2, 2.4

Thus,

$$\frac{1}{\text{Harmonic mean}} = \frac{\text{Sum of } (1/x)}{n}$$

$$= \text{Sum of } (1.25, 0.83, 0.31, 0.63, 1.43, 0.91, 0.37, 0.48, 0.77, 1.11, 0.77, 0.67, 0.91, 1.11, 0.56, 0.45, 0.42) \div n$$

$$= 12.98 \div 17$$

$$= 0.7635$$

Therefore, HM = 1.31 km. If we were to use the distances directly, then we would have obtained the result

$$\text{AM} = \frac{26.8}{17} = 1.58 \text{ km}$$

6.9 PERCENTILE

The value below which a given percentage of observations occur is called a *centile* or *percentile*. We have already pointed out in Section 6.3 that the median divides a distribution into two halves. For this reason, the median is also called the *50th percentile*, or *50th centile*. Other percentiles are similarly interpreted, e.g., the 5th or 10th centile of a birth-weight distribution. The percentiles divide the distribution into 100th but sometimes it is more convenient to use quartiles, which divide the distribution into quarters, or deciles which divide it into tenths, etc. Like the 50th percentile (the median), the 25th and 75th percentiles are sometimes given special names—the lower quartile and upper quartile respectively.

Using the data given in Table 6.2, a few of the centiles are computed as follows:

First quartile or 25th percentile

$$P_{25} = L + \frac{(25N/100 - cf)}{f} \times C$$

$$= L + \frac{(N/4 - cf)}{f} \times C$$

$$= 35 + \frac{100 - 70}{100} \times 10 = 35 + \frac{30}{10}$$

$$= 35 + 3 = 38$$

Third quartile or 75th percentile

$$P_{75} = L + \frac{(75N/100 - cf)}{f} \times C$$

$$= L + \frac{(3N/4 - cf)}{f} \times C$$

$$= 55 + \frac{300 - 280}{80} \times 10 = 55 + \frac{200}{80}$$

$$= 57.5$$

Third percenile

$$P_3 = L + \frac{(3N/100 - cf)}{f} \times C$$

$$= 15 + \frac{12 - 0}{30} \times 10 = 15 + \frac{120}{30} = 15 + 4$$

$$= 19$$

First decile or 10th percentile

$$P_{10} = L + \frac{(10N/100 - cf)}{f} \times C$$

$$= 25 + \frac{40 - 30}{40} \times 10 = 25 + \frac{100}{40} = 25 + 2.5$$

97th percentile

$$P_{97} = L + \frac{(97N/100 - cf)}{f} \times C$$

$$= 65 + \frac{388 - 360}{30} \times 10$$

$$= 65 + 9.3 = 74.3$$

It is often most convenient to first draw a cumulative frequency curve or, better still, a percentage cumulative frequency curve and then determine the centiles as shown in Figure 5.4.

6.10 USE OF COMPUTER

In this era of computers, all the measures discussed in this chapter can be obtained using appropriate program. Details are given in Chapter 25.

QUESTIONS

1. Explain the need for the measures of central tendency.
2. Write explanatory notes on:
   (a) Arithmetic mean
   (b) Median
   (c) Mode
   (d) Percentile
3. What are the guidelines for selecting the appropriate measure of location?
4. Find the arithmetic mean, median, and mode for the following values of serum cholesterol levels (mg/dl) of 15 patients:

   220, 260, 240, 200, 210, 250, 150, 160,
   170, 160, 230, 190, 230, 210, 180

CHAPTER

7 Measures of Dispersion

7.1 INTRODUCTION

The fact that we need an average or a measure of central tendency shows that there is variation among the observations. Variation which is another characteristic of a group of observations has to be considered for describing the group more satisfactorily. A single figure for a group relating to its central tendency does not give any idea about the variability of the observations. It can be verified from the following haemoglobin values of two groups of children that they have the same central tendency (12.5), but the amount of spread of observations in each group differs very much.

Group I: 12.1, 12.2, 12.8, 12.9, 12.3, 12.4, 12.7, 12.6, 12.5

Group II: 12.1, 12.3, 11.7, 11.9, 13.1, 13.3, 12.5, 12.9, 12.7

It is often necessary to give an idea of the quantity of such spread, or variability of the observations in a group while characterising that group. For describing this, more than one measure is available. For a group of observations, a suitable measure is selected. The value given by the suitable measure for central tendency and that for dispersion comprise, under many situations, a fairly adequate presentation of a frequency distribution.

In this chapter some of the measures of dispersion—range, interquartile range, mean deviation, standard deviation and coefficient of variation—are defined and their calculations are illustrated. These measures are also called measures of variation.

7.2 RANGE

The range of a group of observations is the interval between the smallest and the biggest observations. The value of the range is dependent only upon the two extreme observations in the group and does not consider the other observations. The occurrence of rare observations in the group greatly influences the value of the range and so it is not considered ideal as a measure of dispersion even though it is used in certain circumstances. This can be calculated easily and also understood by the layman.

Illustration. The lowest value among the 26 observations in Table 7.1 is 10.4 and the highest value, 14.2. Therefore, the range is 10.4 g%–14.2 g%.

Table 7.1 Haemoglobin Values (g%) of 26 Normal Children

11.8	12.9	12.4	13.3	13.8
11.4	12.3	11.7	12.9	12.2
10.4	10.8	12.7	13.2	
11.6	12.0	12.2	14.2	
10.8	10.5	11.6	13.5	
12.2	11.2	12.6	13.0	

For the frequency distribution in Table 7.2 the accurate range cannot be found out but we can approximately give the lowest and the highest values of the class intervals. Thus, the range is 15 g–85 g.

Table 7.2 Protein Intake of 400 Families

Protein intake/consumption unit/day (gram)	No. of families
15–25	30
25–35	40
35–45	100
45–55	110
55–65	80
65–75	30
75–85	10
Total	400

7.3 INTERQUARTILE RANGE

The interquartile range of a group of observations is the interval between the values of the upper quartile and the lower quartile for that group. Upper quartile of a group is the value above which 25% of the observations fall. Lower quartile is the value below which 25% of the observations fall.

This measure gives us the range which covers the middle 50% of the observations in the group. Unlike the range, the value given by this measure is unaffected by the occurrence of rare extreme values, and makes a good measure of dispersion. When the extreme values have not been recorded exactly as in the case of frequency distribution with open-end class intervals, the median and interquartile range can be adopted to measure the central tendency and dispersion respectively.

Illustration. Let us find out the interquartile range for the example given in Table 7.1. Arranging these observations in ascending order of magnitude, we get

10.4	11.2	11.7	12.2	12.6	13.0	13.8
10.5	11.4	11.8	12.2	12.7	13.2	14.2
10.8	11.6	12.0	12.3	12.9	13.3	
10.8	11.6	12.2	12.4	12.9	13.5	

The lower quartile Q_1 is 11.6, i.e., about 25% of the number of observations fall below the value 11.6. The upper quartile Q_3 is 12.9, i.e., nearly 25% of the number of observations are above the value 12.9. Therefore, the interquartile range is 11.6 to 12.9.

7.4 MEAN DEVIATION

The mean deviation is the arithmetic mean of the deviations of the observations from the arithmetic mean ignoring the sign of these deviations.

The mean deviation is based on all observations in the group. It is easy to grasp the meaning of the procedure involved in this measure. However, it is not widely used because of the availability of a more advantageous measure.

The formula for the mean deviation for ungrouped data is

$$\text{Mean deviation} = \frac{\Sigma | x - \bar{x} |}{n}$$

where $| x - \bar{x} |$ indicates the difference between the value of the observation and the arithmetic mean ignoring the sign of the difference. For the grouped data (frequency distribution),

$$\text{Mean deviation} = \frac{\Sigma f | (x - \bar{x}) |}{\Sigma f}$$

where x is the mid-point of the class interval and f is the frequency.

Illustration. Let us find out the mean deviation for the data given in Table 7.1 (Table 7.3).

Table 7.3 Calculation of Mean Deviation

Hb level (g%)	Deviation from arithmetic mean* (without sign)	Hb level (g%)	Deviation from arithmetic mean* (without sign)
11.8	0.4	11.7	0.5
11.4	0.8	12.7	0.5
10.4	1.8	12.2	0.0
11.6	0.6	11.6	0.6
10.8	1.4	12.6	0.4
12.2	0.0	13.3	1.1
12.9	0.7	12.9	0.7
12.3	0.1	13.2	1.0
10.8	1.4	14.2	2.0
12.0	0.2	13.5	1.3
10.5	1.7	13.0	0.8
11.2	1.0	13.8	1.6
12.4	0.2	12.2	0.0
	Total	317.2	20.8

* Mean = 12.2

Total of the deviations from arithmetic mean (without taking into account the sign) = $\Sigma|(x-\bar{x})| = 20.8$.

$$\text{Mean deviation} = \frac{\Sigma|(x-\bar{x})|}{n} = \frac{20.8}{26} = 0.8 \text{ g\%}$$

Let us find the mean deviation for the example given in Table 7.2 (see Table 7.4).

Table 7.4 Calculation of Mean Deviation

Protein intake/ consumption unit/ day (grams)	No. of families	Mid-point of class interval	Deviation of mid-point from mean*	Absolute Deviation fx
Class interval	f	x	$x-\bar{x}$	$f\|(x-\bar{x})\|$
15–25	30	20	– 27.5	825
25–35	40	30	– 17.5	700
35–45	100	40	– 7.5	750
45–55	110	50	+ 2.5	275
55–65	80	60	+ 12.5	1000
65–75	30	70	+ 22.5	675
75–85	10	80	+ 32.5	325
Total	400			4550

* Mean = 47.5

$$\text{Mean deviation} = \frac{825+700+\cdots+325}{400} = \frac{4550}{400} = 11.375$$

7.5 VARIANCE AND STANDARD DEVIATION

The standard deviation is the square root of the average of the squared deviations of the observations from the arithmetic mean. The deviation from mean is considered without its sign in calculating the mean deviation, but in calculating the standard deviation, it is squared. The standard deviation is the most important measure of dispersion. For some frequency distributions there is a relationship between the range and the standard deviation. The standard deviation together with the arithmetic mean can describe a frequency distribution uniquely. The use of the standard deviation will become very clear in a later chapter which deals with sampling distributions.

The standard deviation of the population is usually denoted by σ, and that of the sample by S.

The formula for the standard deviation for an ungrouped data is

$$\text{Standard deviation } (\sigma \text{ or } S) = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} \text{ or } \sqrt{\frac{1}{n}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]}$$

For the grouped data (frequency distribution)

$$\text{Standard deviation } (\sigma \text{ or } S) = \sqrt{\frac{\Sigma f\,(x-\bar{x})^2}{\Sigma f}} \text{ or } \sqrt{\frac{1}{\Sigma f}\left[\Sigma f x^2 - \frac{(\Sigma f x)^2}{\Sigma f}\right]}$$

To get an unbiased estimate of population standard deviation from small samples, divide by $n-1$ instead of by n. This is denoted as 's'. Thus, the formula to compute standard deviation for small samples is

$$s = \sqrt{\frac{\Sigma\,(x-\bar{x})^2}{n-1}} \text{ or } \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$$

The square of standard deviation is called *variance* which can also be used as a measure of dispersion. Calculations of variance and standard deviation are illustrated in Tables 7.5–7.8.

Illustration.

Table 7.5 Calculation of Standard Deviation for the Data of Table 7.1

Serial No.	Haemoglobin values	Deviation from arithmetic mean 12.2	Square of deviation
1	11.8	– 0.4	0.16
2	11.4	– 0.8	0.64
3	10.4	– 1.8	3.24
4	11.6	– 0.6	0.36
5	10.8	– 1.4	1.96
6	12.2	0.0	0.0
7	12.9	0.7	0.49
8	12.3	0.1	0.01
9	10.8	– 1.4	1.96
10	12.0	– 0.2	0.04
11	10.5	– 1.7	2.89
12	11.2	– 1.0	1.00
13	12.4	– 0.2	0.04
14	11.7	– 0.5	0.25
15	12.7	– 0.5	0.25
16	12.2	0.0	0.0
17	11.6	– 0.6	0.36
18	12.6	0.4	0.16
19	13.3	1.1	1.21
20	12.9	0.7	0.49
21	13.2	1.0	1.00
22	14.2	2.0	4.00
23	13.5	1.3	1.69
24	13.0	0.8	0.64
25	13.8	1.6	2.56
26	12.2	0.0	0
Total	317.2	0	25.40

Arithmetic mean is 12.2

$$\text{Standard deviation} = s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}} = \sqrt{\frac{25.40}{25}}$$

$$s = \sqrt{1.016} = 1.01\,\text{g}\%$$

or

$$\text{Variance} = s^2 = 1.016$$

Table 7.6 Calculation of Standard Deviation for Data of Table 7.2

Protein intake/ consumption unit/day (g)	No. of families	Mid-point of class interval	Deviation of mid-point from arithmetic mean*	Squared deviation	Frequency × sq. deviation
Class interval	f	x	$(x - \bar{x})$	$(x - \bar{x})^2$	$f\,(x - \bar{x})^2$
15–25	30	20	– 27.5	756.25	22687.5
25–35	40	30	– 17.5	306.25	12250.0
35–45	100	40	– 7.5	56.25	5625.0
45–55	110	50	2.5	6.25	687.5
55–65	80	60	12.5	156.25	12500.0
65–75	30	70	22.5	506.25	15187.5
75–85	10	80	32.5	1056.25	10562.5
Total	400				79500.0

*Arithmetic mean = 47.5

From this table, we get

$$\Sigma f\,(x - \bar{x})^2 = 79500.0$$

$$\Sigma f = 400$$

Therefore,

$$\text{Standard deviation} = S = \sqrt{\frac{79500.0}{400}} = 14.10\ \text{g}$$

$$\text{Variance} = S^2 = 198.75$$

7.6 ALTERNATIVE METHOD OF CALCULATING STANDARD DEVIATION

This method can be computationally easier using the same values and the alternative formula given.

Illustration (i)

Table 7.7 Calculation of Standard Deviation for Table 7.1

Serial No.	Haemoglobin values (g%)	Square of haemoglobin values
1	11.8	139.24
2	11.4	129.96
3	10.4	108.16
4	11.6	134.56
5	10.8	116.64
6	12.2	148.84
7	12.9	166.41
8	12.3	151.29
9	10.8	116.64
10	12.0	144.00
11	10.5	110.25
12	11.2	125.44
13	12.4	153.76
14	11.7	136.89
15	12.7	161.29
16	12.2	148.84
17	11.6	134.56
18	12.6	158.76
19	13.3	176.89
20	12.9	166.41
21	13.2	174.24
22	14.2	201.64
23	13.5	182.25
24	13.0	169.00
25	13.8	190.44
26	12.2	148.84
Total	317.2	3895.24

$$\text{Standard deviation, } s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2 / n}{n-1}}$$

$$= \sqrt{\frac{3895.24 - \frac{(317.2)^2}{26}}{25}}$$

$$= \sqrt{1.016}$$

$$= 1.01 \text{ g\%}$$

Illustration (ii)

Table 7.8 Calculation of Standard Deviation for Data of Table 7.2

Protein intake/ consumption unit/day (g)	No. of families	Mid-point of class interval		Square of mid-point of class interval	Frequency × square
Class interval	f	x	$f \cdot x$	x^2	$f \cdot x^2$
15–25	30	20	600	400	12000
25–35	40	30	1200	900	36000
35–45	100	40	4000	1600	160000
45–55	110	50	5500	2500	275000
55–65	80	60	4800	3600	288000
65–75	30	70	2100	4900	147000
75–85	10	80	800	6400	64000
Total	400		19000		982000

$$\text{Standard deviation} = \sqrt{\frac{1}{\Sigma f}\left[\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}\right]}$$

$$= \sqrt{1/400\left(982000 - \frac{(19000)^2}{400}\right)}$$

$$= \sqrt{1/400\,(982000 - 902500)}$$

$$= \sqrt{198.75}$$

$$= 14.10 \text{ g}$$

7.7 COEFFICIENT OF VARIATION

The coefficient of variation is the standard deviation expressed as a percentage of the arithmetic mean. When variability of groups of observations is dependent on the average size of the groups and/or when the observations are in different units of measurement, this measure of dispersion is adopted to facilitate comparison of the relative variability in different groups or of different measurements.

$$\text{Coefficient of variation (per cent), C.V.} = \frac{\text{Standard deviation} \times 100}{\text{Mean}}$$

For the example given in Table 7.1, the standard deviation, s = 1.01 and the arithmetic mean $\bar{x}$ = 12.2, the coefficient of variation is (1.01/12.2) × 100 = 8.28%.

For the example given in Table 7.2, the standard deviation, s = 14.10 and the arithmetic mean $\bar{x}$ = 47.5, the coefficient of variation, therefore, is (14.10/47.5) × 100 = 29.68%.

In Table 7.9 is given an example where the coefficient of variation is used for comparison of variability in different characteristics.

Table 7.9 Arithmetic Mean, Range, Standard Deviation and Coefficient of Variation of Certain Organ Measurements* (South Indian males, Age: 21–30 years)

Serial No.	No recorded	Arithmetic mean	Range	Standard deviation	Coefficient of variation
1. Height (cm)	33	164.6	142.2 – 180.3	7.64	4.7%
2. Weight (kg)	33	43.1	22.0 – 55.1	6.48	15.0%
3. Brain (g)	14	1317.0	1100.0 – 2335	296.1	22.5%
4. Heart (g)	33	249.5	110 – 1000	150.8	60.4%
5. Liver (g)	33	1205.0	540 – 2500	376.3	30.2%
6. Spleen (g)	32	367.2	53 – 2700	561.4	152.9%

* Autopsy Statistics, Pathology Department, Christian Medical College Hospital, Vellore, Tamil Nadu.

7.8 THE USE OF COMPUTER

All the measures of dispersion can be obtained using standard package program. Some clues on obtaining them are given in Chapter 25.

QUESTIONS

1. What is the need for measures of dispersion?
2. Write an explanatory note on various measures of dispersion.
3. Explain the uses of coefficient of variation.
4. Find the range, mean deviation, standard deviation and coefficient of variation for the following values of birth weight (kg):

   2.5, 2.8, 2.5, 2.8, 3.3,
   3.5, 3.2, 3.0, 2.9, 3.5

5. The haemoglobin levels of 10 children fed with protein supplement diet are given below.

   11.4, 12.2, 12.3, 12.1, 11.2, 12.7, 12.3, 12.0, 10.2, 13.6

   Another 10 comparable children fed on normal diet had the following haemoglobin levels:

   11.2, 10.4, 11.1, 11.6, 9.7, 10.0, 11.5, 10.5, 10.7, 13.3

   Calculate appropriate measures of central tendency and dispersion for each group, compare them and give your comments.

CHAPTER

8 Probability

8.1 INTRODUCTION

The word 'probable' is used very commonly in everyday language to mean something which is very likely to happen. Such vague and general description of the notion of chance is not of much help to quantify the chance of occurrence of an event. We need a more precise definition to quantify chance. The concept of probability fills this need adequately. This concept is used widely in biological sciences in the presentation of research findings as well as in drawing conclusions.

One function of statistical methods is to provide techniques for making inductive inferences and also for measuring the degree of uncertainty of such inferences. These techniques have their foundations in the concept of probability which had its beginning with games of chance in the 17th Century.

In games of chance, such as spinning a roulette wheel, rolling dice, tossing a coin, drawing a card etc., the outcome of a trial is unpredictable. However, it is recognised that even though the outcome of any particular trial may be uncertain, there is a predictable long-term outcome. It is known, for example, that in a large number of tossings of an ideal (balanced, symmetrical) coin about one-half of the tossings will result in heads. It is this long term predictable regularity of outcome that enables gaming houses to engage in their business.

We find similar situations in physical and biological phenomena involving an element of uncertainty and a long-term regularity. For example, in the science of genetics it is uncertain whether an offspring will be male or female, but in the long run it is known approximately what per cent of offspring will be male and what per cent will be female. The long-term regularity provides us with a measure of the amount of chance to which a particular trial is subject. This measure of chance is denoted as *probability*.

8.2 THE PROBABILITY SCALE

Chance is measured on a probability scale having zero at one end and unity at the other. The top end of the scale marked unity represents absolute certainty. The bottom end of the scale marked zero, represents 'absolute impossibility'. The other points on the probability scale falling between 0 and 1 would indicate the relative chance of occurrence of the outcome.

When we are unable to place any odds between the occurrence and the non-occurrence of an outcome we say that $P = 1/2$, or that the event can happen or not happen with equal odds.

8.3 MEASUREMENT OF PROBABILITY

8.3.1 *a priori* or Classical Probability

Suppose that we toss an ideal coin. What is the probability of getting heads? Since there are only two ways that the coin can fall—heads or tails—and since the coin is well-balanced, one would expect it to fall heads and tails with about equal frequency; hence, in the long run one would expect it to fall heads about one-half of the time and so the probability of the outcome of heads will be $\frac{1}{2}$. This kind of reasoning prompted the following classical definition of probability:

If there are a total of n mutually exclusive and equally likely outcomes of a trial and if n_A of these outcomes have an attribute A, then the probability of A is the fraction n_A/n.

We shall use this definition to find out the probability of outcomes in the following cases:

1. Suppose, we toss a die. What is the probability of 4 coming up?

Since there are six mutually exclusive and equally likely outcomes out of which 4 is only one, the probability of 4 coming up is 1/6.

2. Suppose we toss 2 coins. We can have the following outcomes: both heads, *HH*; one head and the other tail, *TH* or *HT*; and both tails, *TT* (*H* = Head; *T* = Tail). Suppose we want to know the probability of *HH*.

HH being one of the four equally likely outcomes, the probability of obtaining *HH* is $\frac{1}{4}$.

3. Suppose we throw 2 dice and we want the probability of a total of 7 points.

A total of 7 can come in 6 ways (1–6, 2–5, 3–4, 4–3, 5–2, or 6–1). So the numerator will be 6. Since we have 6 sides for each die, the total number of 'equally likely', 'mutually exclusive' outcomes is $6 \times 6 = 36$. So the chance of getting a total of 7 when we throw 2 dice is 6/36 (or 1/6).

4. Suppose there is a box containing 12 red beads and 8 green beads. You shuffle it and pick one bead with your eyes closed. The chance of your getting a red bead is 12/20.

If you go on picking the beads at random, each time putting back the bead you have chosen, then, in the long run, 12 out of 20 times you should get a red bead.

8.3.2 *a posteriori* or Frequency Probability

In the previous sub-section the evaluation of probability was done in some simple cases making use of our intuitive notion of chance. However, in more complex situations the evaluation of probability will have to be based on observational or experimental evidence.

The estimate of probability of a specified outcome based on a series of independent trials is given by

$$\text{Probability} = \frac{\text{The number of times the outcome occurred}}{\text{Total number of trials}}$$

Sometimes this probability is referred to as *statistical probability*, frequency or empirical probability or a posteriori probability, i.e., after the event.

For example, if we want to know the probability of success of a surgical procedure, a review of past experience of this surgical procedure under similar conditions will provide the data for estimating this probability. The longer the series we have, the closer the estimate would be to the true value.

8.4 LAWS OF PROBABILITY FOR INDEPENDENT EVENTS

There are two important laws of probability which are useful in finding out probabilities in complex situations where the events concerned are independent.

8.4.1 Addition Law

If an event can occur in any one of several mutually exclusive ways, the probability of that event is the sum of the individual probabilities of the different ways in which it can occur.

For example, when we toss a die, what is the probability of getting 2 or 4 or 6?

The probability of 2 = 1/6

The probability of 4 = 1/6

The probability of 6 = 1/6

Probability of 2 or 4 or 6 is

$$\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{3}{6}=\frac{1}{2}.$$

8.4.2 Multiplication Law

The probability of the simultaneous occurrence of 2 or more independent events is the product of the individual probabilities.

For example, in tossing 2 coins

$$\text{Probability of heads-in one coin} = \tfrac{1}{2}$$

$$\text{Probability of heads-in another coin} = \tfrac{1}{2}$$

$$\text{Thus, probability of heads in both coins} = \tfrac{1}{2}\times\tfrac{1}{2} = \tfrac{1}{4}$$

8.5 CONDITIONAL PROBABILITY

In some situations the chance of occurrence of a particular event depends on some other event. The multiplication law explained above is not applicable in case of *dependent* events. For example, the chance that a patient with some disease survives the next year depends, of course, on his having survived to the present time and the current status of his disease. Such probability is called Conditional Probability.

Let $P(A)$ represent the probability of occurrence of event A, and $P(B)$ that of event B. Let $P(AB)$ represent the probability of the simultaneous occurrence of events A and B.

Then, by definition, the probability that event A occurs, given that event B has already occurred, $P(A/B)$, is given by

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

The general rule of multiplication in its modified form in terms of conditional probabilities becomes

$$P(A \text{ and } B) = P(B) \times P(A/B) \text{ or } P(A) \times P(B/A).$$

Example. From past experience with the illnesses of his patients, a doctor has gathered the following information in a population:

5% feel that they have cancer and do have cancer;

45% feel that they have cancer and don't have cancer;

10% do not feel that they have cancer and do have it; and the remaining

40% feel that they do not have cancer and really do not have it.

Denoting the events as

A when the patitent feels he has cancer, and

B when the patient has cancer, we have

$$P(AB) = 0.05$$

$$P(A) = 0.5$$

$$P(B) = 0.15$$

The probability that a patient has cancer, given that he feels he has it, given by

$$P(B/A) = \frac{P(AB)}{P(A)} = \frac{0.05}{0.5} = 0.1$$

The probability he feels he has cancer, given that he does have it is given by

$$P(A/B) = \frac{P(AB)}{P(B)} = \frac{0.05}{0.15} = 0.33$$

8.6 BAYES' THEOREM

Usually the physician knows the *conditional probability* of a particular symptom (or positive test) for a particular disease. However, it is important that he knows the conditional probability of the disease for an individual patient, given the particular symptom (or positive test). A theorem attributed to Thomas Bayes (1702–61) provides the means to derive the latter probability from the former probability. This theorem is illustrated by an example.

An example to develop Bayes' theorem. This example concerns bacteriuria, and pyelonephritis in pregnancy. Suppose it is known that roughly 6 per cent of pregnant women attending a prenatal clinic at a large urban hospital have bacteriuria (bacteria in the urine).

Consider the two events: A a pregnant woman has bacteriuria, and $\bar{A}$ she does not have bacteriuria.

Since A and $\bar{A}$ are mutually exclusive and complementary

$$P(A) = 0.06, \qquad P(\bar{A}) = 1 - 0.06 = 0.94$$

Suppose it is further known that 30 per cent of bacteriuric and 1 per cent of non-bacteriuric pregnant women proceed to develop this disease.

Using B to denote the occurrence of pyelonephritis, and $\bar{B}$ its absence, then

Notation	*Event*	*Probability*
B/A	Pyelonephritis given that the pregnant woman was bacteriuric	0.30
$B/\bar{A}$	Pyelonephritis given that the pregnant woman was non-bacteriuric	0.01

1. With these definitions consider the following probability questions:

(a) What is the chance of a pregnant woman having both bacteriuria and pyelonephritis?

Using multiplicative law:

$$P(A \text{ and } B) = P(B/A)\,P(A) = (0.30)\,(0.06) = 0.0180$$

(b) What is the chance of a pregnant woman not having bacteriuria but having pyelonephritis?

Using multiplicative law:

$$P(\bar{A} \text{ and } B) = P(B/\bar{A})\,P(\bar{A}) = (0.01)\,(0.94) = 0.0094$$

2. What is the chance of pyelonephritis?

In this particular example, pyelonephritis can occur in two mutually exclusive ways, with or without bacteriuria. Hence, application of the additive law to the probabilities determined in 1(a) and 1(b) gives

$$\begin{aligned} P(\text{pyelonephritis}) &= P(B) \\ &= P(A \text{ and } B) + P(\bar{A} \text{ and } B) \\ &= 0.0180 + 0.0094 = 0.0274 \end{aligned}$$

3. Finally, with the knowledge that a pregnant woman has developed pyelonephritis, what is the chance she had been bacteriuric?

Using the notation developed, the question is: What is the probability of A/B, i.e., the presence of bacteriuria given that the pregnant woman has pyelonephritis. From the alternative form for the multiplicative law of the preceding section and the answers to 1(a) and 2 above:

$$P(A/B) = P(A \text{ and } B) / P(B) = 0.0180/0.0274 = 0.6569$$

In other words, if a pregnant woman has developed pyelonephritis there is a 65.7 per cent chance that she had been bacteriuric.

From the above illustration of Bayes' theorem, we find that knowledge of the chances of pyelonephritis when bacteriuria is present and when it is absent along with knowledge of the relative frequency of bacteriuria among pregnant women has permitted the determination of a conditional probability, namely the chance that the pregnant woman had been bacteriuric given that she has developed pyelonephritis.

Substituting letters for numbers and working backward from the expression in the answer to Question 3, it may be seen that

$$P(A/B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A \text{ and } B)}{P(A \text{ and } B) + P(\bar{A} \text{ and } B)}$$

$$= \frac{P(B/A)\,P(A)}{P(B/A)\,P(A) + P(B/\bar{A})\,P(\bar{A})}$$

The last expression is the usual formulation of Bayes' theorem.

8.7 APPLICATION OF BAYES' THEOREM IN DETERMINING DIAGNOSTIC EFFICACY

Bayes' theorem has been used frequently to evaluate the performance of diagnostic and screening tests. Assessing a new test begins with the identification of a group of subjects known to have the disorder or disease of interest, using an accepted reference test known as the gold standard. Let us denote subjects with the disease as D+ and those without as D–. Further, let us denote subjects who are positive for the new diagnostic test as T+ and those negative as T– as shown in Table 8.1.

Table 8.1 Subjects Classified according to Disease Status and Test Results

	Disease		
Test	+	–	Total
+	a	b	$a + b$
–	c	d	$c + d$
Total	$a + c$	$b + d$	N

We can describe the Test performance in terms of the following characteristics:

(i) Sensitivity. This is the probability that a diseased individual will have a positive test result, and hence, the true positive rate (TPR) of the test. In conditional probability notation, sensitivity is written P (T+/D+). From the Table, *Sensitivity* = $(a/a + c)$.

(ii) Specificity. This is the probability that a disease-free individual will have a negative test result, and is the true negative rate (TNR) of the test. In conditional probability notation, specificity is written as P (T–/D–). From Table 8.1, *Specificity* = $(d/b + d)$.

(iii) False negative rate (FNR). This is the probability that a diseased individual will have a negative test result. In conditional probability notation, FNR is written P (T–/D+), and from Table 8.1, FNR = $(c/a + c)$.

(iv) False positive rate (FPR). This is the probability that a disease-free individual will have a positive test result. In conditional probability notation, FPR is written as P (T+/D–), and from the Table, FPR = $(b/b + d)$.

Two other indicators, known as likelihood ratios and very useful in clinical practice, can also be calculated. These are also known as *positive* and *negative predictive values*.

(v) Predictive value positive (PVP). This is the probability that an individual with a positive test result has the disease. PVP is also known as the posterior probability, or post-test probability of disease. In conditional probability notation, PVP is P (D+/T+) and from the table, PVP = $(a/a + b)$.

(vi) Predictive value negative (PVN). This is the probability that an individual with a negative test result does not have the disease. In conditional probability notation, PVN is written as P (D–/T–), and from the table, PVN = $(d/c + d)$.

Published reports of the performance of a diagnostic test typically contain values for sensitivity and specificity, as well as an estimate of the test's predictive value for the particular study setting. The physician who wishes to use the test in a different practice setting must take into account the effect of any differences between the study population and the practice population on the predictive ability of the test. It is especially critical to assess the impact of any alteration in disease prevalence. If prevalence can only be expressed as a range of plausible values, rather than a specific estimate, a sensitivity analysis (i.e., an evaluation of the predictive value of the test over this range) is indicated.

The predictive values, PVP and PVN can also be calculated using Bayes' rule as follows:

$$\text{PVP} = P(\text{D+/T+}) = \frac{P(\text{D+})\,P(\text{T+/D+})}{P(\text{D+})\,P(\text{T+/D+}) + P(\text{D–})\,P(\text{T+/D–})}$$

$$\text{PVN} = P(\text{D–/T–}) = \frac{P(\text{D–})\,P(\text{T–/D–})}{P(\text{D–})\,P(\text{T–/D–}) + P(\text{D+})\,P(\text{T–/D+})}$$

The formula for PVP can be rewritten as

$$\text{PVP} = \frac{(\text{Prevalence})\,(\text{Sensitivity})}{(\text{Prevalence})\,(\text{Sensitivity}) + (1 - \text{Prevalence})\,(1 - \text{Specificity})}$$

and the formula for PVN may be rewritten as

$$\text{PVN} = \frac{(1 - \text{Prevalence})\,(\text{Specificity})}{(1 - \text{Prevalence})\,(\text{Specificity}) + (\text{Prevalence})\,(1 - \text{Sensitivity})}$$

In general, the more sensitive a test, the better its PVN; and the more specific a test, the better its PVP.

Sometimes, it is useful to show graphically the relationship between sensitivity and specificity for a diagnostic test. Such a diagram is known as the *receiver operator characteristic* (*ROC*) *curve*, and provides a simple tool for applying the predictive value method to the choice of a positivity criterion.

The ROC curve is constructed by plotting the true positive rate (sensitivity) against the false positive rate (1 – specificity) for several choices of the positivity criterion as shown in Figure 8.1.

The ROC curves can also be used to compare two diagnostic tests. The area under the curve represents the overall accuracy of a test; the larger the area, the better the test (see Figure 8.1).

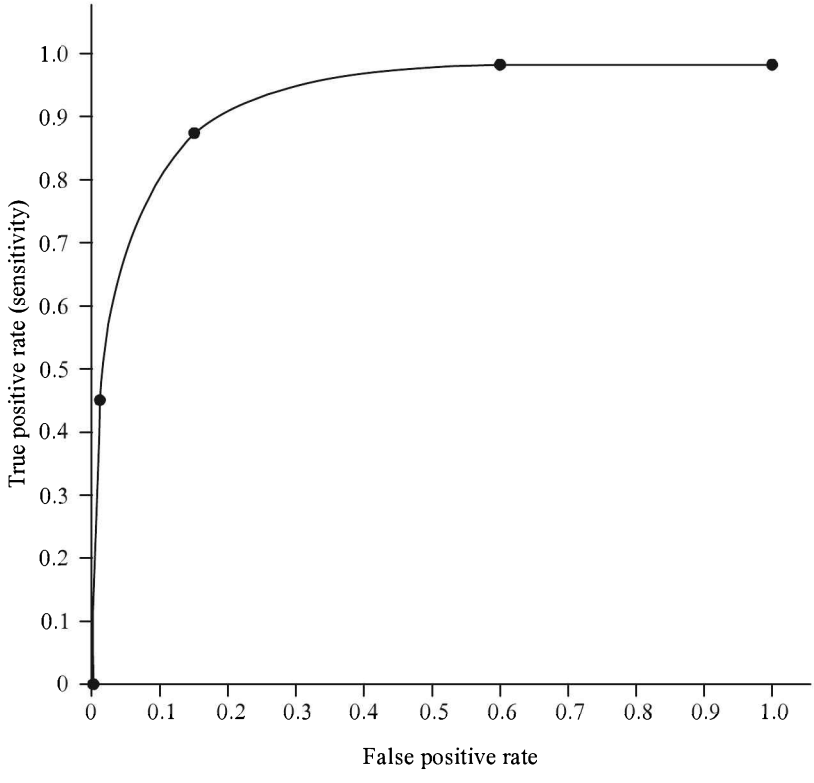


Figure 8.1 Receiver operator characteristic curve.

8.8 OTHER APPLICATIONS OF PROBABILITY

Since its humble beginning at the gambling tables in the seventeenth century, the probability theory has been developed and employed to treat and solve many weighty problems. It is the foundation of the classical decision procedures of estimation and testing of hypothesis. The probability theory has enriched and enlarged our knowledge of several biological phenomena. Two main areas of application of the probability theory are:

1. Description of empirical phenomena through specific probability distribution or models; and

2. Predictions of events based on pertinent theoretical (probabilistic, stochastic) considerations.

Examples of specific uses may be found in the descriptions of population dynamics and estimates of population growth, determination of risks and competing factors in fertility or mortality, description of the course of epidemics and epidemic processes, and in the documentation of clinical and biologic processes such as those encountered in human physiology, pathology or therapeutics. Thus, population growth models based on certain assumptions were postulated regarding the initial population levels, rates of input (or birth), and decay (or death). Similarly, physiologists have described nerve conduction in terms of probability of movement of ions and their concentration at the relevant place on the axon membrane. Pharmacologists have followed up these ideas by describing in probability terms the actions of drugs, especially by invoking receptors connected with the pharmacological response. The clinical process refers to the passage of an individual through the various stages of a disease in probability terms. Disease is a dynamic pathological process reversible in its early stages, e.g., scientists have built models of the evolution of pulmonary tuberculosis by defining various stages and the differential probabilities of movement from one stage to another. Others have tried to describe the mechanism of tumour growth using probabilities for continued genotype variation, adaptation, and selection within the genetic mosaic of cell clones constituting a tumour.

Mendel (1866) postulated that there were certain entities, now called genes, responsible for flower colour and passed on from the plants to their progeny. The present example involves two types of genes, which we may denote by A and a. Each plant has two genes so that there are three kinds of plants, or genotypes: AA, Aa, and aa. A plant (AA), both of whose genes are of type A, will have red flowers. A plant (aa), whose genes are of type a, will have white flowers. Finally, a plant (Aa) with one gene of each type will have pink flowers. Each offspring obtains one gene from each parent to make up its own pair. When we cross an AA plant with another AA plant, the offspring must also be AA. Similarly $(aa) \times (aa) \rightarrow (aa)$, and $(AA) \times (aa) \rightarrow (Aa)$. The varying results are now also explained: in the $(AA) \times (Aa)$ cross, the offspring is sure to get A from the first parent, but may receive either A or a from the other, and thus may be either AA or Aa. Similarly, the $(Aa) \times (Aa)$ cross may produce each of the three types.

To explain the suitable frequencies, we now introduce probability assumptions into the model. We suppose that:

1. from an (Aa) parent, an offspring has probability $\frac{1}{2}$ of receiving each of the genes A and a; and

2. the genes obtained by an individual from his two parents are unrelated.

Breeding experiments also justify another assumption of unrelatedness:

3. the genes passed on by a parent to different offsprings are unrelated.

With these simple assumptions, all the Mendelian frequencies are explained. Thus, in an $(Aa) \times (Aa)$ cross, the probability is $\frac{1}{4}$ that the offspring will be (AA), since there is probability $\frac{1}{2}$ of obtaining the A gene from each parent, and by assumption (2) these probabilities are to be multiplied. Table 8.2 shows the nine possible crosses, and the probabilities of each of the three possible offsprings in each case.

Table 8.2 Probabilities of Genotypes

Father \ Mother	*AA*		*Aa*		*aa*	
AA	*AA*	1	*AA*	$\frac{1}{2}$	*AA*	0
	Aa	0	*Aa*	$\frac{1}{2}$	*Aa*	1
	aa	0	*aa*	0	*aa*	0
Aa	*AA*	$\frac{1}{2}$	*AA*	$\frac{1}{4}$	*AA*	0
	Aa	$\frac{1}{2}$	*Aa*	$\frac{1}{2}$	*Aa*	$\frac{1}{2}$
aa	*aa*	0	*aa*	$\frac{1}{4}$	*aa*	$\frac{1}{2}$
	AA	0	*AA*	0	*AA*	0
	Aa	1	*Aa*	$\frac{1}{2}$	*Aa*	0
	aa	0	*aa*	$\frac{1}{2}$	*aa*	1

The probability theory has become an indispensable tool for all types of formal studies that involve uncertainty. It should be noted that the concept of probability is employed not only for various types of scientific investigations, but also for many problems in everyday life. It is difficult to explain satisfactorily the basic idea of probability and its applications without recourse to some algebra and calculus. Also, in figuring out the 'total' of probabilities, we need to be familiar with the idea of permutations and combinations. Combinations concern forming of groups without any reference to the order within the group, while in permutations, we take the arrangement into account. We shall not go into this aspect here and the reader who is interested may refer to other textbooks.

Two other specific applications in clinical medicine are briefly given.

8.8.1 Decision Analysis

Physicians are increasingly faced with the challenge of making clinical decisions on the basis of a complex and often conflicting array of diagnostic, therapeutic, financial, and social considerations. This analytical procedure selects from among available diagnostic or therapeutic options, based on the probability and predetermined value (utility) of all possible outcomes of those options. The decision analysis procedure consists of six sequential steps:

1. Construct a decision tree, which depicts a map of all relevant courses of action and their associated outcomes.

2. Determine and assign probabilities. The probability of occurrence associated with each chance node may be derived from estimates or from literature values.

3. Assign utilities to each potential outcome. A utility is a numerical estimate of the worth or value of a given outcome.

4. Determine the expected utility. The expected utility of each potential course of action is a function of both the probability of the outcome and its utility.

5. Choose the course of action with the highest expected utility.

6. Evaluate the resistance of the chosen course of action to changes in probabilities and utilities. A clinically valid decision is resistant to credible changes in the assigned probabilities and utilities. If small variations in these values alter the decision, no course of action is clearly preferable to the others and the decision is a toss-up.

8.8.2 Genetic Counselling

The second practical use of probability is in the area of human genetics and genetic counselling: Genetic counselling represents a unique type of diagnostic assessment, which utilizes the rules of probability and a knowledge of the family history of a particular genetic disease to predict future occurrences of the disorder. Bayes' Rule is particularly useful in situations where estimates of the likelihood of inheritance are revised on the basis of new knowledge, e.g., in predicting the probable disease status of future siblings following the birth of an affected child. The probability theory is also used in the outcomes of mating patterns, mutations and selection of genes, the discussion about the gender of an unborn baby, etc.

The application of the basic concepts of probability in the formation of probability distributions, that are the foundations of statistical inference, are described in the next chapter.

QUESTIONS

1. Define probability and give two examples.

2. Explain how probability is measured.

3. State the laws of probability. Explain with examples.

4. (a) If the probability of low birth weight is 0.3, what is the probability of normal birth weight?

(b) If it is known that 30% and 35% pregnant women have hypertension and anaemia respectively, what is the probability that a pregnant woman has

(i) either hypertension or anaemia?

(ii) both hypertension and anaemia?

CHAPTER

9 Probability Distributions

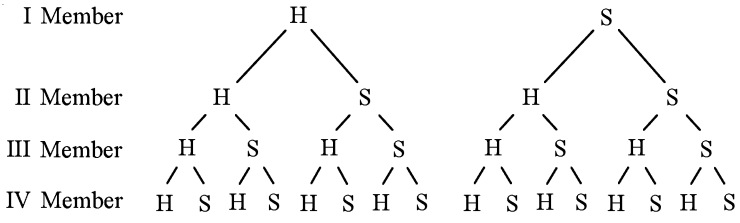
9.1 INTRODUCTION

For making inferences from a sample, we found that we have to think in terms of the part played by chance. This is done by considering the sampling distribution and calculating the probabilities. It would be much easier and more time saving if these probabilities are available for all sampling distributions. All sampling distributions are not of the same type. They differ according to the type of the characteristic studied, the type of population and the size of the sample. For each type of situation, a sampling distribution can be formed using a mathematical model. The sampling distribution formed by using the mathematical model is called theoretical distribution. The sampling distribution formed by actually taking the sample from the population, as done in the last section, is called observed sampling distribution. In several situations the theoretical sampling distributions are very close approximations of the observed sampling distributions. When the theoretical distribution is known, we need not form an observed sampling distribution. From the theoretical distribution the necessary evaluation of a sample can be done. Mathematical models have been developed, the required probabilities calculated and made available in tables for certain types of families of distributions. Three such families which are fundamental in the theory of statistics are Binomial distribution, Poisson distribution and Normal distribution.

9.2 BINOMIAL DISTRIBUTION

Very often, we are interested in knowing what proportion of individuals in a population possess a particular character; for example, the proportion of persons of a locality who are sick at a particular point of time. As discussed in the next chapter, an estimate of this proportion is calculated on the basis of a suitably drawn sample from this population and the corresponding sampling distribution. In this type of problem the sampling distribution is given by a theoretical frequency distribution known as *Binomial distribution.*

Let us consider the following example. In a morbidity survey in a village, it is found that the proportion of sick persons is 40%. A sample of 4 persons can be any one of the five types having no sick person in the sample or having 1, 2, 3 or 4 sick persons. Assuming random sampling, there are sixteen ways in which we will get such a sample as shown in the diagram:



H stands for a healthy person and S for a sick person. The first member selected may be either a healthy person or a sick person. The second also may be either a healthy or a sick person and so on. All such possibilities are listed in this diagram. If we follow each line starting from either the first H or S, we get one composition. So, we get 16 such compositions. The probability of a sick person being chosen is 0.4 and the probability of a healthy person being chosen is 0.6. The probabilities of the various compositions are obtained by using the addition and multiplication theorems on probability (Table 9.1).

Table 9.1 Calculation of Probabilities for a Binomial Distribution

Composition of the sample	Ways of occurrence	Probability of occurrence	Probability of getting such sample
No sick persons	H H H H	$0.6 \times 0.6 \times 0.6 \times 0.6$	0.1296
One sick and three healthy persons	S H H H H S H H H H S H H H H S	$0.4 \times 0.6 \times 0.6 \times 0.6$ $+ 0.6 \times 0.4 \times 0.6 \times 0.6$ $+ 0.6 \times 0.6 \times 0.4 \times 0.6$ $+ 0.6 \times 0.6 \times 0.6 \times 0.4$	0.3456
Two sick and two healthy persons	S S H H S H S H S H H S H S S H H S H S H H S S	$0.4 \times 0.4 \times 0.6 \times 0.6$ $+ 0.4 \times 0.6 \times 0.4 \times 0.6$ $+ 0.4 \times 0.6 \times 0.6 \times 0.4$ $+ 0.6 \times 0.4 \times 0.4 \times 0.6$ $+ 0.6 \times 0.4 \times 0.6 \times 0.4$ $+ 0.6 \times 0.6 \times 0.4 \times 0.4$	0.3456
Three sick and one healthy person	S S S H S S H S S H S S H S S S	$0.4 \times 0.4 \times 0.4 \times 0.6$ $+ 0.4 \times 0.4 \times 0.6 \times 0.4$ $+ 0.4 \times 0.6 \times 0.4 \times 0.4$ $+ 0.6 \times 0.4 \times 0.4 \times 0.4$	0.1536
All sick persons	S S S S	$0.4 \times 0.4 \times 0.4 \times 0.4$	0.0256

In the last column of the above table are shown the probabilities of getting the five types of samples when the proportion of sick in the population is 40%. This type of distribution is called *Binomial distribution*. The individual probability of each type of sample is given by the formula

$$\binom{n}{x} p^x q^{n-x}$$

where x is the number of occurrences of the event, p the probability of occurrence of the event, q the probability of non-occurrence of the event and n the size of the sample.

$\binom{n}{x}$ indicates the number of ways in which x occurrences can appear in a sample of size n.

9.3 POISSON DISTRIBUTION

There are situations in which the number of times an event occurs can be counted but the number of times, the event did not occur cannot be counted. Actually, the latter would be meaningless. For example, the amount of radioactive fallout during a certain duration, following a nuclear explosion can be measured and it is meaningless to think of the amount of nuclear dust that has not fallen during this period. The number of goals scored by a football team in a series of matches can be counted but not the number of goals not scored. Similarly, the number of spells of diarrhoea observed in a group of infants over a predetermined period can be counted but not the number of spells that did not occur. The probability of observing one spell, two spells, etc., in a given sample in such cases, can theoretically be found out by the use of Poisson distribution. The probabilities are generated by the mathematical function

$$\frac{e^{-m}m^x}{x!}$$

where x is the number of times the event occurs, m is the mean of the distribution and $x! = 1 \times 2 \times 3 \times 4 \cdots \times x$. If only the mean is known, all the frequencies can be found out theoretically.

Example. The number of yeast cells is counted with the help of a haemocytometer. The number of cells found in each tiny square in the haemocytometer is given in Table 9.2.

Table 9.2 Example of a Poisson Distribution—Number of Yeast Cells Counted in 400 Squares of Haemocytometer

No. of yeast cells in the square	No. of squares		Theoretical probability of getting the given number of cells
	Observed	Theoretical (given by Poisson distribution)	
1	2	3	4
0	103	106	0.2650
1	143	141	0.3525
2	98	93	0.2325
3	42	41	0.1025
4	8	14	0.0350
5	4	5	0.0125
6	2		
Total	400	400	1.0000

In Column 2, the actual numbers of squares found with the given number of cells are given. In Column 3, the frequencies generated by the mathematical formula are given. Column 4 gives the probabilities given by the formula.

9.4 NORMAL DISTRIBUTION

The Binomial and Poisson distributions deal with the occurrence of distinct events such as the number of sick, number of accidents, number of cells, number of coins, etc. There are many situations in which the quantity that is studied is continuous in magnitude. Body temperature, intelligence quotient, height, weight, etc., are examples of quantities that vary continuously. For such characteristics many populations, and also their sampling distributions are very close to a pattern of frequency distribution known as *Normal* (*Gaussian*) *distribution*. The mathematical function that generates the probabilities is

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

where x is the variable in question (characteristic studied), μ is the mean and σ is the standard deviation. An example of normal distribution is given in Table 9.3 and some of the interesting properties are enumerated.

Table 9.3 Example of a Normal Distribution—Distribution of 1000 Men in a Village According to Their Height

Height (in inches)	No. of men of given height			
61–62	2			
62–63	5			
63–64	17			
64–65	43			
65–66	86			
66–67	152			
67–68	193	690	955	997
68–69	197			
69–70	148			
70–71	91			
71–72	45			
72–73	16			
73–74	4			
74–75	1			
Total	1000			

The graphical presentation of the data as shown in Figure 9.1 follows a normal distribution not entirely but fairly closely.

The characteristics of an ideal normal distribution are as follows:

1. The curve of the distribution is bell shaped and symmetrical around the mean of the distribution. We can define symmetry as that property where the total

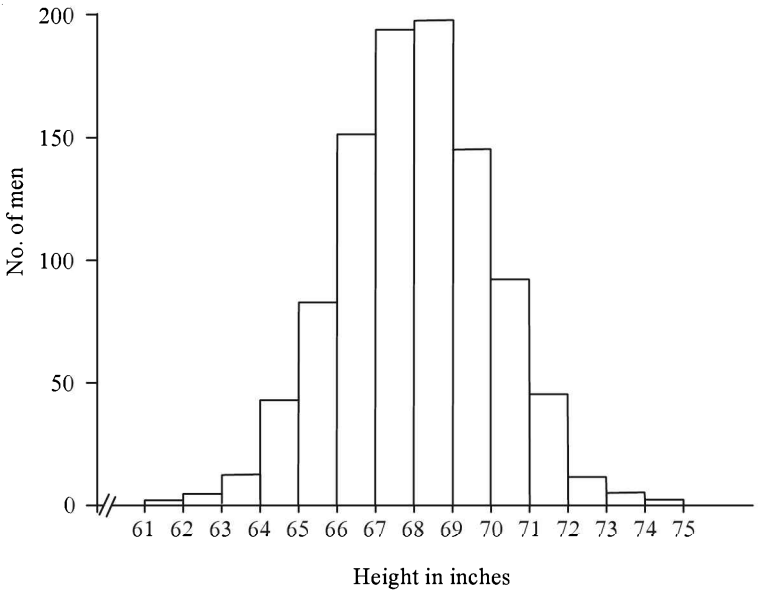


Figure 9.1 Histogram of data given in Table 9.3.

frequencies on either side of the mean value are the same; also the frequencies at equal distances from the mean on either side are the same.

2. The mean, median and mode of the distribution coincide.

3. If we consider a certain number of standard deviations from either side of the mean, we can theoretically state the proportion of frequencies that lie within these limits. To be more specific, we have about 68.27% of the observations lying within one standard deviation on either side of the mean, 95.45% of the observations lie within 2 standard deviations from either side of the mean and about 99.73% of the observations lie within 3 standard deviations from either side of the mean. From the example that we have presented, we notice that nearly 1/3 of the observations are outside one standard deviation of the mean. This corresponds to the complement of the proportion within one standard deviation. Similarly, if we consider the number of observations outside 2 standard deviations, we notice that it is 100% – 95% = 5%, and also only three in a thousand differ from the mean by more than three times the standard deviation. Thus, in a normal distribution, values that differ from the mean by more than twice the standard deviation are fairly rare; for only about 1/20 of the observations will do so. The values that differ from the mean by more than three times the standard deviation are very rare for only about one in 370 observations will do so. We will consider the application of normal distribution in a later chapter.

The normal distribution plays a very predominant part in medical and biological research. In the first place, many populations encountered in the course of research in many fields have a normal distribution to a good degree of approximation. For example, when measurements are made of some biological phenomena such as

heights of men, the frequency distribution will often plot a symmetrical bell-shaped curve. Not all biological measurements follow a normal distribution, but many do, and we could make use of this particular feature in the statistical theory. The other consideration which favours the normal distribution is the fact that sampling distributions of the statistics calculated from samples taken from the parent (population) normal distribution are, in most of the cases, known. The third consideration, of course, is that we have a mathematical basis for calculation of various descriptive measures. When the distribution is normal, the mean and standard deviation describe the distribution completely. If the distribution is approximately normal, no essential details are lost by considering it to be normal. Particular use of this normal distribution is made whenever we have a large sample since many other theoretical distributions approximate to a normal curve. A table showing the various frequencies under the normal curve to various multiples of standard deviation is shown in Appendix 4.

9.5 '*t*' DISTRIBUTION

The '*t*' distribution, also known as Student's *t*-distribution, is a sampling distribution derived from the parent normal distribution, introduced by W.S. Gosset and developed by R.A. Fisher. The statistics of small samples were revolutionized by this distribution.

Take a sample (of size 10 for example) from the 500 Hb values given in Appendix 1 and calculate the mean. Calculate the quantity '*t*' which is given by the equation

$$t = \frac{|\bar{x} - \mu|}{s/\sqrt{n}}$$

where $\bar{x}$ is the mean of the sample; μ the mean of the population; n the size of the sample; $s/\sqrt{n}$ is the estimate of the standard error of mean and s is given by the formula

$$\sqrt{\frac{\text{Sum of squares of deviations of observations from mean (sample)}}{n-1}}$$

If many samples of the same size are taken from the normal population then the quantity t (both positive and negative values) will follow this distribution.

This *t*-distribution is also a symmetric distribution about its mean. Like in all symmetric distributions, mean, median and mode coincide. Unlike the normal distribution, the shape of *t*-distribution varies according to the size of the sample or according to the degrees of freedom. The size of the sample minus the number of parameters estimated is called degrees of freedom. In the present case, only one quantity is used as an estimate. Therefore, the number of degrees of freedom is $n - 1$. The number of degrees of freedom is also defined as the number of independent members in the sample. When the size of the sample is large, the *t*-distribution closely approximates the normal distribution.

9.6 OTHER DISTRIBUTIONS

Many other distributions are also developed from the parent normal distribution, e.g., '*F*' distribution, χ^2 distribution, etc. These also have practical utility. Each one of these distributions has a mathematical formula or theoretical basis from which the probabilities are computed. Tables which give such probabilities of the occurrence of the event are available in standard forms in several textbooks. These tables are widely used in tests of significance which are described in Chapter 12.

QUESTIONS

1. Write a brief note on probability distributions.
2. Give two examples for each of the following:
   (a) Binomial distribution
   (b) Poisson distribution
3. State the characteristics of a normal distribution.
4. What are the uses of normal distribution?

CHAPTER

10 Sampling

10.1 INTRODUCTION

Most of the investigations in medicine are carried out on a small number of subjects or individuals and the results are generalised to the class of all such subjects. For instance, a study of the effect of insulin on blood sugar is carried out on a group of diabetic patients in a ward. The conclusion drawn from the study is assumed to be true for the diabetic patients in the ward, for all such patients in the hospital, and believed to be applicable to all diabetic patients in that town and also proclaimed to be true for all patients in the world. At each stage of generalisation the area covered is larger than the previous one. Similarly, the superiority of one type of surgical operation over other forms of treatment for treating peptic ulcer is observed in a few scores of patients by observing them for a certain number of years and the results are assumed to be true for all peptic ulcer patients in the hospital and also for all such patients yet to come. Again, the part played by an infectious disease on hearing is studied on a few pretested persons and the conclusions are extended to the hearing of all persons affected by that particular disease. These are some examples where generalisations are made from a particular group; in other words, the inference derived from the 'study' group is generalised to all such groups. This type of inference is called inductive inference and is used very often in the world of scientific research. At every stage of generalisation, the research worker should pause and think whether he is justified in doing so.

10.2 DEFINITIONS

The small group that is chosen for study is called a sample and the whole group which it is believed to represent is called population. More generally, the population and sample are defined thus: the totality or aggregate of all individuals with the specified characteristic is a *population* (the word 'universe' is also synonymously used) and a group of individuals chosen from that population is a *sample*. The number of individuals or observations in the sample is termed the *sample size*. Always we study the sample but it is the population whose characteristic we seek to know. *Sampling* refers to the choosing of a sample from a population.

10.3 TYPES OF POPULATION

The primary object of any investigation or study is to find out or confirm certain facts about the population. Before starting the investigation one should specify

what characters are to be studied. Very often one is interested in detailed study of the population. Limited by time, money and personnel one should think realistically, and clearly put forth the characteristics that are to be studied. At this stage, it is important to specify the population that is to be studied. If it is the population of patients, mention is to be made, among other things, of the type of patients and the stage of disease. Clear specification of the population will help in eliminating any fallacies which may affect the interpretation of findings.

There are three types of population: finite, infinite and hypothetical. A finite population, however large it may be is one in which there are a finite number of members. The number of citizens in a country, the number of leprosy patients in a community, and the number of students in a university are examples of a finite population.

If a population contains an infinite number of members, it is called an infinite population. For instance, all possible Hb values in a given interval (Hb can take any value in a given interval); all possible heights within the range 150 cm–160 cm—an infinite number of values of heights can be spotted out in this range.

A hypothetical population is one which is assumed for theoretical purposes. Suppose, a few guinea pigs are given vitamin A-deficient diet and they are watched for vitamin A deficiency symptoms within a predetermined period. The result observed is generalised to all such groups of guinea pigs fed on a vitamin A-deficient diet. All such groups constitute a population. This is only an imaginary population which exists hypothetically. Such a population is called a *hypothetical population*.

10.4 SAMPLE

Sometimes, the characteristics that we wish to probe are studied on the whole population. Census is taken to find out a few characteristics of all people. When a university wants to test the knowledge of its students, it examines all students belonging to its various affiliated colleges. In such cases, given resources (time, money and personnel), one can study the whole population. There are instances when one cannot study the whole population. If a doctor wants to know the number of red blood cells in a person he should be satisfied with the estimate based on a few drops (sample) of blood; he cannot think of extracting all the blood (population) from the body. Similarly, when the quality control officer wants to test the length of life of electric bulbs produced in a factory he can only take a sample of bulbs, burn them out, and then base his opinion about the whole lot on this experience. He cannot, for this purpose, burn out all the bulbs coming out of the factory. Thus, it is clear that certain situations would not permit us to study the whole population and we are forced to be content with testing a sample and making an inductive inference.

Even though in many investigations, study of the entire population is possible, sampling is resorted to because of limited resources. A good sample carefully selected and studied can give satisfactory information on the characteristics of the population. If the number of individuals to be studied is small, the investigator would

be able to give extra attention and record the facts more accurately. The heights and weights of a small group of people can be actually measured at the clinic and recorded, but if the number is large, the investigator has to record the heights and weights given by the patients without going into their accuracy. Thus, there are occasions when a few carefully recorded observations are far better and more useful than a mass of data collected haphazardly. Whenever sampling is resorted to it is important that the sample should reflect all the qualities found in the population. The problem is how to select a sample. This is more difficult than it appears to be.

10.5 SAMPLING VARIATION AND BIAS

The principal object of sampling is to get maximum information about the population with minimum effort or with limited resources. Any group of individuals taken from a population is a sample. We are not satisfied with any sample but only a representative (unbiased, unselected, random) sample. Sometimes, nature itself helps us in selecting such a sample. To study the Hb content of the blood of a person a little blood taken from the fingertip is enough. Blood in the body is well mixed and it is true that one drop of blood tells the same story as any other. There is no difficulty in selecting a sample (a few drops of blood) from the population (all blood in the body) because there is not much variation in the population. Sampling would be easy if all populations were like this. We seldom come across such populations. In almost all other populations, there is an endless variation among individuals as well as their environment. Consequently, one sample will differ from another taken from the same population. The natural biological variation between one individual and another will, in fact, be present in the sample also. The variation between one sample and another can be categorised as 'non-sampling' or 'sampling' variation (or bias) and may arise out of different types of sampling methods, inconsistencies in definition, tools, measurement process, nonresponse, carelessness, recording practices, etc. Such variations or biases should be removed through adequate planning, training, standardisation, supervision and monitoring. Sampling variation, on the other hand, is the inherent variation from one sample to another due to random or chance factors. This cannot be eliminated completely but can be minimised by the adoption of suitable methods of sampling.

Another factor that affects the representative nature of the sample is the bias in sampling. A human being is a very poor instrument for drawing an unbiased sample from a population. His consciousness and the psychological make-up will, in one way or another, force him to choose a biased sample. It is natural for an investigator who is collecting opinions from a particular location to select respondents from among those whom he already knows so that his job becomes easier.

This and other pragmatic or intuitively appealing nonprobability methods of sampling have been widely used and have many biases and limitations in generalisations.

10.6 NONPROBABILITY SAMPLING TECHNIQUES

10.6.1 Convenience Sampling

This is also known as *accidental*, *accessibility* or *haphazard sampling*. The major reason is administrative convenience, and the sample is chosen with ease of access being the sole concern. The pitfalls in terms of lack of 'representativeness' are obvious.

10.6.2 Purposive Sampling

This is also known as *judgmental sampling*. The attitude here is quite different. Recognising that the population may well contain different types of individuals, with differing measures and ease of access, the experimenter exercises deliberate subjective choice in drawing what he regards as a 'representative' sample. Judgmental sampling aims at the elimination of anticipated sources of distortion; but there will always remain the risk of distortion due to personal prejudices, or lack of knowledge of certain crucial features in the structure of the population.

10.6.3 Quota Sampling

Here, judgment and convenience are combined, and quota sampling is usually more structured than straight accessibility or judgmental sampling. A proper statistical design may have been used to determine what numbers are needed in each of the quotas, but some subjectivity and arbitrariness of choice of their constituent members cannot be avoided even with the most careful instruction of interviewers.

While nonprobability sampling may be convenient and economical, the major problem is not that it may lead to unrepresentative samples, but that its results are unconvincing (and too easily dismissed if unpalatable) because there is no yardstick against which to measure 'representativeness' or to assess the propriety or accuracy of estimators based on such a sampling principle. Such a yardstick is vital.

For this reason, we are compelled to introduce an element of 'randomness' into sampling procedures and to draw our samples according to some imposed probability mechanism. Such an approach is essential if we are to have a rationale for describing the respresentativeness and precision of our survey and its resulting estimators. A variety of probability sampling schemes have been devised, evaluated, compared, and utilised, and we shall study these in some detail as the only sound basis for survey sampling.

10.7 PROBABILITY (RANDOM) SAMPLING TECHNIQUES

All investigations are carried out to ascertain a particular character of the population. Often it is enough if an estimate of that character is found with the desired degree of accuracy through an unbiased sample drawn from that population. Five methods of drawing unbiased samples are described here:

1. Simple random sampling
2. Stratified random sampling

3. Systematic sampling
4. Cluster sampling
5. Multistage sampling

10.7.1 Simple Random Sampling

In this method, each member (sampling unit) of the population has an equal chance of being selected in the sample. The procedure is easy to apply if a list of all sampling units is available. The randomness of the sample is ensured by any one of several procedures such as use of lottery and a table of random numbers.

In the lottery method identification details of all the listed members of the population are written down on small slips of paper of uniform size, mixed well in a container or drum and then the required number of slips are picked up at random. Instead of slips identical cards may be used, shuffled well and the required number of cards drawn from them. These two procedures work well when the population is relatively small.

By far, the easiest method is that which uses a table of random numbers. A specimen list of random numbers is given in Appendix 2. These numbers, prepared by using certain randomising machines and then arranged in rows and columns, can be used to select either a single digit, double digit, three or four digit numbers. The randomness of the sample is ensured by the proper use of this table. This procedure is illustrated by an example. In Appendix 1, Hb values of 500 children are given. It is required to find out the arithmetic mean of these values. The arithmetic mean of these values can be found either by adding all the 500 values and dividing by 500 or by taking a sample of a suitable size, say 20, from the population of 500 Hb values and computing the sample arithmetic mean which will be an estimate of the population arithmetic mean. The latter method is easier. Let us assign serial numbers 1 to 500 to these 500 Hb values in an order; 11.6 is assigned number 1; 10.3, 2; 9.7, 3 and so on. Now, turn to the table of random numbers and arbitrarily select any number, or make your pencil point fall on the table and start with the number on which it fell. Let us assume that it fell on the number 2 which is in the junction of the 6th row from the top and 7th column from the left. We need to take three digit numbers in this situation and hence read all three digit numbers starting from this number. The numbers are 261, 266, 705, 690, 023, 599 etc. Drop all the numbers which are greater than 500 and take only 20 (sample size) numbers smaller than 500 in that order. Our sample consists of the 261st observation which is 10.2, the 266th observation which is 12.8 and so on. These give an arithmetic mean 10.75 which is an estimate of the population arithmetic mean.

10.7.2 Stratified Random Sampling

Very often the population that is investigated is not homogeneous. In order to make the sample more representative in that situation we would also like to consider different sections of the population which are homogeneous within themselves. In such cases, the stratified random sampling procedure is used. The population is divided into a number of sections called strata. A sample is drawn independently from each stratum by the simple random sampling method. Since a random sample is taken

from each stratum the whole population is adequately represented. The basic idea is that the heterogeneous population is divided into several sections each of which is more or less homogeneous; thus the variability in each stratum that exists in the population is adequately represented in the sample also. However, careful consideration is necessary to make appropriate stratification before selecting the sample.

10.7.3 Systematic Sampling

In this sampling method, every *n*th case is chosen for the study from a list of cases. Suppose, we wish to take a sample of 20 cases from the 500 Hb values given in Appendix 1. All the 500 values are arranged in some order and one observation is taken from each of the 25 observations systematically. The first value to be selected among the first 25 observations is to be determined by lot or from the table of random numbers. Suppose, it is 10. Then our sample consists of the 10th value, 35th, 60th, 85th, etc. and the last item is 485th value.

In this method, all sections of the population are adequately represented. Even when a list is not available this method is convenient and simpler and perhaps results in an unbiased sample in many instances. Consider this example where the list is not available: We want to pick out 150 houses from a total of 1500 houses in an area. We have to select one house in every ten. The initial number should be within the first 10 and is determined randomly. Let us assume that we have got the number 8. The investigator can start at one end of a street and proceed taking the 8th, 18th, 28th house etc. into the sample as he goes along. Nevertheless, there are a few problems involved, particularly when a locked or vacant house is chosen. Also at times, it might happen that every 10th number chosen is at the corner of a street and is a children's playground or a post office and not a house.

This method is often much more convenient than random or stratified random sampling in extensive sampling schemes, since the need for preparing a frame for selection is avoided. It must, however, be clearly recognised that the responsibility for the judgment that the material is such that systematic sampling will give satisfactory results rests with the investigator.

10.7.4 Cluster Sampling

This procedure is selected for use primarily for administrative convenience when the whole population consists of many natural groups (or clusters). In an immunisation status survey of first standard children in a town, a child may be regarded as a sampling unit and a random sample of the required number of children can be chosen using the nominal rolls kept in each school. Alternatively, the school itself can be considered as the sampling unit and a random sample of the schools selected. Thus, all the first standard children in these selected schools are included in the sample. This method is relatively simpler to administer than the random sample of first standard children in each school but is not so efficient. Another natural group or cluster is a family or a house or a hostel, etc. While most groups may be natural, one could also form artificial clusters. Thus, any method of sampling wherein a group is taken as a sampling unit is known as *cluster sampling*.

A well-known application of cluster sampling technique is the evaluation of immunisation coverage of children, wherein 30 clusters of at least 7 children in the age group of 12 to 23 months in each cluster are randomly selected to provide information on what proportion of those eligible are being properly immunised. Such a sample is expected to meet the following standards of reliability:

1. The data obtained from the survey will have a level of accuracy of plus or minus 10%. For example, if the survey shows immunisation coverage of 70% in the sample, the actual coverage in the target population will be between 60% and 80%.

2. The level of confidence is 95%, which means there is a 95% probability that the survey results will fall within the range listed above (plus or minus 10% of the coverage in the target population.

3. The survey data will reflect coverage of a period of 1 year prior to the date of survey.

4. The results will reflect coverage in the area as a whole. Comparisons between clusters are not valid.

The procedures start with a listing of all villages or sectors/wards of cities and towns included in the area for which service coverage is to be evaluated. Against the name of each village or ward, the individual population is noted, and the cumulative population of each village or ward determined. The sampling interval is determined using the formula:

$$\frac{\text{Total cumulative population}}{\text{30 clusters}}$$

A random number is then selected which is less than, or equal to, the sampling interval. This identifies the community in which cluster 1 is located. The community in which cluster 2 is located is obtained by adding the sampling interval to the Random Number. Subsequent clusters are similarly identified till 30 random clusters are found. In *cluster sampling*, probability is proportional to size (pps). More details are available in other books on Sampling.

10.7.5 Multistage Sampling

There are many ways in which the cluster sampling method may be modified or extended to cope with the specific demands of more complicated situations.

A situation arises where the selected clusters in the primary cluster sample are themselves sampled, rather than fully inspected. Thus, if we choose a stratified random sample of m of the M clusters (primary units) which comprise the population, we may then take stratified random samples of sizes $n_1, n_2, ..., n_m$ of the secondary units in the chosen clusters. Typically, we have $n_i < N_i$ ($i = 1, 2, ..., m$), in contrast to the sampling scheme of the previous section where $n_i = N_i$. The process of drawing samples from the selected clusters is called sub-sampling; the resulting total sample of size $n = \Sigma n_i$ is called a two-stage cluster sample.

If the secondary units are the individual members of the study population, there is no point in going further. If instead they consist of groups of population members,

then we might either use their constituent members, or proceed to further stages of subsampling. In this latter case, we encounter multistage cluster samples, corresponding to progressively higher levels of subsampling.

As described, the probability sampling mechanism at each stage is simple random sampling.

The use of suitable sampling techniques not only eliminates the bias but also enables us to find out the degree of precision of the estimate of the population character. The amount of sampling variation which influences the estimate can be measured for most of these methods based on techniques developed from appropriate mathematical models. In general, choosing the sample based on these techniques keeps the investigator on the safe side when he extends the findings from the sample to the population.

10.8 LISTING OF POPULATION

Most sampling techniques necessitate the use of a list of individuals in the population which is otherwise known as a '*frame*'. The population consists of many units—individuals or observations—called *sampling units*. In an investigation, the sampling unit may be a house, street, patient, family or guinea pig. The units must be defined clearly. If the sampling unit is an individual, the definition should give clearly what type of individual is included in the population. In the case of a house, the investigator should clearly know whether or not to include a school or a church in the population. The general principle guiding the definition of the sampling unit is that there should be no overlapping of the units and all units taken together constitute the population. Sometimes, a list may already be available (prepared by somebody else and for some other study) and such a list can be used with necessary modifications provided the definition of the sampling unit employed in the earlier study is the same as in the present one.

It is not always necessary to prepare a list. A good map of the place with serially numbered houses is enough to choose a sample of houses. Alphabetically or chronologically arranged out-patients cards are sufficient to get a sample of out-patients. These are used as substitutes for lists or frames.

10.9 SAMPLE SIZE

Often, the research worker is faced with the question of how many individuals or cases he should take in his study. What is the size of the sample which would be sufficient to answer the problem under investigation? Sometimes, results are kept in abeyance for want of enough cases. The number (sample size) depends on the type of problem investigated, required precision and to a certain extent, the resources available. Putting one man in orbit and bringing him back is sufficient proof that man can conquer space; but tens of thousands of cases are necessary to find evidence that smoking causes lung cancer. Specific problems on sample size are discussed in Chapter 23.

QUESTIONS

1. Write brief notes on:
   (a) Types of population
   (b) Sampling variation
2. Distinguish between population and sample.
3. Is there a need for studying a sample? What are the advantages of studying a sample over a population?
4. Explain the procedure of any three probability sampling techniques.

CHAPTER

11 Basis of Statistical Inference

11.1 INTRODUCTION

In Chapter 10, we discussed the importance of sampling in scientific investigations and described certain sampling techniques. A sample is expected to provide a description of the population from which it was drawn. Can we say that what we find in a sample of patients with a particular disease will be true of all other patients with that disease? Will another sample of patients present the same picture? If samples are subject to the vagaries of chance, and no two samples are alike, with how much confidence can we speak of the population from studying one sample? The procedures followed in drawing conclusions from sample values are known as *statistical inference.*

11.2 SAMPLING DISTRIBUTION

Very rarely do sample values coincide with the population value (parameter). The discrepancy between the sample estimate and the parameter is known as *sampling error*, when this discrepancy is the result of random sampling. Fortunately, these errors behave systematically and have a characteristic distribution. We shall illustrate this by taking a large number of samples from a known population.

Let the 500 haemoglobin values given in Appendix 1 constitute our population. The mean of the entire group is 10.96 g/100 ml. Take a random sample of size 10 and calculate its mean. Sample mean is an estimate of population mean. For example, let the sample consist of the following values: 9.5, 8.0, 10.7, 11.4, 11.8, 9.6, 12.0, 13.2, 7.8 and 15.3; the mean of these values is 10.93 g/100 ml. Take another sample of the same size (10) and calculate its mean; let us assume it is 11.4 g/100 ml. Let us repeat the same procedure and get 400 samples of the same size and calculate the means. We shall have four hundred estimates of the population mean. Table 11.1 gives the frequency distribution of 400 such estimates in such a sampling experiment. This distribution is called the *sampling distribution of means.*

It is found that most of the sample means cluster around the population mean (10.96 g/100 ml). Very few sample means deviate seriously from the population mean. The standard deviation of the sampling distribution is known as *standard error*, which in this case has the value 0.707 g/100 ml. If we have taken samples of larger size than 10 as in this case, we will find a smaller scatter (standard error) of the sample means.

Table 11.1 Frequency Distribution of Means of 400 Samples of Size 10 Taken from the 500 Hb Values

Mean Hb value (g/100 ml)	No. of samples
9.0–9.4	1
9.5–9.9	13
10.0–10.4	54
10.5–10.9	159
11.0–11.4	108
11.5–11.9	56
12.0–12.4	6
12.5–12.9	3
Total	400

Table 11.2 shows the proportion of sample means within the specified ranges. It can be seen that sample means below 10 g/100 ml and above 12 g/100 ml are rare.

Table 11.2 Distribution of Means of 400 Samples of Size 10 Taken from the 500 Hb Values

Mean Hb value (g/100 ml)	Proportion of samples
9.0–9.4	0.0025
9.5–9.9	0.0325
10.0–10.4	0.1350
10.5–10.9	0.3975
11.0–11.4	0.2700
11.5–11.9	0.1400
12.0–12.4	0.0150
12.5–12.9	0.0075
Total	1.0000

11.3 STATISTICAL INFERENCE

The statistical inference consists of two aspects: Estimation of a population value (or parameter), and testing of the hypothesis.

11.3.1 Estimation

Our main interest is to estimate the population parameter from a sample of observations. In fact, it should be the best unbiased estimate. The estimate may be a single statistic such as the mean which is called the *point estimate* or a range with attached probability, called *confidence interval*. For example, suppose the mean and standard error of a large sample of cases taken to estimate the population Hb value are 11.4 g/100 ml and 0.7 respectively. According to theoretical statistics, the sample mean is the best estimate of the population mean, so the point estimate of the

population mean is 11.4 g/100 ml. The interval with 1.96 times standard error below and above the mean, i.e., 11.4 ± 1.96 (0.7) is 10.0 g to 12.8 g. This interval is the 95% confidence interval. We say that the population mean lies between 10.0 g and 12.8 g with 95% confidence. In other words, if we were to find many such intervals each from a different sample, drawn in the same way, then in the long run, 95% of such intervals would include the population mean.

11.3.2 Testing a Hypothesis

This involves making an assumption about the parameter and checking the plausibility of that assumption using sample data.

We can also identify two types of research procedures with reference to hypothesis testing:

1. Comparison of two or more populations, e.g., patients with disease A and disease B.

2. Relationships between two or more factors in a population, e.g., smoking and lung cancer.

The role of sampling variation was considered in estimating the population value with a sample from the population. In hypothesis testing also, we have to deal with sampling variation as most often we have to make an inference on the basis of samples.

The process of statistical inference involves the following principles. We have to determine the sampling variation when the samples are drawn from statistically identical populations, or determine the correlation between factors which the samples are likely to reveal when, in the population, they are not related. Only if the observed sample difference or correlation is larger than that which the sampling variation would produce, can we conclude that differences (or correlation) exist in the populations. The assumption of no difference between populations (or no association between factors in the population) is known as *null hypothesis*.

The testing of the null hypothesis is fundamental in all tests of significance and consists of the following reasoning:

Compare the actual difference that we observe in our experiment with the corresponding sampling distribution, under the null hypothesis.

1. If the probability of observing that and greater differences in the sampling distribution is small, *reject the null hypothesis*. In this case, we say that the difference between the two treatments is *significant*. This means that the samples came from statistically different populations. Usually, *a probability of 5% or less is considered as rare or small* and sometimes 1% is used.

2. On the other hand, if the probability of observing that and greater differences as seen from the sampling distribution is large, we *do not reject* the null hypothesis, and say that the difference is *not significant*. This means that the difference that is observed in the actual experiment is not big enough to reject the null hypothesis.

11.4 TYPE I AND TYPE II ERRORS

While drawing conclusions as stated above, we are likely to commit two types of errors. These two types of errors are involved in the test procedure itself. We may conclude that the difference is significant when in fact there is no real difference in the population; and so reject the null hypothesis when it is true. This error is called *Type I error* whose magnitude is usually denoted by the greek letter α. Therefore, the risk of rejecting the null hypothesis when it is true is α. If in an experiment, Type I error is fixed at 0.05 (or 5%), limits are fixed in such a way that 5% of random samples will fall outside these limits if the null hypothesis is true. Therefore, we shall reject the null hypothesis when in fact it is true in 5% of the samples, thus making Type I error in 5% of the samples. On the other hand, we may give our final verdict that the difference is not significant when in fact there is real difference between the populations; that is, the null hypothesis is not rejected when actually it is false. This error is called the error of the second kind or *Type II error* whose magnitude is denoted by β. Type II error gives the magnitude of the risk of failure to detect the difference between the populations when there is a difference.

	Actual situation	
Decision	Null hypothesis is true	Null hypothesis is false
Null hypothesis rejected	Type I error	No error
Null hypothesis not rejected	No error	Type II error

Lets us suppose that treatment A produces 30% cure and treatment B produces 55% cure. Now, let us conduct a study, taking a pair of samples of size 30 each and administering treatment A to one group and treatment B to another group. Use the test of significance with the Type I error of 5% and judge whether there is any significant difference between the treatments based on these samples. Suppose that we have repeated such studies a large number of times and have judged that there is significant difference between the treatments only in 40 per cent of the studies. This means that in 60% of the pairs of samples we have failed to detect a difference that is large enough to reject the null hypothesis. We call this 60% as Type II error. The magnitude of risk of this error is related to the actual difference between the populations.

There is another useful concept closely associated with Type II error. If the Type II error is 60%, its complement, i.e., 40% is known as the 'power' of the test. The power is a numerical value indicating the sensitivity of a test.

Every investigator is anxious to keep both Type I and Type II errors at the lowest but it is not possible to reduce both the errors simultaneously. For a given sample size, if one is reduced, the other automatically increases.

Usually, the Type I error is fixed at a tolerable limit and the Type II error is minimised. After fixing the Type I error, the Type II error can be decreased by increasing the size of the sample.

QUESTIONS

1. Write an explanatory note on sampling distribution.
2. Explain two aspects of statistical inference: estimation and test of significance.
3. Briefly describe the steps followed in a test of significance.
4. Write brief notes on:
   (a) Standard error
   (b) Null hypothesis
   (c) Type I and Type II errors

CHAPTER

12 Tests of Significance and Estimation

12.1 INTRODUCTION

There are various types of problems for which the tests of significance are used for drawing conclusions. Different types of problems need different tests but the basis of all tests and the steps involved in the procedure are the same. The common types of problems are:

1. Comparison of sample mean with population mean
2. Comparison of two sample means
3. Comparison of sample proportion with the population proportion
4. Comparison of two sample proportions.

12.2 PROCEDURE

The steps involved in the procedure of tests of significance are as follows:

1. Finding out the type of problem and the question to be answered
2. Stating the Null Hypothesis
3. Determining the correct sampling distribution and calculating the standard error of the statistic used
4. Calculating the critical ratio. Generally, this is given by

$$\frac{\text{Difference between the statistics}}{\text{Standard error}}$$

This transforms the comparison to a standard form so that the available tables can be consulted.

5. Taking into account the distribution of the critical ratio and comparing the value observed in the experiment with that at the predetermined significant level given by the table; and
6. Making inferences.

One example from each of the 4 types mentioned above is treated here to illustrate the procedure.

12.3 FOR LARGE SAMPLES: EXAMPLES

12.3.1 Comparison of Sample Mean with Population Mean

Problem

In a Health Survey of schoolchildren, it is found that the mean haemoglobin level of 55 boys is 10.2 per 100 ml with a standard deviation of 2.1. Can we consider this group as taken from a population with a mean of 11.0 g/100 ml?

Solution

(a) *Question to be answered.* Is the mean Hb level of 10.2 g per 100 ml significantly different from the population mean of 11.0?

Given Data:

Size of sample	: $n = 55$
Mean of sample	: $\bar{x} = 10.2$ g/100 ml
Standard deviation of sample	: $S = 2.1$ g/100 ml
Mean of population	: $\mu = 11.0$ g/100 ml

(b) *Null hypothesis.* The sample is taken from the population with mean Hb level of 11.0 g/100 ml.

(c) *Standard error of mean.* In this problem, the means are compared. We, therefore, need the sampling distribution of the mean. It is known that if the population standard deviation is σ and the size of the sample is n, the standard error of mean is given by the formula $\sigma/\sqrt{n}$. Here the population standard deviation is not given, but the size of the sample is large, so we can assume that the sample standard deviation adequately represents the population standard deviation. Therefore, the standard error of mean is $S/\sqrt{n} = 2.1/\sqrt{55} = 0.283$.

(d) *Critical ratio.*

$$\text{Critical ratio} = \frac{\text{Difference in means}}{\text{Standard error of mean}}$$

$$= \frac{|\bar{x} - \mu|}{S/\sqrt{n}} = \frac{|10.2 - 11.0|}{0.283}$$

$$= \frac{0.8}{0.283} = 2.83$$

(e) *Comparison with the theoretical value.* It is known theoretically that the critical ratio follows a normal distribution. Turn to Appendix 4 where the probability of normal distribution is given (or see Ready Reckoner—Appendix 5). The 5% level given by the table is 1.96 and 1% level is 2.576. This means the probability of observing by chance a value greater than 2.576 is less than 1%. In this problem, our observed critical ratio is 2.83 which is greater than 2.576. This indicates that the probability of getting by chance a value as much as or greater than 2.83 is less than 0.01. Hence, we reject the null hypothesis ($P < 0.01$).

(f) *Inference.* The sample is not taken from the population with mean 11.0 g/100 ml.

12.3.2 Comparison of Means of Two Samples

Problem

In a survey on hearing levels of schoolchildren with normal hearing it was found that in the frequency 500 cycles per second, 62 children tested in the sound-proof room had a mean hearing threshold of 15.5 decibels with a standard deviation of 6.5. 76 comparable children who were tested in the fields had mean threshold of 20.0 decibels with a standard deviation of 7.1. Test, if there is any difference between the hearing levels recorded in the sound-proof room and in the field.

Solution

(a) *Question to be answered.* Is there any difference between the hearing levels recorded in the sound-proof room and in the field?

Given data:

	Sample 1	*Sample 2*
Sample size	$n_1 = 62$	$n_2 = 76$
Arithmetic mean	$\bar{x}_1 = 15.5$ decibels	$\bar{x}_2 = 20.0$ decibels
Standard deviation	$S_1 = 6.5$ decibels	$S_2 = 7.1$ decibels

(b) *Null hypothesis.* There is no difference between the means of the hearing thresholds taken in the sound proof room and in the field, that is, the two samples have come from the same population.

(c) *Standard error of the difference in means.* The null hypothesis states that the two samples have come from the same population. First the standard deviation of population, S, is to be calculated. This is given by the formula

$$S = \sqrt{\frac{n_1 S_1^{\,2} + n_2 S_2^{\,2}}{n_1 + n_2}}$$

$$S = \sqrt{\frac{62 \times 42.25 + 76 \times 50.41}{62 + 76}}$$

$$= 6.837 \text{ decibels}$$

The standard error of the difference in means is denoted by $S_{\bar{x}_1 - \bar{x}_2}$ and given by the formula

$$S_{\bar{x}_1 - \bar{x}_2} = S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= 6.837\sqrt{\frac{1}{62} + \frac{1}{76}} = 1.17$$

(d) *Critical ratio.*

$$\text{Critical ratio} = \frac{\text{Difference in means}}{\text{Standard error of difference in means}}$$

$$= \frac{|\bar{x}_1 - \bar{x}_2|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{|15.5 - 20.0|}{1.17} = \frac{4.5}{1.17} = 3.846$$

(e) *Comparison with the theoretical value.* In this case also the critical ratio follows a normal distribution. The 5% level is again 1.96 and the 1% level is 2.576. Our observed value is 3.846. The probability of observing this value (3.846) or a greater value by chance is less than 1%. Hence, the rejection of the null hypothesis ($P < 0.01$).

(f) *Inference.* There is evidence to believe that the hearing level tested in the sound-proof room is different from the hearing level tested in the fields.

12.3.3 Comparison of Sample Proportion with Population Proportion

Problem

In an otological examination of schoolchildren, out of 146 children examined 21 were found to have some type of otological abnormalities. Does it conform with the statement that 20% of the schoolchildren have otological abnormalities?

Solution

(a) *The question to be answered.* Is the sample taken from a population of children with 20% otological abnormalities?

Given data:

Size of sample : $n = 146$

Sample proportion : $p = \frac{21}{146} \times 100 = 14.4\%$

Population proportion : $P = 20\%$

(b) *Null hypothesis.* The sample has come from a population with 20% otologically abnormal children.

(c) *Standard error of proportion.* If the population proportion is P and $Q = 1 - P$, then the standard error of proportion is $\sqrt{PQ/n}$. In this problem, it is

$$\sqrt{\frac{20 \times 80}{146}} = 3.31$$

(d) *Critical ratio.*

$$\text{Critical ratio} = \frac{\text{Difference in proportions}}{\text{Standard error of proportion}}$$

$$= \frac{|14.4 - 20.0|}{3.31}$$

$$= 1.69$$

(e) *Comparison with theoretical value.* This critical ratio also follows a normal distribution whose 5% level is 1.96 and 1% level is 2.576. Our observed value is 1.69 which is less than the 5% level, that is, the probability of observing a value equal to or greater than 1.69 by chance is more than 5%. We, therefore, do not reject the null hypothesis.

(f) *Inference.* There is no evidence to show that the sample is not taken from a population of children with 20% abnormalities.

12.3.4 Comparison of Two Sample Proportions

Problem

In a hearing survey among 246 town schoolchildren 36 were found with conductive hearing loss and among 349 village schoolchildren 61 were found with conductive hearing loss. Does this present any evidence that conductive hearing loss is as common among town children as among village children?

Solution

(a) *Question to be answered.* Is there any difference in the proportion of hearing loss between children living in towns and villages?

Given data:

	Sample 1	*Sample 2*
Size	: $n_1 = 246$	$n_2 = 349$
Proportion with hearing loss	: $p_1 = (36/246) \times 100$	$p_2 = (61/349) \times 100$
	$= 14.6\%$	$= 17.5\%$

(b) *Null hypothesis.* The two samples have come from the same population; that is, there is no difference between the proportions of conductive hearing loss cases among the town children and among the village children.

(c) *Standard error of difference between proportions.* The proportion with conductive hearing loss in the population from which these two samples are drawn is given by

$$p = \left(\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \right)$$

assuming the null hypothesis is true. Here,

$$p = \left(\frac{36+61}{246+349}\right) \times 100 = 16.3\%$$

The standard error of the difference between the proportions is given by

$$S_{p_1-p_2} = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$$

where $q = 1 - p$, ($q = 100 - p$ if p is given in percentage).

$$S_{p_1-p_2} = \sqrt{\frac{(16.3 \times 83.7)}{246} + \frac{(16.3 \times 83.7)}{349}}$$

$$= 3.075$$

(d) *Critical ratio.*

$$\text{Critical ratio} = \frac{\text{Difference in proportions}}{\text{Standard error of the difference in proportions}}$$

$$= \frac{|p_1 - p_2|}{\sqrt{pq/n_1 + pq/n_2}}$$

$$= \frac{|14.6 - 17.5|}{3.075} = 0.94$$

(e) *Comparison with the theoretical value.* In this case also the critical ratio follows a *normal distribution*. The 5% level is 1.96 and the 1% level is 2.576. Our observed value is 0.94. The probability of observing by chance this value or a value greater than this is more than 5%. Therefore, we do not reject the null hypothesis.

(f) *Inference.* There is no evidence to show that the two samples are not drawn from the same population.

12.4 FOR SMALL SAMPLES: EXAMPLES

In all the four problems, the critical ratio followed the normal distribution. This is because the sample is large enough to find out the standard deviation of the population or the standard deviation of the population is known. If the size of the sample is small, the sample standard deviation does not adequately represent the population standard deviation even if the population from which the sample is drawn is a normal distribution. It is, therefore, necessary to get the estimate of the population standard deviation with the help of the sample standard deviation. Readily available mathematical formulae help us in getting the estimate of the population standard deviation from sample values. If the estimate of the standard error is substituted in the denominator of the critical ratio for the actual standard error, the critical ratio follows another sampling distribution known as *t*-distribution.

12.4.1 Comparison of Sample Mean with the Population Mean

Problem

A health status survey in a few villages revealed that the normal serum protein value of children in that locality is 7.0 g/100 ml. A group of 16 children who received high protein food for a period of six months had serum protein values shown in Table 12.1. Can we consider that the mean serum protein level of those who were fed on high protein diet is different from that of the general population?

Table 12.1 Serum Protein Level (g%) of 16 Children Fed on High Protein Food

Serial (child) No.	Protein level (g%)	Serial (child) No.	Protein level (g%)
1	7.10	9	7.36
2	7.70	10	6.59
3	8.20	11	6.85
4	7.56	12	7.90
5	7.05	13	7.27
6	7.08	14	6.56
7	7.21	15	7.93
8	7.25	16	8.56

Solution

(a) *Question to be answered.* Is the mean serum protein level of the children fed on high protein food significantly different from the population mean of 7.0 g%?

Given that:

Size of the sample : $n = 16$

Mean of the sample : $\bar{x} = 7.4$ g%

Mean of population : $\mu = 7.0$ g%

(b) *Null hypothesis.* The sample is taken from the population with mean serum protein level of 7.0 g%.

(c) *Standard error of the mean.* The estimate of the population standard deviation is given by s. In this problem, the estimate of the population standard deviation is 0.5589 g%. Dividing this value by $\sqrt{n}$ as was done in Section 12.3.1, the estimate of the standard error of the mean is obtained. This is found to be

$$\frac{0.5589}{\sqrt{16}} = 0.1397$$

(d) *Critical ratio.*

$$t = \frac{\text{Difference in means}}{\text{Estimate of the standard error of mean}}$$

$$= \frac{7.4 - 7.0}{0.1397} = 2.863$$

(e) *Comparison with the theoretical value.* The critical ratio, denoted by t, follows a t-distribution with $n - 1$ (i.e., $16 - 1 = 15$) degrees of freedom. In Appendix 6, values of t-distribution for certain probability level are given for various degrees of freedom. The t-distribution for 15 degrees of freedom gives the 5% level as 2.131 and the 1% level as 2.947. Our observed value is 2.863, which is more than the 5% level but less than the 1% level. This means the probability of getting by chance a value as much as 2.863 or greater is less than 5%. Hence, we reject the null hypothesis ($P < 0.05$).

(f) *Inference.* There is evidence to believe that the mean serum protein level of the children fed on high protein food is different from the mean serum protein level of the general population.

12.4.2 Comparison of Means of Two Independent Samples

Problem

In a feeding trial, 17 children were given high protein food supplement to their normal diet and 15 comparable children were kept under normal diet. The total calories of intake per child per day of the high protein group is 1296 and of the control group is 1293. They were kept on this feeding trial for a period of seven months. At the end of the study, the changes (initial – final) in the haemoglobin (g%) level of the two groups are assessed and given in Table 12.2. Does it provide any evidence to say that the change in the haemoglobin level of the children who received high protein food is different from the control group?

Table 12.2 Changes in Haemoglobin Levels of Children in High Protein Diet and Control Group

High protein diet group		Control Group	
Serial (child) No.	Change in Hb.	Serial (child) No.	Change in Hb.
1	– 0.50	1	0.77
2	– 2.20	2	1.60
3	– 1.90	3	0.81
4	– 0.67	4	2.00
5	– 0.73	5	– 0.94
6	0.23	6	1.77
7	1.08	7	0.45
8	1.22	8	0.54
9	– 0.41	9	0.45
10	– 0.16	10	1.22
11	– 4.59	11	– 2.30
12	– 0.35	12	0.20
13	– 0.98	13	4.90
14	– 1.00	14	1.70
15	– 1.01	15	1.06
16	2.27		
17	– 2.73		

Solution

(a) *Question to be answered.* Is there any difference between the means of the change in the haemoglobin of the children fed on high protein food and those who were fed on normal food?

Given data:

	High protein diet group	*Control group*
Sample Size	$n_1 = 17$	$n_2 = 15$
Mean (calculated)	$\bar{x}_1 = -0.69\%$	$\bar{x}_2 = 0.9487$ g%

(b) *Null hypothesis.* The two samples have come from the population with same mean. In other words, there is no difference in means of the change in haemoglobin values between the children fed on the high protein diet and normal diet.

(c) *Standard error of the difference in means.* This estimate of the standard error of difference in means is given by the formula

$$SE = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$s = \sqrt{\frac{(n_1 - 1)\, s_1^{\,2} + (n_2 - 1)\, s_2^{\,2}}{n_1 + n_2 - 2}}$$

s_1 = standard deviation of the first sample

s_2 = standard deviation of the second sample

In this problem,

$$s = \sqrt{\frac{41.2704 + 33.8646}{17 + 15 - 2}} = 1.5826\,\text{g}\%$$

(d) *Critical ratio.* The critical ratio is given by

$$t = \frac{|x_1 - x_2|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{|-0.69 - 0.9487|}{1.5826\sqrt{1/17 + 1/15}}$$

$$= 2.923$$

(e) *Comparison with the theoretical value.* This critical ratio denoted as t follows a t-distribution with $n_1 + n_2 - 2 = (17 + 15 - 2 = 30)$ degrees of freedom. The t-distribution for 30 degrees of freedom gives the 5% level as 2.042 and 1% level as 2.750. Our observed value is 2.923 which is greater than the 1% level. This

means the probability of getting by chance a value as much as 2.923 or greater is less than 1%. Therefore, we reject the null hypothesis ($P < 0.01$).

(f) *Inference.* This experiment provides evidence to show that the mean change in the haemoglobin level of the children fed on high protein diet is different from the mean change in the haemoglobin level of the children fed on normal diet.

This *t*-test assumed that the variances of the two groups are the same. If there is evidence against this assumption, a modified *t*-test has to be applied. This is described by Snedecor and Cochran (1967).

12.4.3 Comparison of Means of Two Correlated Samples (i.e., with the Same Subjects in Both Groups): The Paired *t*-test

Problem

Twelve pre-school children were given a supplement of multi-purpose food for a period of four months. Their skin-fold thickness (in mm) was measured before the commencement of the programme and also at the end. The values obtained are given in Table 12.3. Test if there is any change in their skin-fold thickness.

Table 12.3 Skin-Fold Thickness of the Children Who Received Multi-Purpose Food

Serial (child) No.	Skin-fold thickness (mm) At the beginning	At the end	Difference 'd'	$d - \bar{d}$ *	$(d - \bar{d})^2$
1	6	8	2	1.25	1.5625
2	8	8	0	− 0.75	0.5625
3	8	10	2	1.25	1.5625
4	6	7	1	0.25	0.0625
5	5	6	1	0.25	0.0625
6	9	10	1	0.25	0.0625
7	6	9	3	2.25	5.0625
8	7	8	1	0.25	0.0625
9	6	5	− 1	− 1.75	3.0625
10	6	7	1	0.25	0.0625
11	4	4	0	− 0.75	0.5625
12	8	6	− 2	− 2.75	7.5625
Total	79	88	9	0	20.2500

*$\bar{d}$ = 0.75

Solution

(a) *Question to be answered.* Is there a difference between the measurements of skin fold thickness taken at the beginning and at the end of the experiment?

(b) *Null hypothesis.* The sample is taken from the population in which there is no difference in the skin-fold thickness.

(c) *Standard error of the mean of difference.* The mean of the differences is 0.75 mm. The estimate of the population standard deviation is given by

$$s = \sqrt{\frac{\Sigma (d - \bar{d})^2}{n - 1}}$$

where n stands for the number of subjects included in the study. In the problem,

$$s = \sqrt{\frac{20.25}{11}} = 1.357 \text{ mm}$$

The estimate for the standard error of mean is

$$\frac{s}{\sqrt{n}} = \frac{1.357}{\sqrt{12}} = 0.3917 \text{ mm}$$

(d) *Critical ratio.*

$$t = \frac{\bar{d} - 0}{s/\sqrt{n}} = 1.91$$

(e) *Comparison with the theoretical value.* This critical ratio, t, follows a t-distribution with $n - 1$ (12 – 1 = 11) degrees of freedom. The 5% level is 2.201 and the 1% level is 3.106 for 11 degrees of freedom. The value 1.91 is less than the 5% level. This means the probability of getting by chance a value as much as 1.91 or greater is more than 5%. Therefore, we do not reject the null hypothesis.

(f) *Inference.* This experiment does not provide any evidence to say that there is a difference between the initial and final values of skin-fold thickness.

12.5 ESTIMATION: EXAMPLES FOR LARGE SAMPLES

Probem 1

In a health survey of 55 schoolboys, it was found that the mean haemoglobin level was 10.2 g per 100 ml with a standard deviation of 2.1. Estimate the mean haemoglobin level of the population of such schoolboys.

Solution

It is theoretically proved that the mean of the sample is the best estimate of the mean of the population. Therefore, the estimate of the population mean = 10.2 g/100 ml. This is a point estimate because it specifies one value (one point) as the estimated haemoglobin level. We can also get the confidence interval as an interval estimate. The confidence interval of mean is given by

$$\bar{x} - Z_p \text{ (SE of mean) to } \bar{x} + Z_p \text{ (SE of mean)}$$

or

$$\bar{x} \pm Z_p \text{ (SE of mean)}$$

where Z_p is the value in the normal distribution corresponding to the confidence probability p.

To get 95% confidence interval, look at the last two columns of Appendix 4, and note that 95% of the observations of the normal distribution lie between $Z = -1.96$ and $Z = 1.96$. So, $Z_p = 1.96$ (also from Appendix 5). Therefore, (using data from Section 12.3.1) the 95% confidence interval is

$$10.2 - 1.96\ (0.283) \text{ to } (10.2) + 1.96\ (0.283)$$

or

$$9.65 \text{ to } 10.75$$

This indicates that the population mean will lie in the interval 9.65 to 10.75 with 95% probability.

To determine 99% confidence interval, again, find from Appendix 4 that 99% of the observations of normal distribution lie between $Z = -2.576$ and $Z = 2.576$. So, $Z_p = 2.576$. Therefore, the 99% confidence interval is

$$10.2 - 2.576\ (0.283) \text{ to } 10.2 + 2.576\ (0.283)$$

or

$$9.47 \text{ to } 10.93$$

We say that population mean will lie in the interval 9.47 to 10.93 with 99% confidence.

Problem 2

Using the data from the problem of 12.3.2, find the 95% confidence interval for the difference in means.

Solution

Following the procedure used in the previous problem, for 95% confidence interval Z_p is obtained as 1.96. The confidence interval for difference in mean is

Difference in means ± 1.96 (SE difference in means)
using data from Section 12.3.2, the 95% confidence interval is

$$4.5 - 1.96\ (1.17) \text{ to } 4.5 + 1.96\ (1.17) \text{ or } 2.21 \text{ to } 6.79$$

The 95% confidence interval for difference in means is from 2.21 to 6.79.

Problem 3

Find the 99% confidence interval for the proportion of children with otological abnormalities, using data from Section 12.3.3.

Solution

The Z_p for 99% confidence interval is given by 2.576 (see Problem 1). The 99% confidence interval of the proportion is

$$p \pm 2.576 \text{ (SE of proportion)}$$

where p = proportion with otological abnormalities, and

$$\text{SE of proportion} = \sqrt{\frac{14.4 \times 85.3}{146}} = 2.90$$

i.e.,

$$14.4 \pm 2.576\ (2.9) \text{ or } 6.93 \text{ to } 21.87$$

The 99% confidence interval of the population proportion is 6.93% to 21.87%.

12.6 ESTIMATION: EXAMPLES FOR SMALL SAMPLES

Problem 1

Find the best estimate of the mean and the 95% confidence interval of the mean using the data given in Section 12.4.1.

Solution

The best estimate of the population mean is the sample mean, which is 7.4 g%.

The 95% confidence interval is given by $\bar{x} \pm t_{0.05}$ SE of $\bar{x}$, where $t_{0.05}$ is obtained from the t-distribution given in Appendix 6 for n_1 (= 15) degrees of freedom. Here, $t_{0.05} = 2.131$ and the confidence interval is, thus,

$$7.4 \pm 2.131\ (0.1397) \text{ or } 7.1 \text{ to } 7.7$$

Problem 2

Find the 95% confidence interval of the difference in means for the data of Section 12.4.2.

Solution

The 95% confidence interval of difference in means is given by

$$\text{Difference in means} \pm t_{0.05}\ (\text{SE of difference in means})$$

the $t_{0.05}$ with $n_1 + n_2 - 2$ (= 30) degrees of freedom is 2.042, which is obtained from the t distribution given in Appendix 6. Since the difference in means is

$$\bar{x}_1 - \bar{x}_2 = 1.6387$$

the 95% confidence interval is

$$1.6387 \pm 2.042\ (0.5607) \text{ or } 0.49\text{–}2.78$$

12.7 READING COMPUTER OUTPUT

Many investigators use statistical package programs to do tests of significance. The computer output gives generally, the values of critical ratio such as Z or t or χ^2 or F and the probability (with many decimal places) of obtaining that critical ratio. The investigator should ascertain the significance by reading the probability displayed or printed in the output. If the probability (or P) given by the computer is less than 0.05, then the difference (or association) is declared significant and the null hypothesis is rejected ($P < 0.05$). If the probability is less than 0.01, then the difference or the association is declared significant and the null hypothesis is rejected ($P < 0.01$). If

the probability is more than 0.05, the difference (or association) is declared not significant and the null hypothesis is not rejected. Sometimes, the computer output gives $P = 0.000000$ (the number of zeros displayed depends on the command in the built-in program) which should be considered as very small and the null hypothesis should be rejected.

While presenting the P values, some journals accept presenting it as $P < 0.05$ or $P < 0.01$, etc. but other journals prefer to specify the exact value of the probability up to three or four decimal places. For example, in Section 12.4.1 the value of t was calculated as 2.863. The computer gives the exact probability for this t with 15 degrees of freedom as $P = 0.01185058$. This could be presented as $t = 2.863$ and $P = 0.0119$ (correct to four decimal places) in the journals.

QUESTIONS

1. What are the common types of problems for which tests of significance are used? Give an example for each.
2. Explain the methodology followed in the tests of significance.
3. Explain the procedure of estimating population mean from:
   (a) small sample
   (b) large sample
4. Distinguish between point estimate and interval estimate.
5. What is confidence interval?
6. Give the meaning of P value used in test of significance.
7. The birth weights (kg) of two random sample of infants belonging to rural and urban areas are given below:

   Rural: 3.1, 2.9, 2.8, 3.0, 2.7, 3.1, 2.6, 2.8, 2.9, 3.0

   Urban: 3.5, 3.0, 3.1, 3.2, 2.9, 3.4, 3.0, 3.4, 2.8, 3.4

   Test whether the difference between the two means is statistically significant.

CHAPTER

13 Linear Regression and Correlation

13.1 INTRODUCTION

In the foregoing chapters, we have considered observations made on only one characteristic in each of a group of individuals. However, one of the chief objectives of science is to estimate values of one factor by reference to the values of an associated factor; and very often we find that research workers are interested in more than one characteristic of a group of individuals and the relationships existing among them. In this chapter, we shall briefly discuss a method of investigating the relationship between two characteristics, both of which are quantitative in nature. In such investigations, we shall have two observations—one for each characteristic made for each individual. For example, age and blood pressure of a group of individuals may have been recorded in the course of an investigation. Are these two characteristics related? If so, what is the nature of this relationship? Does the knowledge of this relationship help us predict more accurately the value of one characteristic of an individual, when the value of the other characteristic of that individual is known?

13.2 SCATTER DIAGRAM

An analysis of the relationship may be started by representing the values of the two characteristics of each individual by a point on a graph known as *scatter diagram*. Such a graph is illustrated in Figure 13.1 for the data shown in Table 13.1.

Table 13.1 Gestational Age, GA (Weeks) and Crown-Heel Length, CHL (cm) of 20 Newborns

Serial No.	GA (weeks)	CHL (cm)	Serial No.	GA (weeks)	CHL (cm)
1	34	46.8	11	31	46.4
2	30	47.0	12	38	47.2
3	32	47.0	13	34	47.4
4	28	46.2	14	40	47.8
5	35	47.0	15	42	48.4
6	37	47.0	16	37	46.6
7	40	47.4	17	35	47.8
8	29	46.6	18	40	48.2
9	38	47.8	19	42	48.0
10	32	46.6	20	41	48.0

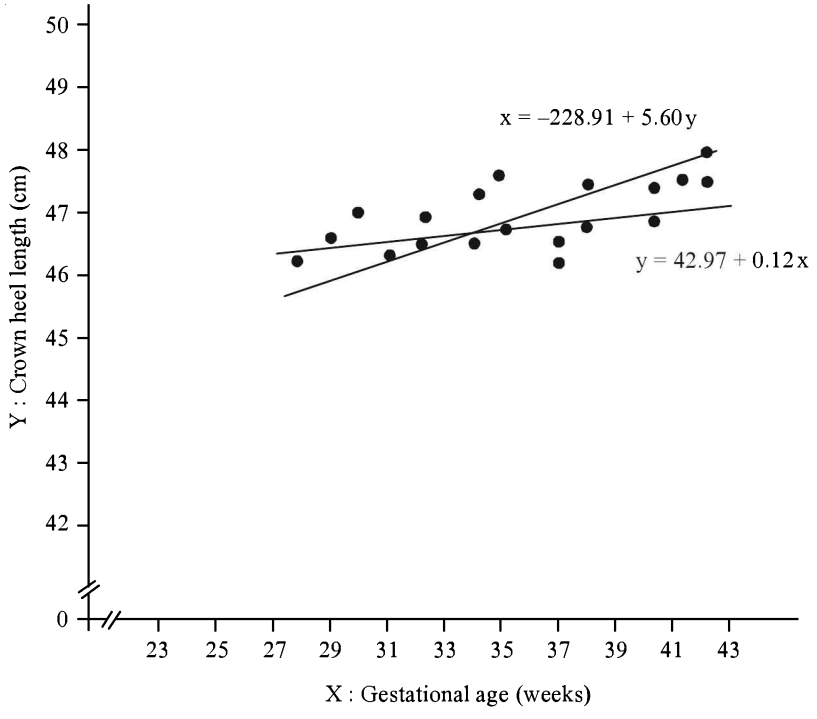


Figure 13.1 Relationship between Gestational Age and Crown-Heel Length.

The configuration of the points on the graph will indicate the nature of relationship. If the points lie scattered randomly over the ranges of observations of the two characteristics, absence of any relationship between these characteristics, is suggested. If, on the other hand, these points lie clustered along a curve or a straight line, a relationship is suggested and can be described by means of an appropriate curve or straight line around which the points cluster. We come across several types of such curves of relationship. Many growth curves are either of the logistic or exponential type. The seasonal curves are usually cyclic. A straight line is also a usual type. It must be remembered that when the range of observation is limited to a small section in the above cases, it may be difficult in most cases to differentiate the pattern observed from a straight line.

13.3 CORRELATION AND REGRESSION

There has been much confusion on the subject of correlation and regression. Quite frequently, correlation problems are treated as regression in scientific literature, and the converse is equally true. There are several reasons for this confusion. First of all, the mathematical relations between the two methods of analysis are quite close and mathematically one can easily move from one to the other. Hence, the temptation to do so is great. Secondly, earlier texts had not made the distinction between the two approaches sufficiently clear, and this liability has still not been entirely overcome. Finally, while the approach chosen by an investigator may be correct in terms of his

intentions, the data available for analysis may be such as to make one or the other of the techniques inappropriate.

Let us look at the intentions or purposes behind the two types of analysis. In correlation, we are concerned whether two variables are interdependent or co-vary, that is, whether they vary together. We do not express one as a function of the other. There is no distinction between independent and dependent variables. It may well be that of the pair of variables whose correlation is studied, one is the cause of the other but we neither know nor assume this. A more typical (but not essential) assumption is that the two variables are both effects of a common cause. What we wish to estimate is the degree to which these variables vary together. Thus, we might be interested in the correlation between biparietal diameter and femur length in a population of human foetuses or between birth weight and maternal height or between drug intake and number of days taken for cure.

In regression, we intend to describe the dependence of a variable on an independent variable. We employ regression equations to lend support to hypotheses regarding the possible causation of changes in y by changes in x; for purposes of prediction, of y in terms of x; and for purposes of explaining some of the variations of y by x, by using the latter variable as a statistical control. Studies of the effects of temperature on heartbeat rate, protein intake on growth rate in a child, age of person on blood pressure, or dose of an insecticide on mortality of the insect population are all typical examples of regression for the puposes named above.

13.4 CORRELATION COEFFICIENT

There are many situations wherein it is impossible to decide which one of the two characteristics is dependent on the other. In such cases, the relationship is rather mutual. A measure of the strength of the relationship between two variables is provided by the coefficient of correlation, denoted by r, if the relationship between the two variables is of the linear form. It is also called the *coefficient of linear correlation* or *Pearson's product-moment correlation*. When *perfect relationship/ correlation* between the two characteristics is indicated (by a one-to-one relation between the values of these characteristics) the value of one characteristic of an individual is sufficient to determine exactly the value of the other characteristic of that individual. The correlation coefficient calculated for data arising from such situations will have a value + 1 or – 1. When higher values of one characteristic are associated with higher values of the other, this is termed *positive correlation* and has a value of + 1 if the relationship is perfect. Alternatively, when higher values of one characteristic are associated with lower values of the other, the correlation is termed *negative* and has a value of – 1 in case the relationship is perfect.

In situations where one characteristic is not correlated with the other, the value assessed by one characteristic of the individuals will not have any relevance to the values of the other characteristic in these individuals. The correlation coefficient will tend to have a value, either positive or negative, near zero.

In the case of many biological characteristics, we do not come across perfect correlation which will result in a coefficient of correlation of + 1 or – 1. Often, there is sufficient basis to look for some association between the two and the correlation

coefficient gives a value which is an estimate of the extent of this association. Such a situation may be identified for some data only after a test of significance.

The square of the correlation coefficient, r^2, is called the *coefficient of determination* or *explained variation*, i.e., the proportion of the total variation in the dependent variable that is explained by the regression line. For example, blood pressures show a fair degree of variation from one person to another due to many factors, one of which is body weight. This can be referred to as the total variation in the variable. Some of the total variation in blood pressure measurements can be explained by the variation in body weights (the 'explained' variation), while the rest of the variation must be due to other factors which were not considered explicitly or are unknown (the 'unexplained' variation). The larger the first component relative to the second, the stronger is the relationship between the two variables.

Coefficient of correlation, r, is calculated by the formula

$$\frac{(1/n)\,\Sigma\,(x-\bar{x})\,(y-\bar{y})}{(\text{SD of } x)\,(\text{SD of } y)} \quad \text{or} \quad \frac{(1/n)\,\Sigma\,(xy)-\bar{x}\,\bar{y}}{(\text{SD of } x)\,(\text{SD of } y)}$$

where x denotes the value taken by one characteristic, y the value taken by the other characteristic and $\bar{x}$ and $\bar{y}$ the arithmetic means of the values of x and y respectively; the term

$$((1/n)\,\Sigma\,xy)-\bar{x}\,\bar{y} \text{ is known as } \textit{covariance.}$$

The method of calculation of the correlation coefficient is illustrated in Table 13.2.

Table 13.2 Calculation of Correlation Coefficient between Gestational Age (Weeks) and Crown-Heel Length (cm) of 20 Newborns (Data of Table 13.1)

	GA (x)	CHL (y)	x^2	y^2	xy
	34	46.8	1156	2190.24	1591.2
	30	47.0	900	2209.00	1410.0
	32	47.0	1024	2209.00	1504.0
	28	46.2	784	2134.44	1293.6
	35	47.0	1225	2209.00	1645.0
	37	47.0	1369	2209.00	1739.0
	40	47.4	1600	2246.76	1896.0
	29	46.6	841	2171.56	1351.4
	38	47.8	1444	2284.84	1816.4
	32	46.6	1024	2171.56	1491.2
	31	46.4	961	2152.96	1438.4
	38	47.2	1444	2227.84	1793.6
	34	47.4	1156	2246.76	1611.6
	40	47.8	1600	2284.84	1912.0
	42	48.4	1764	2342.56	2032.8
	37	46.6	1369	2171.56	1724.2
	35	47.8	1225	2284.84	1673.0
	40	48.2	1600	2323.24	1928.0
	42	48.0	1764	2304.00	2016.0
	41	48.0	1681	2304.00	1968.0
Total	715	945.2	25931	44678.00	33835.4
Mean	35.75	47.26	1296.55	2233.9	1691.77

Standard deviation of $x = \sqrt{1296.55 - (35.75)^2} = 4.30$

Standard deviation of $y = \sqrt{2233.9 - (47.26)^2} = 0.63$

Correlation coefficient, $r = \dfrac{1691.77 - (35.75)(47.26)}{4.30 \times 0.63} = 0.82$

The formula for *standard error* of the correlation coefficient is $\sqrt{\dfrac{1-r^2}{n-2}}$ which is 0.14 for this correlation coefficient.

The statistical significance of the correlation coefficient is found by referring to Appendix 9 with $n - 2$ degrees of freedom. The table shows that for 18 degrees of freedom, a correlation of 0.561 or more is significant at $p < 0.01$ level.

Alternatively, one could use the t-distribution and determine $t = 0.82/0.14 = 5.86$. A reference to Appendix 6 gives the same conclusion as that for $p < 0.01$.

When a large series of observations is collected, the data may be expressed as a frequency table as shown in Table 13.3.

Table 13.3 Overcrowding and Infant Mortality, Example of Correlation Table

Infant mortality rate per 1000 (x)	Number of towns and percentage of population in private houses where more than two persons stay in a room (y)						
	1.5–	4.5–	7.5–	10.5–	13.5–	16.5–19.5	Total
36–46	5						5
46–56	9	1					10
56–66	10	4	1			1	16
66–76	4	7	5	2			18
76–86	2	5	4	1	1		13
86–96		2	2	2		1	7
96–106		1	2	2	1	1	7
106–116		1		1			2
Total	30	21	14	8	2	3	78

The calculation of the correlation coefficient is quite similar and the formula can be expressed as

$$r = \frac{(1/n)\,\Sigma f \cdot (x - \bar{x})(y - \bar{y})}{(\text{SD of } x)(\text{SD of } y)} \quad \text{or} \quad \frac{[(1/n)\,\Sigma f \cdot xy] - (\bar{x}\,\bar{y})}{(\text{SD of } x)(\text{SD of } y)}$$

The computations are similar to those used when dealing with frequency distributions. The mid-points of each class-interval are first obtained for the first variable x (infant mortality) and used in determining the means and standard deviations. Likewise, for the second variable y (overcrowding). For the covariance, each frequency is multiplied by the mid-points of the corresponding row and column class-intervals, added and divided by n. Thus

For infant mortality (x),

$$\bar{x} = 71.64, \qquad s_x = 17.43$$

For overcrowding index (y),

$$\bar{y} = 6.69, \qquad s_y = 3.99$$

$$\text{Covariance} = \frac{40722}{78}$$

Therefore,

$$r = 0.62$$

13.5 REGRESSION EQUATION

Let us consider here the case of linear relationship only which is characterised by a tendency of the points on the Scatter Diagram to cluster along a straight line, known as the *regression line*. This line,

$$y = a + bx$$

is used to describe the dependence of one characteristic (y) upon the other characteristic (x), both x and y representing the value of the two characteristics, a and b are constants computed from the data. It will be evident that two regression lines can be computed for every set of data—one each to describe the dependence of one characteristic on the other.

b is known as *regression coefficient* which shows a change expected in y for unit change in x. The above line describes the dependence of y on x; and b is the regression coefficient of y on x.

For the data of Table 13.1, the regression equation of y on x gives the estimated value of PVC for a specified value of haemoglobin. The regression coefficient b is estimated using the method of least squares (this method is not explained here).

$$b = r\ (\text{SD of } y)/(\text{SD of } x)$$

or

$$b = \frac{(1/n)\,\Sigma\,(x - \bar{x})\,(y - \bar{y})}{(\text{SD of } x)^2}$$

An easy way of writing the equation is

$$\frac{y - \bar{y}}{\text{SD}_y} = r\,\frac{x - \bar{x}}{\text{SD}_x}$$

$$\frac{y - 47.26}{0.63} = 0.82\,\frac{x - 35.75}{4.30}$$

$$y - 47.26 = 0.120\ (x - 35.75)$$

$$y = 0.120x + 47.26 - 4.29$$

or

$$y\ (\text{CHL}) = 42.97 + 0.120x\ (\text{GA})$$

For example, the estimated CHL for a baby with GA 40 is found by substituting 40 for x in this equation. The estimated CHL is 47.77.

Similarly, the regression equation of x on y gives the estimated value of GA for a specified value of CHL. Its equation is

$$\frac{x-\bar{x}}{\mathrm{SD}_x} = r\frac{y-\bar{y}}{\mathrm{SD}_y}$$

$$\frac{x-35.75}{4.30} = 0.82\frac{y-47.26}{0.63}$$

$$x - 35.75 = 5.60(y - 47.26)$$

$$x = 5.60y + 35.75 - 264.66$$

$$x\ (\mathrm{GA}) = -228.91 + 5.60y\ (\mathrm{CHL})$$

The *standard error* of regression coefficient b is given by

$$S_b = \sqrt{\frac{1}{n-2}\left[\frac{\Sigma(y-\bar{y})^2}{\Sigma(x-\bar{x})^2} - b^2\right]}$$

In the regression equation $y = 0.12x + 42.97$ (for data of Table 13.2) the regression coefficient is 0.12. Its standard error is

$$S_b = \sqrt{\frac{1}{20-2}\left[\frac{0.3924}{18.4875} - 0.014\right]}$$

$$= 0.02$$

13.6 RESTRICTIONS

There are certain precautions to be taken while using the regression line and correlation coefficient:

1. The regression lines should not be extended beyond the range of the data on which they are based unless there is sufficient justification for doing so on other grounds. For instance, the regression line of height on age of school-age children cannot be extended further to predict the heights of adults.

2. While predicting the value of the dependent characteristic corresponding to a value of the independent characteristic lying between two observed values, there is an assumption that there is uniform trend in the dependent characteristic between these values.

3. In interpreting a correlation coefficient, we have to remember that a correlation between two characteristics need not suggest that they are related directly, i.e., a value observed for one of these characteristics is due to the value observed for the other characteristic, of an individual. In the case of the correlation observed between the heights of brothers and sisters it is obvious that a third set of common factors operate to bring about this association. A tall brother is not the cause of a tall sister.

4. The correlation coefficient gives the intensity of relationship only when the nature of relationship is linear but not for other types of relationships.

5. If x and y are independent, the coefficient of correlation r is zero, but if $r = 0$, x and y are not necessarily independent. They are merely uncorrelated (or have no linear relationship).

6. The data should be reasonably homogeneous. If the dots in the scatter diagram show a tendency to cluster in two or more groups, a spuriously high value of r may result, due merely to the heterogeneity of the data.

7. Even if x and y are independent of each other, it may be that both vary roughly in the same way (or in opposite ways) with time, and hence, if observations are spread over a long interval of time, a *spurious* correlation may be found. Yule observed a very high correlation ($r = 0.951$) between the proportion of marriages celebrated in Anglican Churches (in England and Wales) from 1866 to 1911 and the standardised mortality rate over the same period. The reason is that both these variables show a well-marked downward trend over the period. Yule called correlations such as this *nonsense correlations*. Another example of this was pointed out by the Norwegian statistician L.V. Charlier, who found a correlation of 0.86 between the size of the stork population in Oslo over a period of about 40 years and the number of babies born there each year.

8. When there is a real correlation between two variables, it may be that a change in the one variable is the cause of a change in the other. If so, the former would be taken as x and the latter as y, and we should be interested in the regression of y on x. Sometimes, however, it is not clear which variable corresponds to cause and which to effect.

13.7 NONLINEAR REGRESSION

There are several techniques available now to handle situations when the scattergram suggests a nonlinear set of data points or some kind of curve. Transformation of the data is a common solution to the problem. In some cases, taking the logarithm of the variables may lead to a linear relationship. For example, the curvilinear relationship $Y = bX^2$ can be linearised by transforming it to $\log Y = \log b + 2 \log X$. Many exponential relationships can be made linear by such a method. Another approach to curvilinear regression is to run a polynomial regression where, without transformation, a dependent variable Y might be related to an independent variable X with a polynomial of the following form

$$Y = a + bX + cX^2$$

with terms in X^3 and X^4 ... . If necessary, in such analyses, more than one regression coefficient must be estimated.

13.8 MULTIPLE REGRESSION

The concepts and techniques for analysing the association among three or more variables are natural extensions of those explored in the bivariate situation discussed

so far. In the multiple regression model, we assume that a linear relationship exists between some variable Y, which we call the dependent variable, and k independent variables, X_1, X_2, ..., X_k. The independent variables are sometimes referred to as *explanatory* variables, because of their use in explaining the variation in Y; or as *predictor* variables, because of their use in predicting Y. In general, we ought to be able to improve our predicting ability by including more independent variables in such an equation.

The strength of the relationship between the dependent variable and the explanatory variables may also be estimated by calculating the *coefficient of multiple correlation* (R). This is analogous to the simple correlation coefficient r and as before, its square measures the proportion of the total variation in the dependent variable which can be 'explained' by variations in the explanatory variables. The 'unexplained' variation may be due, of course, to other variables which have not been included in the regression equation. If these variables can be identified, then a new multiple regression equation can be calculated with these additional variables included. By calculating what are called *partial correlation coefficients*, the strength of the relationship between the dependent variable and any one of the independent variables may be calculated, assuming the other independent variables are held constant. These partial correlation coefficients differ from the simple bivariate (or zero order) correlation coefficients described earlier. In the former, the simultaneous influence of other independent variables is taken into account; in the latter, the correlation between two variables is calculated without any explicit attempt to remove the possible influence of other variables.

Various methods are employed to choose the 'best' (in some sense) set of independent variables for inclusion in a multiple regression equation. *Forward* (*stepwise*) inclusion methods search through all the possible independent variables not already in the equation and add into the equation, at each step, the variables that increase the coefficient of multiple correlation to the greatest extent. (Other statistical criteria may also be used before a variable is entered into the equation.) The process is then continued until enough variables have been entered or the inclusion criteria cannot be satisfied. Other techniques such as *backward elimination*, where one starts with all the variables in the equation deleting them one by one, or even more complex methods, may also be used to obtain the 'best' set of independent variables. Due to their computational complexity, multiple regressions are usually performed on a computer.

It must be stressed, however, that multiple regression techniques cannot be used blindly, including all possible variables in an analysis. In essence, a multiple regression analysis requires some model to be postulated that would determine beforehand which variables could be appropriately included on theoretical grounds.

When two independent variables are highly correlated, the effect of one on the dependent variable will often totally mask the effect of the other. This can result in one of these variables having a small and non-significant partial regression coefficient. How to deal with such *multicollinearity* is a problem. Usually, only one of a set of such variables will be included in the final equation, since one variable may act as a proxy for all the others. Variables excluded for such reasons, however, may have important effects on the dependent variable.

13.9 OTHER EXTENSIONS

So far, the regression techniques discussed have examined the relationship between a quantitative dependent variable, and one or more quantitative independent variables. When a quantitative independent variable is also to be included in such an analysis, the multiple regression approach can be modified to deal with the situation. This results in some cases in an *analysis of covariance* or combinations of both analysis of variance and multiple regression. Another method is to use *logistic* or *multiple logistic regression* techniques. Basically, one works with a transformed dependent variable, running a multiple regression on the transformed variable. These sophisticated but relevant techniques are beyond the scope of this book but are described in several excellent textbooks.

QUESTIONS

1. Define correlation. How is it measured and interpreted?
2. What are the uses and restrictions of regression equations?
3. Write brief notes on:
   (a) Scatter diagram
   (b) Spurious correlation
   (c) Multiple regression.
4. Calculate the correlation coefficient for the following values of Apgar score and birth weight (kg) of 10 infants and interpret.

Apgar score	3	2	3	4	5	2	3	5	4	5
Birth weight	3.6	2.9	3.2	3.8	3.9	3	3.5	3.8	3.9	3.9

CHAPTER

14 The Chi-Square Test

14.1 INTRODUCTION

Investigations often deal with testing a certain hypothesis regarding the population ratio. Thus, a physician may have a hypothesis that a certain disease requiring hospitalisation is equally common among men and women. In a sample of 900 hospital cases he finds 480 men and 420 women. Do these results support or contradict his hypothesis? (Incidentally, this is an example in which the sampled population may differ from the target population. Although good medical practice may prescribe hospitalisation, there are often cases which for one reason or another do not come to a hospital and, therefore, could not be included in his sample).

To answer such questions two results are needed, a measure of the deviation of the sample from the hypothetical population ratio, and a means of judging whether this measure is an amount that would commonly occur in sampling, or on the contrary, is so great as to throw doubt upon the hypothesis. Both results were furnished by Karl Pearson in 1889. He devised an index of dispersion or test criterion denoted as *chi-square* and obtained the formula for its theoretical frequency distribution, when the hypothesis in question is true. Like the binomial distribution, the chi-square distribution is another of the basic theoretical distributions much used in statistical work.

14.2 THE FORMULA FOR CHI-SQUARE

The deviations of the *observed* numbers from those specified by the hypothesis form the basis of the chi-square. Each deviation is squared, each square is divided by the *hypothetical* or *expected* number, and the results are added. The expected numbers appear in the denominators in order to introduce sample size into the quantity—it is the relative size that is important.

In general terms, if O_1 and O_2 are the sample counts of individuals who do and do not possess the attribute being investigated, and if the corresponding hypothetical or expected frequencies are E_1 and E_2, then the two deviations are $O_1 - E_1$ and $O_2 - E_2$, so that chi-square is given by the formula

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \Sigma \frac{(O - E)^2}{E}$$

where Σ denotes summation.

14.3 DISTRIBUTION OF CHI-SQUARE (χ^2) AND DEGREES OF FREEDOM

It is obvious that greater differences between the two percentages in the two categories will be reflected in higher values of chi-square. The chance of this observed or greater deviation is given by the chance of getting this or greater value for chi-square. This chance can be determined by referring to the table of χ^2 (Appendix 8) where the probability of obtaining a given value of χ^2 or greater is computed for different degrees of freedom (d.f.). Given a fixed total, the number of independent cells or cells that can be varied freely will be the number of degrees of freedom.

In the example of 14.1, with 900 cases, the numbers of male or female cases expected on the hypothesis are each 450. The deviations, then, are

$$480 - 450 = +30$$

$$420 - 450 = -30$$

the sum of the two being zero. The value of chi-square is given by

$$\chi^2 = \frac{(+30)^2}{450} + \frac{(-30)^2}{450} = 2 + 2 = 4$$

The observed χ^2 of 4 is greater than the 5% level of χ^2 with one degree of freedom (3.84, Appendix 8). So, we conclude that hospitalisation is not equally common among men and women ($P < 0.05$).

In the above example, there are only 2 cells of which only one can be varied freely, and hence the degrees of freedom (df) will be only 1. The Chi-square test can easily be extended to any number of rows(r) and columns(c), and the degrees of freedom can be calculated as $(r-1) \times (c-1)$, and the appropriate P values calculated from the table of Chi-square.

14.4 SOME APPLICATIONS OF CHI-SQUARE TEST

The Chi-square test is useful in at least 3 types of problems, viz. in testing for independence (or association) between 2 variables, in testing for homogeneity (or similarity), and for testing the goodness-of-fit between an observed and expected distribution. Each type is illustrated with an example.

14.4.1 Testing Independence (or Association)

This is by far the most common use.

Example The severity of a disease and blood type were studied in a research project. The findings are given in Table 14.1, known as the 2 × 2 *contingency table.*

It will be noted that 9.4% of the 'O' group had severe disease as against 6.1% of the others. Can this difference be attributed to sampling error, or does it indicate a real association of 'O' groups with severity of the disease? The null hypothesis is that the proportions severe, 51/540 and 59/960, are estimates of the same quantity or there is *no* association between 'O' groups and severity of disease.

Table 14.1 Severity of a Disease Classified by Blood Group in 1500 Patients

Severity	Blood groups		Total
	'O'	Others	
Yes	51	59	110
No	489	901	1390
Total	540	960	1500
Severity %	9.4	6.1	7.3

The test can be performed by χ^2. As usual,

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

where the O's are the observed numbers 51, 59, 489, 901 in the four cells. The E's are the numbers that would be expected in the four cells if the null hypothesis were true.

The E's are computed as follows: If the proportions 'severe' are the same for the two groups based on blood type, our best estimate of this proportion is the proportion, 110/1500 found in the combined sample. Since there are 540 of 'O' group, the expected number severe, on the null hypothesis, is

$$\frac{(540)\,(110)}{1500} = 39.6$$

The rule is: to find the expected number in any cell, multiply the corresponding column and row totals and divide by the grand total. The expected number of 'O' group who are not severe is

$$\frac{(540)\,(1390)}{1500} = 500.4$$

and so on. Alternatively, having calculated 39.6 as the expected number of 'O' group who are severe, the expected number not severe is found more easily as 540 – 39.6 = 500.4. Similarly, the expected number of severe in non-O group will be 110 – 39.6 = 70.4. Finally, the expected number of non-O group who are not severe is 960 – 70.4 = 889.6. Thus, only one expected number need be calculated; the others are found by subtraction. The observed numbers, expected numbers and the differences $(O - E)$ appear in Table 14.2.

Except for their signs, all four deviations $(O - E)$ are equal. This result is true for any 2 × 2 table.

Table 14.2 Values of *O* (Observed), *E* (Expected) and (*O* – *E*) in the Four Cells

O		*E*		*O* – *E*
51	59	39.6	70.4	11.4 – 11.4
489	901	500.4	889.6	– 11.4 + 11.4

Since $(O - E)$, without sign, is the same in all cells, χ^2 may be written as

$$\chi^2 = (O-E)^2 \sum_{i=1}^{4} \frac{1}{E} = (11.4)^2 \left[\frac{1}{39.6} + \frac{1}{70.4} + \frac{1}{500.4} + \frac{1}{889.6}\right]$$

$$= (129.96)\ (0.0426) = 5.536$$

A table of reciprocals is useful in this calculation, since the four reciprocals can be added directly.

How many degrees of freedom has χ^2? Since all four deviations are the same except for sign, this suggests that it has only 1 d.f., as was proved by Fisher. With 1 d.f., the table of χ^2 (Appendix 8) shows that a value of χ^2 greater than 5.54 occurs with probability between 0.02 and 0.01. Thus, the observed association between severity of disease and blood group cannot be explained as being due to sampling errors. Therefore, we conclude that the two factors are associated and not independent.

Sometimes an alternative 'short-cut' formula for computing may be preferred. The 2 × 2 table may be represented as follows:

a	b	$a + b$
c	d	$c + d$
$a + c$	$b + d$	$a + b + c + d = N$

Then,

$$\chi^2 = \frac{N\,(ad - bc)^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)} = \frac{1500\,(51 \times 901 - 59 \times 489)}{110 \times 1390 \times 540 \times 960}$$

$$= 1500\ (45951 - 28851) = 1500\ (17100)$$

$$= 5.536$$

which is the same result as obtained earlier.

14.4.2 Test of Homogeneity (or Similarity)

A second common use of chi-square test is in the testing of homogeneity or similarity between frequency distributions or groups. One common example is in assessing the similarity between nonresponders and responders in any survey. This is illustrated in Table 14.3.

Table 14.3 Age Distribution of Responders and Nonresponders

Age (years)	Responders		Nonresponders		Total
<20	76	(82)*	20	(14)	96
20–29	288	(289)	50	(49)	338
30–39	312	(310)	51	(53)	363
40–49	187	(185)	30	(32)	217
50 and above	77	(73)	9	(13)	86
Total	940		160		1100

*Expected values are shown in brackets.

$$\chi^2 = \frac{(-6)^2}{82} + \frac{(+6)^2}{14} + \frac{(-1)^2}{289} + \frac{(+1)^2}{49} + \frac{(+2)^2}{310} + \frac{(-2)^2}{53} + \frac{(-2)^2}{185} + \frac{(-2)^2}{32} + \frac{(+4)^2}{73} + \frac{(-4)^2}{13}$$

$$= 0.439 + 2.571 + 0.003 + 0.020 + 0.013 + 0.075 + 0.022 + 0.125 + 0.219 + 1.231$$

$$= 4.718$$

$$\chi^2_{4,0.05} = 9.49$$

Since the computed χ^2 is less than the tabled value, the null hypothesis is not rejected. Thus, it is concluded that the distributions are similar, i.e., nonresponders do not differ from responders.

14.4.3 Test for Goodness-of-Fit

The third of Chi-square test is in determining whether an observed pattern is as good as a hypothetical or theoretical pattern expected on the basis of some experience. Sometimes, based on the observed frequencies, a theoretical distribution is fitted, and the expected frequencies obtained. Then for each point in the curve, the difference between Observed and Expected frequencies computed, and the Chi-square test used as usual, and interpreted based on the Table of Chi-square (Appendix 8) with appropriate degrees of freedom. An example is given.

The incidence of Diarrhoea among children in a village was noted each month as follows:

The Chi-square test calculations are also shown along with the observed values, on the null hypothesis that there is no seasonal variation.

Table 14.4 Seasonal Variation in Childhood Diarrhoea

Month	Observed	Expected	$O - E$	$(O - E)^2$	$(O - E)^2/E$
Jan	8	6	2	4	0.67
Feb	9	6	3	9	1.50
Mar	12	6	6	36	6.00
Apr	4	6	2	4	0.67
May	2	6	4	16	2.67
Jun	3	6	3	9	1.50
Jul	6	6	0	0	0.00
Aug	9	6	3	9	1.50
Sep	10	6	4	16	2.67
Oct	4	6	2	4	0.67
Nov	3	6	3	9	1.50
Dec	2	6	4	16	2.67

The sum of the last column gives the Chi-square value as 22.02. Since there are 12 months, the degrees of freedom will be 11. Referring to the Table of Chi-square (Appendix 8), our value lies between the probabilities for 0.05 and 0.01.

Hence we reject the null hypothesis of no seasonal variation, with $0.02 < P < 0.05$.

14.5 YATES CORRECTION

The above χ^2 has not been corrected for continuity. A correction is appropriate because the exact distribution of χ^2 is continuous and a 2 × 2 table is discrete. The correction for continuity is made by deducting 0.5 from $| O - E |$. The formula for corrected χ^2 is given by Yates as follows:

$$\chi^2 = \Sigma \frac{(|\text{Observed value} - \text{Expected value}| - 0.5)^2}{\text{Expected value}}$$

In small samples, the correction makes a substantial difference. In the short-cut formula,

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

The subtraction of $N/2$ represents the *correction for continuity*. For the above problem (Table 14.1)

$$\chi^2 = 5.059, \quad 0.2 < P < 0.05$$

Still significant at 5% level but not at 2%.

14.6 EXTENSION OF 2 × 2 TABLE

The Chi-square test is so versatile that the formula can be extended to any number of rows and columns, depending on how the variables studied can be subdivided. For example, in testing for independence of association between Socioeconomic status (5 levels), and the performance of children in school (3 levels), we would have a 3 × 5 table with observed values which could be contrasted with expected values based on some hypothesis. As mentioned earlier, the degrees of freedom will be the number of cells which can be freely filled in, given the marginal totals, and is given by the formula $(r - 1) \times (c - 1)$, in this case, $(3 - 1) \times (5 - 1)$ or $2 \times 4 = 8$ degrees of freedom.

Another worked example is given.

Considering the earlier example, extension of blood groups and severity into more categories resulted in Table 14.5.

Table 14.5 Associations between Blood Group and Severity

	Blood group				
Condition	O	A	B	AB	Total
Severe	51	40	10	9	110
	(39.6)	(49.1)	(11.7)	(9.5)	
Moderate	105	103	25	17	250
	(90.0)	(111.7)	(26.7)	(21.7)	
Mild	384	527	125	104	1140
	(410.4)	(509.2)	(121.6)	(98.8)	
Total	540	670	160	130	1500

Note: The expected frequencies are in parenthesis.

The null hypothesis is that the severity of condition is independent of the blood group. Then,

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

$$= \frac{(51-39.6)^2}{39.6} + \frac{(40-49.1)^2}{49.1} + \cdots + \frac{(104-98.8)^2}{98.8}$$

$$= 3.2818 + 1.6866 + \cdots + 0.2737$$

$$= 13.35$$

Degrees of freedom (d.f.) $= (r-1)(c-1)$

$$= (3-1)(4-1) = 6$$

$\chi^2_{6,0.05} = 12.59$; so REJECT the null hypothesis, since computed χ^2 is more than the tabled value at 5% level of significance. Thus, we conclude that the data are compatible with the hypothesis that severity of the condition and blood types are associated and not independent.

14.7 PRECAUTIONS IN USING THE CHI-SQUARE TEST

A few precautions are to be taken in the proper use of the chi-square test.

1. The calculation of χ^2 must be for the actual number observed in each category; neither proportion nor percentage must be used in its place. Suppose, we are told that 80% of the persons in a sample are abnormal, and asked to compute chi-square. If we guess that the sample contained 100 persons, then

$$\chi^2 = \frac{(80-75)^2}{75} + \frac{(20-25)^2}{25} = \frac{25}{75} + \frac{25}{25} = 1.33$$

But if the sample actually contained only 10 persons, then

$$\chi^2 = \frac{(8-7.5)^2}{7.5} + \frac{(2-2.5)^2}{2.5} = \frac{0.25}{7.5} + \frac{0.25}{2.5} = 0.133$$

If the sample had 1000 persons, a similar calculation finds $\chi^2 = 13.33$. For a given percentage abnormal, the value of chi-square can be anything from almost zero to a very large number.

2. The χ^2 test is only a test of significance. It is not a measure of association. Thus, a large χ^2 is significant only in terms of a low probability of obtaining such a result by chance alone and not that there is a greater association.

3. The χ^2 test is a large-sample approximation, based on the assumption that the distributions of the observed numbers O_i in the classes are not far from normal. This assumption fails when some or all of the observed numbers are very small. Historically, the advice most often given was that the expected number in any class should not be less than 5, and that, if necessary, neighbouring classes should be combined to meet this requirement. Later research showed that this restriction is

too strict. Moreover, the combination of classes weakens the sensitivity of the χ^2 test. If more than one cell contains an expected number of less than 1, the χ^2 test should not be applied.

4. In an investigation involving 2 categories for each of the two characteristics (i.e., 2 categories in the rows and two categories in the columns), Yates correction for continuity must be applied especially when the numbers involved in the study are small.

5. For all the observations in the data, uniform classification must have been adopted. Each observation must have been independent of the others.

6. The investigator must ensure that no bias—personal or any other, is present in the data. In interpreting the results of these χ^2 tests in non-experimental studies, caution is necessary, particularly when χ^2 is significant. The two groups being compared may differ in numerous ways, some of which may be wholly or partly responsible for an observed significant difference.

Textbooks on statistics may be referred to for other uses of χ^2, and also for methods of analysis when the number of observations is small.

Several forms of χ^2 tables have been published and one variety is shown in condensed form in Appendix 8. The body of the table shows the value of χ^2 corresponding to specified degrees of freedom and the probability of obtaining that value or greater. For example, with 1 d.f. (degree of freeom) the probability of obtaining a value of $\chi^2 = 4.3$ lies between 0.05 and 0.025. With 6 d.f., the probability of obtaining the same value of $\chi^2 = 4.3$ is greater than 0.5.

Symbolically, sometimes the degrees of freedom and a given probability level are shown as subscripts. Thus, $\chi^2_{0.05,(6)}$ indicates the χ^2 value with 6 d.f. and $P = 0.05$.

The mean of a χ^2 distribution equals its degrees of freedom Thus, the mean of $\chi^2_{(6)}$ distribution is 6. In the table, for $\chi^2_{(6)}$, the 5% value is 5.35. This illustrates the asymmetry of the χ^2 distribution, the mean being to the *right* of the median.

QUESTIONS

1. Explain three uses of Chi-square test giving an example for each.
2. What are the precautions to be considered while interpreting the results of Chi-square Test?
3. Write short notes on:
   (a) Yate's correction
   (b) Chi-square test for goodness of fit
4. Among 50 rural children 20 were malnourished and among 50 urban children 10 were malnourished. Test whether malnutrition and area of residence are associated.

CHAPTER

15 Analysis of Variance

15.1 INTRODUCTION

In the chapter on tests of significance, the *t*-distribution was used in the comparison of means of two samples. Investigations, however, may not always be confined to comparisons of two samples only. For example, more than two food supplements may be tested, each on a different group of children; certain treatments may be tried on patients at different stages of a disease. The statistical inference has to be made from the responses of all groups of children or categories of patients. A statistical technique known as Analysis of Variance (ANOVA) is used when such simultaneous comparisons are made of measurements from more than two samples. When measurement data are influenced by several kinds of effects operating simultaneously, this statistical technique is adopted to decide which effects are important, and to estimate such effects.

15.2 THE *F*-TEST

We shall consider a simple situation of comparing the means of samples from two or more populations.

Assume that the groups which are studied represent random samples from statistically identical populations. In this case, each sample will provide an estimate of the same population parameter. The variance of the population can be estimated from each sample. Combining all these estimates, we can compute a *single* estimate of variance of the population: this is called the *mean square within samples*. We found in the chapter dealing with sampling distribution that means of random samples from the same population exhibit variation and that the variation of sample means is related to the variation of the individual values in the population and the sample size. Utilising this relationship, we can compute from the sample means another estimate of the population variance: this is known as the *mean square between samples*.

When our assumption that the samples represent samples from identical populations is true, these two estimates of the population variance are comparable. Any difference observed between them is expected to be within the range of their sampling errors. The following *variance ratio* (*F*) has under the above assumption, the sampling distribution known as Snedecor's *F*:

$$F = \frac{\text{Mean square between samples}}{\text{Mean square within samples}}$$

When the populations from which the samples were drawn have different means, the mean square between samples is expected to provide a higher value reflecting, in addition to the population variance, the difference between the population means also.

The results from an actual study are analysed and the observed variance ratio, F_{obs} is calculated. If the probability of F as found from the Table of F-Distribution, (F_{theor}) is small, i.e., less than say, 0.05, the study suggests existence of significant difference between the population means. Further analysis may be done to investigate the differences detected.

The analysis of variance technique can also be used when more than one factor affects the responses.

15.3 AN ILLUSTRATION OF ONE-WAY ANOVA

The method of calculation and drawing of inference is illustrated with the following example: The haemoglobin levels of three groups of children fed three different diets are given in Table 15.1. Test whether the means of these three groups differ significantly.

Table 15.1 Haemoglobin Levels (g%) of Children Fed with Three Different Diets

	Group I	Group II	Group III
	11.6	11.2	9.8
	10.3	8.9	9.7
	10.0	9.2	11.5
	11.5	8.8	11.6
	11.8	8.4	10.8
	11.8	9.1	9.1
	12.1	6.3	10.5
	10.8	9.3	10.0
	11.9	7.8	12.4
	10.7	8.8	10.7
	11.5	10.0	
		9.7	
No. of subjects	11	12	10
Total	124.0	107.5	106.1
Mean	11.27	8.96	10.61
		Total No. of subjects	33
		Grand total	337.6
		Common mean	10.23

Question to be answered. Do the mean haemoglobin levels of the three groups differ significantly?

In other words, is the variation of the mean levels of the different groups as measured from the common $\bar{x}$, significant? (the mean of the whole set of 33 haemoglobin levels is $\bar{x} = 10.23$).

Null hypothesis. The three groups are samples from the same population.

Procedure. First, we have to calculate three sums of squares:

1. Total sum of squares: i.e., sum of squares of deviations from the common mean.

Sum for the 33 individuals = Σx = 11.6 + 10.3 + 10.0 + ... + 12.4 + 10.7

= 337.6

Sum of squares of individuals = Σx^2 = $(11.6)^2 + (10.3)^2 + (10.0)^2 + ...$
$+ (12.4)^2 + (10.7)^2$

= 3516.32

$$\text{Correction term} = \frac{(\Sigma x)^2}{n} = \frac{(337.6)^2}{33} = 3453.75$$

$$\text{Total sum of squares} = \Sigma (x - \bar{x})^2$$

$$= \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 3516.32 - 3453.75 = 62.57$$

2. Sum of squares between groups, i.e., between the common mean and the mean for each group, with due allowance for the number of observations in each group.

Square the total for each group and divide each square by the number of observations in that group. Total the quotients and subtract the correction term.

Sum of squares between groups is

$$\frac{124.0^2}{11} + \frac{107.5^2}{12} + \frac{106.1^2}{10} - 3453.75 = 32.81$$

3. Sum of squares within groups is found by subtracting item 2 from item 1, i.e., sum of squares within group = total sum of squares – between group sum of squares = 62.57 – 32.81 = 29.76.

Degrees of freedom. As in the *t*-test, the degrees of freedom are one less than the number of items that compose the sum of squares.

1. Total sum of squares: No. of observations = 33; d.f. 32.

2. Sum of squares between groups: No. of groups = 3; d.f. 2.

3. Sum of squares within groups: This is got by subtracting 2 from 1, i.e., 32 – 2 = 30.

Mean Square = Sum of squares/No. of degrees of freedom. These results are shown in Table 15.2.

Table 15.2 The Analysis of Variance Table

Variation	Sum of squares	d.f.	Mean square
Between groups	32.81	2	16.405
Within groups (Residual)	29.76	30	0.992
Total	62.57	32	. .

$$\text{Variance Ratio, } F_{obs} = \frac{\text{Mean square between groups}}{\text{Mean square within groups (Residual)}}$$

$$= \frac{16.405}{0.992} = 16.54$$

Comparison with the theoretical value. The observed variance ratio, F_{obs}, is distributed like F_{theor} and the 5% and 1% theoretical values can be had from the tables of F-distribution for the required degrees of freedom. (see Appendix 7). In our case, the degrees of freedom are $n_1 = 2$ and $n_2 = 30$. The table of F-distribution gives the 5% value as 3.32 and 1% value as 5.39. Our observed value 16.54 is greater than the 1% value. Therefore, we say that the difference between the groups is significant at the 1% level. This means that the probability of observing a variance ratio as much as 16.54 or greater is less than 0.01.

Inference. The between group variability is greater than the within group variability. From the F-test, we found that the difference between these two variations is significant ($P < 0.01$). Therefore, we reject the null hypothesis that all the three groups are from the same population; that is, there is evidence to show that all the three groups are not from populations having identical means.

15.4 TWO-WAY ANOVA

The basic principles of one way analysis of variance can be extended to two-way classification or two-factor analysis of variance. The mean squares are calculated in a similar manner but extended for two factors. The formulas needed to calculate sums of squares and other statistics are not given here as computer programs are extensively used for ANOVA. Two examples are given here.

Example 1. The number of days of hospitalization in a year is classified by age-group and gender to find out if the hospitalization days differ by age and by gender. The data are given in Table 15.3.

Table 15.3 Days of Hospitalization during a Year

Gender	Age group				
	20–29	30–39	40–49	50–59	60+
Males	8	8	9	10	21
Females	10	11	10	12	17

Two null hypotheses are tested using the data.

1. The means of hospitalization days of all age-groups are equal.
2. The means of hospitalization days of the genders are equal.

The results of two-way analysis of variance are summarized in Table 15.4.

Table 15.4 Results of ANOVA of Hospitalization Days

Source of variation	Sum of squares	df	Mean square	F	Significance
Age-groups	141.400	4	35.350	9.182	0.027
Gender	1.600	1	1.600	0.416	0.554
Residual	15.400	4	3.850		
Total	158.400	9	17.600		

Note: The last column, significance, gives the probability of obtaining the F assuming the null hypothesis is true.

Inference

1. *Variation between age groups:* As the probability of obtaining the F is very small (less than 0.05), the null hypothesis is rejected with $p = 0.027$. So the means of hospitalization days between different age-groups are significantly different.

2. *Variation between the genders:* As the probability of obtaining the F is not small (more than 0.05), the null hypothesis is not rejected. So the means of hospitalization days between the genders are not different.

Through ANOVA the inference obtained was that the means (of age-groups) are different but it does not specify which mean of the age-group is different from which mean of the age-group. To get this post-hoc contrast (available in the computer program) has to be performed.

In this example only one observation was made for a combination of age-group and gender. Usually, many observations are made in each age-group and gender combination. In this situation the interaction (joint effect) of age-group and gender may be studied.

Example 2. The hospitalization days of 30 persons were observed and classified by their age-group and gender. There were 3 observations in each age-group and gender combination. The data were analysed to find out if there is any difference between the age groups, difference between the genders and also presence of interaction between age-group and gender.

In addition to the two null hypotheses stated in the first example, another null hypothesis that there is no interaction between age-groups and gender is also tested simultaneously.

The analysis of the data resulted in Table 15.5.

Table 15.5 Results of ANOVA of Hospitalization Days

Source of variation	Sum of squares	df	Mean square	F	Significance
Age-groups	563.200	4	140.800	76.800	0.000
Gender	4.800	1	4.800	2.618	0.121
Age* Gender (Interaction)	29.200	4	7.300	3.982	0.016
Residual	36.667	20	1.833		
Total	633.867	29	21.857		

Inference

1. *Variation between age groups:* As the probability of obtaining the F is very small (less than 0.001), the null hypothesis is rejected with $p = 0.000$. So the means of hospitalization days between different age-groups are significantly different.

2. *Variation between the genders:* As the probability of obtaining the F is not small (more than 0.05), the null hypothesis is not rejected. So the means of hospitalization days between the genders are not different.

3. *Interaction between age and gender:* As the probability of obtaining the F is small (less than 0.05), the null hypothesis is rejected with $p = 0.016$. So the interaction between age and gender is significant.

Extension of ANOVA for more than 2-way (factors) can be attempted successfully, especially when good computer software is available. The analysis of variance can also be extended to more than one dependent variable. This is called multiple (multivariate) analysis of variance, MANOVA, the details of which are available in relevant literature.

15.5 ASSUMPTIONS OF ANOVA

Before carrying out any analysis of variance on an actual research problem you should reassure yourself that the following assumptions seem reasonable:

1. All ANOVAs require that sampling of individuals be at random. Adequate safeguards to ensure random sampling during the design of an experiment, or when sampling from natural populations, are essential.

2. The second assumption is that the items be independent. This may not be true if there is correlation in time or space.

3. Equality of variances in a group of samples is an important precondition for several statistical tests.

4. The residual component should be normally distributed.

5. The 'additivity' effects in a 2-way or higher order ANOVA should be small.

If the above assumptions are not fulfilled, you should carry out one of several possible alternative steps to remedy the situation. Sometimes, departure from the assumptions of analysis of variance can be rectified by 'transformation' of the original data by use of a new scale. When transformations are unable to make the data conform to the assumptions of analysis of variance, other techniques of analysis, analogous to the intended ANOVA must be employed.

QUESTIONS

1. Explain the uses of ANOVA giving an example.

2. What are the assumptions of ANOVA?

3. Write brief notes on

(a) Residual

(b) F-test

(c) Interaction

CHAPTER

16 Non-parametric or Distribution-free Statistical Tests

16.1 INTRODUCTION

The statistical methods of inference which were discussed in the preceding chapters make certain assumptions about the populations from which the samples are drawn. For example, the assumptions may be that the populations are normally distributed, have the same variance, etc. Population values, as we have seen, are known as *parameters*; the statistical techniques which make assumptions about the parameters are called *parametric* techniques.

It can be remembered that in the computations of the parametric tests, the arithmetic processes of addition, division, and multiplication were used. For instance, the computation of arithmetic mean required addition and then division. When the observations (or the values) in the sample studied are not truly numerical, such arithmetic processes will introduce distortions in the data. Only when all the assumptions involved in a test are true will the probability statement arrived at by the test be meaningful.

In many biological investigations, the research worker may not know the nature of the distribution or other required values of the population. On many occasions, the number of observations made for the study may not be sufficient to test the assumptions regarding the population. Also in many instances, when the observation is represented by a numerical figure, the scale of measurement may not be really numerical. Some biological measurements are necessarily crude in their early stages of investigation. Grades of severity of an illness and ranks given to anaesthetic agents for their effectiveness are examples of data which, though expressed numerically, do not possess the characteristics necessary for arithmetical processes. The parametric tests may not be suitable in such situations.

More recently, a large number of techniques of inference have been developed which do not make stringent assumptions about the population characteristics (or parameters). The investigator will have to choose, in the above circumstances, from the newer *distribution-free* or *non-parametric* techniques which impose much less qualifications. In many of the non-parametric tests, the data are changed from measurements or scores to ranks or even to signs.

16.2 ADVANTAGES OF NON-PARAMETRIC TESTS

1. Probability statements obtained from most non-parametric tests are exact probabilities.

2. If samples are of sizes as small as 6, there is no alternative to using a non-parametric test unless the nature of the population distribution is precisely known.

3. There are suitable tests for treating observations from samples drawn from several *different* populations.

4. Tests are available to treat data which are inherently in ranks as well as data whose seemingly numerical scores have only the strength of ranks.

5. Methods are available to treat data which are simple classificatory.

6. These tests are much easier to learn and apply than parametric tests.

16.3 DISADVANTAGES OF NON-PARAMETRIC TESTS

1. If all the assumptions of the parametric test are in fact met in the data, and if the measurement is of the required strength, then non-parametric tests are wasteful of data.

2. There are no non-parametric methods for testing interactions in the analysis of variance.

3. Tables of critical values may not be easily available.

16.4 SOME NON-PARAMETRIC TESTS

The circumstances in which some of the non-parametric tests are applicable are given below:

1. When we have to test an assumption about the population distribution with a random sample from the population:

(a) *Binomial Test* can be used when the data are in two categories and the sample size is small.

(b) *Chi-square Test* can be used when the data are in discrete categories (2 or more) and the expected frequencies in the sample in these categories (based on our assumption) are sufficiently large.

(c) *Kolmogorov–Smirnov Test* will be applicable when the variable has a continuous distribution.

2. When we have to test if two random samples are likely to have come from populations with the same mean:

(a) *Randomisation Test* can be used with small samples when the data consist of measurements in a numerical scale; and the *Kolmogorov–Smirnov Test* with weaker measurements.

(b) *Mann Whitney U Test* may be used with large samples when the data represent weaker mesurements.

(c) *Median Test* is used when the data cannot even be meaningfully ranked, and when it is possible only to dichotomise the observations as above or below the combined median. While applying the median test the *Fisher's Test* must be used to analyse the data if the combined sample size is small.

(d) When more than 2 samples are considered, the *Kruskal–Wallis Test* can be applied to test whether they all belong to the same population.

(e) *Fisher's Exact Probability Test* is used when the scores from the independent random samples all fall into one or other of mutually exclusive classes. The test determines whether the 2 groups differ in the proportion with which they fall into the 2 classifications.

3. When we have to find out the statistical significance of differences in matched pair comparisons:

(a) *McNemar Test* is used when the data are frequencies in different categories.

(b) *Sign Test* is used when the data are on a variable which have underlying continuity, but can be measured only in a gross way. In this case, the ranks within the pair are used. If, however, the differences observed for the various matched pairs can be meaningfully ranked, *Wilcoxon Test* can be applied.

(c) When the data are measurements in a numerical scale, and when the sample size is sufficiently small, the *randomisation test* can be applied.

4. When we have to measure the correlation as the observations are ranked:

(a) Kendall's τ (tau) or

(b) Spearman's ρ (rho) or r_s is used.

Details of these and many other non-parametric techniques are given in sufficient detail in several books.

16.5 SOME ILLUSTRATIONS

16.5.1 McNemar's Test for Concordance

Table 16.1 presents the ranks given by eight judges for two food preparations. Does the result provide sufficient evidence to say that one food preparation is preferred over the other preparation?

Table 16.1 Ranks Given to Two food Preparations by 8 Judges

Judge	Preparation I	Preparation II
1	2	1
2	1	2
3	1	2
4	1	2
5	2	1
6	2	1
7	1	2
8	1	2

The statistic used to test the null hypothesis that both preparations are preferred equally is:

$$\chi^2 = \frac{(|a-b|-1)^2}{a+b}$$

where

a = number favouring the first preparation

b = number favouring the second preparation

$$\chi^2 = \frac{(|5-3|-1)^2}{8} = \frac{1}{8} = 0.125$$

This statistic follows the χ^2-distribution with one degree of freedom. Referring to the distribution, it is found that the probability of the null hypothesis is more than 0.05. Therefore, the null hypothesis is not rejected on the basis of the evidence provided by the data.

16.5.2 Sign Test

As the name suggests, the sign test uses only the sign of the differences but not the quantity of the differences.

Example The blood pressure levels of 10 adults before and after administering a new anti hypertensive drug are given below. Instead of using the '*t*' test for paired observations, the sign test is used.

Use the + sign to indicate increase and – sign to indicate decrease.

Blood pressure level

Subject	Before	After	Sign
1	140	132	–
2	130	124	–
3	122	110	–
4	150	132	–
5	130	110	–
6	142	122	–
7	122	128	+
8	112	110	–
9	132	140	+
10	130	120	–

Number of – signs: 8

Number of + signs: 2

Null hypothesis: P(+) = P(–) = 0.5

Referring to Appendix 10, for $N = 10$ and $X = 2$ the corresponding probability is 0.055.

This indicates that the change in the blood pressure level was not statistically significant.

16.5.3 Wilcoxon's Signed-rank Test for Differences

In Table 16.2 is shown the duration of endurance of pain by eleven mice before and after administration of a drug (adrenaline 0.04 mg/20 gm body-weight). Is there sufficient evidence in the data to say that the drug increases the duration of endurance of pain?

Table 16.2 Calculations for Wilcoxon's Signed Rank Test

Duration of endurance of pain (seconds)				
Before drug	After drug	Difference (2) – (1)	Rank of difference ignoring signs	Ranks (4) with signs attached
(1)	(2)	(3)	(4)	(5)
15.5	21.2	+ 5.7	8	+ 8
12.7	20.1	+ 7.4	11	+ 11
14.8	17.2	+ 2.4	7	+ 7
16.7	22.7	+ 6.0	9	+ 9
20.1	20.0	– 0.1	1	– 1
22.0	19.8	– 2.2	6	– 6
20.2	19.8	– 0.4	3	– 3
18.1	18.8	+ 0.7	5	+ 5
17.6	17.9	+ 0.3	2	+ 2
17.4	24.3	+ 6.9	10	+ 10
19.1	18.6	– 0.5	4	– 4
Sum of the negative ranks = 14 Sum of the positive ranks = 52				

Ranks are assigned to the differences between measurements on the same mouse ignoring the sign of the difference (Column 4) and then the sign of the difference is added to the ranks (Column 5). The null hypothesis that the drug has no effect is tested using the smaller of the sums of positive and of negative ranks (T) which in this case is 14. Referring to the table (Appendix 11), we find that for the number of pairs (n) = 11, the critical value for T at $P = 0.05$ is 11. Greater values of T have higher probabilities. The value of T in our example being more than 11, we do not reject the null hypothesis.

16.5.4 The Kruskal–Wallis Test for Differences in more than Two Populations

The Kruskal–Wallis one-way analysis of variance by ranks is an extremely useful test for deciding whether k independent samples are from different populations.

Computation. If there are k samples, and varying number n_i is studied in each sample, let n = total studied = $\sum_{i=1}^{k} n_i$. Let all the values from all the k samples be combined and ranked in a single series. The smallest value is replaced by rank 1, the next to smallest by rank 2, and the largest by rank n. n = the total number of independent observations in the k samples. When this has been done, the sum of the ranks in each sample (column) is found. The Kruskal–Wallis test determines whether these sums of ranks are so disparate that they are not likely to have come from samples which were all drawn from the same population. It can be shown that if the k samples actually are from the same population or from identical populations, (Null hypothesis), the H (the statistic used in the Kruskal–Wallis test and defined by formula given below) is distributed as Chi-square with d.f. = $k - 1$, provided that the sizes of the various k samples are not too small. That is,

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

is distributed approximately as Chi-square with d.f. = $k - 1$, for sample sizes sufficiently large (at least 5 in each group). For smaller sample sizes, special tables are available. When ties occur between two or more values, each value is given the mean of the ranks for which it is tied. Since the value of H is somewhat influenced by ties, one may wish to correct for ties in computing H. To correct for the effect of ties, H is computed by the formula given and then divided by

$$1 - \frac{\Sigma T}{n^3 - n}$$

where $T = t^3 - t$ (when t is the number of tied observations in a tied group of values).

The effect of correcting for ties is to increase the value of H and thus to make the result more significant than it would have been if uncorrected. Therefore, if one is able to reject the null hypothesis without making the correction, one will be able to reject it at an even more stringent level of significance if the correction is used.

Example. In a study of cerebrovascular disease, patients from 3 socio-economic backgrounds were thoroughly investigated. One characteristic measured was the diastolic blood pressure (mm/Hg). Is there any reason to believe that the 3 groups differ with respect to this characteristic?

Group A	Group B	Group C
100	92	81
103	97	102
89	88	86
78	84	83
105	90	99
	95	

Null hypothesis. There is no difference among the diastolic pressures of the 3 groups.

There are 16 values. Pooling all these values, arranging them from lowest to highest, ranking them and putting them back into the 3 groups, we find the following ranks in each group:

A	B	C
13	9	2
15	11	14
7	6	5
1	4	3
16	8	12
	10	
$R_1 = 52$	$R_2 = 48$	$R_3 = 36$

$$H = \frac{12}{16(16+1)}\left[\frac{(52)^2}{5} + \frac{(48)^2}{6} + \frac{(36)^2}{5}\right] - 3(16+1)$$

$$= 52.2 - 51.0 = 1.2$$

H is distributed as a χ^2 with 2 d.f. and for the above value of H, P is > 0.5, i.e., the differences observed can occur by chance alone more than 50% of the time. The null hypothesis is, therefore, *not* rejected.

There are at least four other non-parametric tests for k independent samples. These four are rather specialised in their usefulness and the reader might find one of them most valuable in meeting certain specific statistical requirements. The Kruskal–Wallis test seems to be the most efficient of the non-parametric tests for k independent samples. It has power efficiency of $3/\pi = 95.5$ per cent, when compared with the F-test, the most powerful parametric test.

16.5.5 The Spearman's Rank Correlation Coefficient

Of the many non-parametric measures of correlation available, the Spearman's rank correlation coefficient, r_s is attractive because it is simple to calculate. The procedure consists of ranking the two sets of values of X and Y, and computing the difference d of each pair. The d's are then squared and added. The Spearman's rank correlation coefficient is defined as

$$r_s = 1 - \frac{6\,\Sigma d^2}{n\,(n^2 - 1)}$$

The computed value is compared with the critical values, shown in Appendix 12, taking into account n, the number of pairs. The calculations of r_s are shown in Table 16.3.

Tied observations present a problem and if there are many such problems, they must be suitably handled. Details are available in other textbooks.

Table 16.3 An Example of Calculating Rank Correlation between Fasting Blood Glucose Level and Systolic Blood Pressure in 10 Diabetics

Serial No. of patient	Fasting blood glucose level (mg/dl)	Rank (R_1)	Systolic BP level (mmHg)	Rank (R_2)	$d = R_1 - R_2$	d^2
1	90	1	136	3	− 2	4
2	92	2	140	4	−2	4
3	98	3	142	5	− 2	4
4	112	4	130	1	+ 3	9
5	120	5	148	7	− 2	4
6	121	6	135	2	+ 4	16
7	126	7	150	8	− 1	1
8	132	8	170	10	− 2	4
9	143	9	145	6	+ 3	9
10	145	10	165	9	+ 1	1

$$r_s = 1 - \frac{6 \times 56}{10 \times 99} = 1 - 0.339 = 0.661$$

Referring to Appendix 12, for $n = 10$, we find $0.02 < P < 0.05$. Therefore, the null hypothesis of no correlation is rejected and it is concluded that fasting blood glucose and systolic blood pressure are correlated.

QUESTIONS

1. What is a 'non-parametric' test? When is it used?
2. Explain the advantages and disadvantages of non-parametric test.
3. Describe the method of testing the concordance using a nonparametric technique.
4. Find the rank correlation coefficient from the following data of ranks given by two nutritionists for 10 children in a well-baby contest.

Child Name (coded)	A	B	C	D	E	F	G	H	I	J
Nutritionist 1	3	5	1	2	4	7	6	9	10	8
Nutritionist 2	4	3	2	1	6	5	7	9	8	10

CHAPTER

17 Vital Statistics

17.1 INTRODUCTION

In recent years, the term 'Vital Statistics' has been used to denote facts systematically collected and compiled in numerical form relating to or derived from records of vital events namely, live birth, death, foetal death, marriage, divorce, adoption, legitimation, recognition, annulment, or legal separation. These events may be required to be permanently registered with authorised offices for legal and documentary purposes as well as to provide national statistical data. In essence, vital statistics are derived from legally registrable events without including population data or morbidity statistics.

17.2 USES OF VITAL STATISTICS

In general, the uses of vital statistics may be summarised as follows:

1. to describe the level of community health;
2. to diagnose community ills;
3. to discover solutions to health problems and find clues for administrative action;
4. to determine priorities for health programmes;
5. to direct or maintain control during execution of programmes;
6. to develop procedures, definitions, classifications, and techniques such as recording systems, sampling schemes, etc.;
7. to promote health legislation;
8. to create administrative standards of health activities;
9. to determine the met and unmet health needs;
10. to disseminate reliable information on the health situation and health programmes;
11. to determine success or failure of specific health programmes or undertake overall evaluation of public health work; and
12. to demand public support for health work.

To take a specific example, if the disease or underlying cause for each death were carefully recorded for a city over a number of years, we could provide correct answers to the following questions:

1. What is the leading cause of death in this city? Is it malaria, tuberculosis, cancer, or accidents, or some other cause as yet unsuspected?

2. At what age is the mortality highest, and from what disease?

3. What sections of the city are the most unhealthy and what is the outstanding cause of death there?

4. How much reduction in mortality might one expect from such public health measures as measles vaccination or malaria eradication?

5. In comparison with cities of similar size in other countries, is this city healthy or not?

6. In what sections of the city, as judged by the death rate, is improvement being achieved? In which parts on the other hand, is disease on the increase?

7. Are certain diseases affecting specified groups of the population more than others? This might apply for example, to women or children or to individuals following a particular occupation.

8. To what extent are the hospital facilities, staffing, drugs, and equipment able to cope with the disease problem?

9. Is there any season at which various diseases have a tendency to break out? If so, can these be distinguished?

Similarly, the birth rates and their analysis give an indication of the changes due to births in the population and in its various segments. For example, we could provide correct answers to the following questions:

1. Is the rate at which births occur in the population the same for various socio-economic levels?

2. Is the rate at which births occur in a given area relatively high or low as compared to that in other areas?

3. What factors are associated with higher or lower birth rate?

4. What correlation exists between a given birth rate and a given death rate?

The data are studied first as actual counts (total population, births, deaths, etc.) to form an idea of the magnitude of the health problem and secondly as indices (rates, ratios, etc.) for purposes of making comparisons with other areas.

17.3 MECHANISM FOR COLLECTION OF VITAL STATISTICS

It is by no means an easy task to set up a system that will ensure accurate recording of information for each vital event occurring in a country, in each and every town, village, institution, and individual home. In many of the less developed countries a system of this kind does not exist; and even where considerable advances have been made, statistical services have expanded only after many years of effort, and improvements are still being effected. There are thus many countries in which

statistics have yet to be 'sold' to the government and the people. It has to be clearly understood that this is the only way in which the level of health can be measured, and effective comparisons made.

In countries where the machinery for the collection of statistics for rural areas has been set up, the village is generally the smallest administrative unit for which an agent is appointed to report the events, and in some cases he may also be responsible for registering them. The responsibility for registration in rural areas generally rests with the head of the household. The primary responsibility for each village may rest with the village headman, the police officer, village watchman or revenue official. Sometimes one of the local inhabitants is appointed registrar of vital statistics whose duty is not only to keep himself informed about all vital events occurring in the area, but also to be the custodian of legal records.

Regulations for reporting and registering vital events differ from one country to another.

The Health Survey and Development Committee (Bhore Committee) in 1946 brought out that vital statistics in India had three main defects, viz.,

1. incompleteness in the recording of the events;
2. inaccuracy in the cause of death; and
3. faulty compilation.

These defects were attributed to poor collection of vital occurrences at the periphery (in villages) by illiterate chowkidars or by the lowest grades of local body employees, registration by agencies such as police, village headman and the passage of the recorded vital statistics through a series of stages before they reached the Directorate of Public Health of each province. Thus, the data on vital events were poorly collected, improperly handled and scarcely used for either operating or evaluating projects in the country. The Bhore Committee gave several recommendations to rectify these defects.

Following the recommendations of the Health Survey and Development Committee in 1946, the Office of the Registrar General of India and Ex-officio Census Commissioner (RGI) was created in 1951 and population census and vital statistics were brought under his control for a coordinated development of population statistics. The registration of births and deaths was made compulsory and uniform all over India through central legislation with the enactment of the Registration of Births and Deaths Act, 1969. The Act also provides for a statutory authority at the Centre and in each State and prescribes definitions, principles, personnel and penalties connected with enforcement.

The existing registration systems for collection of vital statistics in India are described in the following sections.

17.3.1 Civil Registration System

The civil registration system is the traditional system for registration of vital events in India. Registration in rural areas is carried out by different agencies—Health, Revenue, Police or Panchayats—in different States. Local officials of the department are appointed as registrars and they carry out registration of births and deaths at

village level in addition to their normal duties. In urban areas, the responsibility for civil registration rests generally with the Health Department and health officials like sanitary inspectors, vaccinators and health assistants carry out the registration in municipal towns and cities.

Heads of households are responsible for reporting any vital event occurring in households to the registrar within a prescribed period. The registrar records all vital events occurring along with particulars such as date of occurrence of event, sex, age of mother, age at death, cause of death, etc., in the registers maintained separately for births, deaths and stillbirths for each calendar year.

The registrars send monthly returns to the Chief Registrar (generally Director of Economics and Statistics, Director of Health Services or Director of Panchayats) who is the chief executive authority in the state for registration, either directly or through a District Registrar or Taluk Officer. The Chief Registrar prepares a consolidated annual statistical report for the state government and transmits a copy to the Registrar General of India (RGI) who in turn consolidates the state reports.

Publications. At the All India level, the statistics based on civil registration are consolidated and published by the RGI annually in "Vital Statistics of India".

These statistics are deficient in quality due to under-registration of vital events and non-reporting by the registering units and gross misreporting of age. The extent of under-registration varies from state to state and as such spatial comparison is not advisable, but a temporal comparison can be made.

17.3.2 Sample Registration System

Sample Registration of Births and Deaths, i.e., Sample Registration System (SRS), launched in 1964–65 by RGI, is an important, source of vital rates in India.

The sample registration of births and deaths was initiated by the RGI in 1964–65 in order to provide reliable estimates of birth and death rates separately for rural and urban areas at the state and national levels. Besides, the scheme also provides a variety of other useful demographic data such as sex, age, marital status, medical attention, religion, etc.

Sample design. The sample design adopted was a unistage stratified simple random sample without replacement. The rural areas in each state were divided into a number of natural regions. The natural divisions conform to the regions adopted by the National Sample Survey Organisation in their 32nd round. These natural divisions or regions were formed by grouping contiguous districts having similar topography, population density, cropping pattern and rainfall, etc. Each natural division within a state has been stratified by grouping villages into four population size classes: (a) less than 500, (b) 500 to 999, (c) 1000 to 1999 and (d) 2000 and above. In order to cover the village by one part-time enumerator, large size villages having population of more than 2,000 were segmented into two or more segments of equal size. A simple random sample of villages and segments was selected without replacement in each State/Union Territory. In urban areas, stratification was done on the basis

of size class of the towns/cities. The towns/cities were grouped into five classes according to population size, viz: (a) Below 20,000, (b) 20,000–49,999, (c) 50,000–99,999, (d) 100,000–999,999, and (e) 1,000,000 and above. Each city with population of 1,000,000 or more was treated as a separate stratum. The sampling unit in urban areas was a Census Enumeration Block. A simple random sample of these enumeration blocks was selected without replacement from each of the size classes of towns/cities in each State/UT. The total sample size is 6022 (Rural 4149, Urban 1873).

Publications

(a) *Sample registration bulletin.* The "Sample Registration Bulletin" is published by the office of the RGI bi-annually in June and December. It gives the latest available half-yearly and/or annual estimates of birth and death rates at the state and national levels separately for rural and urban areas collected through the Sample Registration System (SRS).

Besides, related data such as estimates of mid-year population, age-specific fertility rates, infant mortality rates, and age-specific death rates are given and technical notes/papers on different aspects of SRS are also included.

(b) *Sample registration system.* This is an annual publication describing the inter-State and rural-urban differentials in age composition, fertility and mortality. It contains data on distribution of the population by broad age-groups, the general and age-specific rates of fertility and mortality. It also describes the type of medical attention at birth and before death.

17.3.3 Survey of Causes of Death (Model Registration Scheme)

The objectives of the survey of causes of death (initially known as "Model Registration Scheme") are to build up statistics of causes of death in rural India, using "Lay diagnosis" reporting methods. It covers headquarter villages (1033 villages in 1991) of selected Primary Health Centres (PHCs). The paramedical staff such as sanitary inspectors or health inspectors attached to the PHC are the primary workers, who upon being informed by their local contact men, visit households and collect information on symptoms, conditions, anatomical site and duration of illness. Based on such information, the major cause-group and sub-cause for death are determined. An independent six-monthly crosscheck of events is carried out by an agency other than the field agent to ascertain all the deaths in the village during the previous six months. These deaths are then matched with those recorded by the field visit. These statistics are then transmitted to the State headquarters, either the Directorate of Medical Services and Public Health or the Directorate of Economics and Statistics who compile the data at State level and in turn transmit them to the RGI for consolidation and analysis.

Publications. The publication "Annual Report of Survey of Causes of Death" is brought out by the Office of the RGI every year. This publication contains the percentage distribution of the deaths under major cause groups and sub-cause groups

within the major cause groups by sex and age for India. At State level, the percentage distribution is available only for major cause groups by age.

The major cause groups for which information is given are as follows:

1. Accidents and injuries
2. Childbirth and pregnancy
3. Fevers
4. Digestive disorders
5. Coughs (disorders of the respiratory system)
6. Disorders of the central nervous system
7. Diseases of the circulatory system
8. Other clear symptoms
9. Causes peculiar to infancy
10. Senility.

The data in this publication are based on ascertainment of cause of death by paramedical personnel through a non-medical list of causes of deaths based on symptoms etc., after the death. Hence, these are no substitute for statistics of the medically certified cause of death. Further, the data are obtained from selected villages which do not form a representative sample of villages or even of PHC headquarter villages and, therefore, cannot be generalised for all rural areas. The data published are thus not comparable with data on cause of death available from Civil Registration, National Sample Survey and hospital death records.

17.4 BASIC FORMULAE FOR CALCULATION OF VITAL STATISTICS

The number of events that occur in any place are to a large extent dependent on the people living in that area. Therefore, in assessing the state of health or change, the vital events must be considered in terms of the population that may have given rise to these events. Hence, vital statistics are expressed usually either as (a) 'rate' or as (b) 'ratio'. A ratio is a *static* quantity; it is a proportion such as a/b, describing the state of affairs at a particular instant. A rate on the other hand, describes the rapidity with which something is happening during a stated period of time. Thus, a rate is the proportion a/b per unit of time, usually one year.

Ideally, a rate should satisfy the following requirements:

1. The numerator should have arisen from the denominator.

2. Each one included in the denominator should contribute to, or have an opportunity to appear in, the numerator.

In reality, however, these two conditions are not strictly satisfied due to practical considerations, as will be discussed later.

Crude rates. A crude rate, sometimes known also as a general rate, total rate, or gross rate, is computed with respect to the total population, without regard to any

specific characteristic of that population. Such rates measure the proportion prevailing between the number of events under consideration and the total population. The general formula for a crude rate, irrespective of the vital event to which it refers, is:

$$\text{Crude rate per thousand} = \frac{\text{Number of events which occurred among the population of a given geographic area during a given period}}{\text{Mid-period total population of the given geographic area during the same period}} \times 1000$$

Specific rate. Rates defined in terms of one or more characteristics of the population are known as specific rates. Any relevant characteristic of the population may be the basis for a specific rate. Those found most useful and feasible are, age, sex, birth order, occupation, marital status, etc.

$$\text{Specific rate per thousand} = \frac{\text{Number of events which occurred among a specific group of the population of a given geographic area during a given period}}{\text{Mid-period population of the specified group in the given geographic area during the same period}} \times 1000$$

The specific rate is characterised by its reference to only one special segment of the population.

Annual rate. Although a crude rate or a specific rate may be computed for any period of time, it is a general rule to express it on an annual basis. An annual crude rate is obtained when the term 'period' in the numerator and denominator of the above definition of the crude rate is replaced by the term 'year'—preferably 'calendar year'. The same substitution in the definition of the specific rate defines the annual specific rate.

Should the period be more or less than one year, it is customary, in order to facilitate comparison, to adjust the rates to an annual basis. Adjustment is achieved by multiplying the rate by an appropriate factor obtained by dividing the year by the period in question.

Quinquennial rate. When long-term trends are studied, it is convenient to calculate rates for five yearly periods. If a quinquennial rate is required, the total events that occurred in a five-year period should be divided by five, to make the rate comparable with the annual rate. The general formula is:

$$\text{Quinquennial rate} = \frac{\text{Number of vital events which occurred among the population of a given geographic area during the period of five years}}{\text{Population estimated at the middle of the five years}} \times 1/5 \times 1000$$

Decennial rate. This is required when a ten-year period is considered:

$$\text{Decennial rate} = \frac{\text{Number of vital events which occurred among the population of a given geographic area during the period of ten years}}{\text{Population estimated at the middle of the ten years}} \times 1/10 \times 1000$$

Weekly rate. In the study of seasonal variation of diseases, rates are calculated for periods shorter than a calendar year. To express it on an annual basis, the weekly rate is multiplied by 52. The formula is:

$$\text{Weekly rate} = \frac{\text{Number of vital events occurring in one week}}{\text{Total mid-year estimated population}} \times 52 \times 1000$$

This is not much affected by the use of the estimated mid-year population in the denominator, as generally variations in population size are relatively small in the course of one year. Furthermore, as a rule only one annual estimate of the total population, i.e., the mid-year population, is published.

Monthly rate. Similarly, monthly rates are calculated by using an adjustment factor equal to 12 or more exactly, using the fraction

$$\frac{365}{\text{Number of days in the month considered}}$$

Quarterly rate. For computation of a quarterly rate, the adjustment factor is 4.

Half-yearly rate. Use is made of an adjustment factor of 2 for a half-yearly rate.

17.5 MORTALITY RATES

Crude death rate. To measure the decrease of population due to death the rate commonly used is the crude death rate described below:

$$\text{Crude death rate} = \frac{\text{Number of deaths which occurred among the population of a given geographic area during a given year}}{\text{Mid-year total population of the given geographic area during the same year}} \times 1000$$

Example. During 1992, there were 426 deaths in town 'A' and 504 deaths in town 'B'. The estimated mid-year populations for 1992 for towns A and B were 34080 and 42711 respectively. The crude death rates will be:

$$\text{Town A} = \frac{426}{34080} \times 1000 = 12.5 \text{ per } 1000.$$

$$\text{Town B} = \frac{504}{42711} \times 1000 = 11.8 \text{ per } 1000.$$

The crude death rate is perhaps the most widely used of all vital statistics. This is because it is relatively easy to compute, requiring only total deaths and total population, and it has value as an index to numerous demographic and public health problems. Within broad limits, the crude death rate represents the probability of death for persons in the population. As such, it can be interpreted in terms of public health since the rate of death is the first approximate measure of the health status of a population.

Care is needed in the use of the crude death rate for inter-area comparisons. Mortality varies with age. The crude rate which masks all age differences may, therefore, be misleading, if the age-sex structures of the populations being compared are not similar. Populations composed of a high proportion of persons of older ages, where mortality is higher will naturally show a higher crude death rate than a 'younger' population. Some deaths escape registration. The degree of under-registration of deaths also greatly influences the value of the rate.

The crude death rate is valid for comparisons for the same area from year to year since changes in the composition of the population of an area occur rather slowly. If a long-term trend is studied, the effect of population change must be allowed for. Where it is known that the population distributions are approximately similar, or where the crude rate differences are large as in many international comparisons, the crude rate has value as an index of mortality.

Specific death rate. Crude death rates alone are not enough for a detailed study of the mortality conditions in a community. We often need to know more about deaths occurring in different sections of the population. For instance, people interested in infant or child welfare work study deaths taking place under 1 year of age or in age groups such as 1–4 years, 5–9 years, etc. Those interested in maternal health, study how many deaths occurred among women of child-bearing age. Insurance companies are interested in deaths occurring at different ages of the population. Public health officers study deaths caused separately by different diseases or injuries. All these require separate classifications of the total number of deaths.

Death rates specific for age, sex, etc., can be calculated in a manner similar to the crude death rate. Computation of age-specific death rates is shown in the following example:

Example. The mid-year estimated populations in an area for different age groups and the numbers of deaths occurring in each of them are given in Table 17.1.

Table 17.1 Population and Deaths by Age

Age (years)	Mid-year estimated population	Number of deaths	Age-specific death rate per 1000
Under 5	11,150,000	223,946	20.1
5–9	9,560,000	19,998	2.1
10–14	8,710,000	10,339	1.2
15–19	8,530,000	21,472	2.5
20–24	7,710,000	36,124	4.7
25–29	6,160,000	33,096	5.4
30–34	5,190,300	26,600	5.1
35–39	5,060,000	28,364	5.6
40–44	4,490,000	29,503	6.6
45–49	4,000,000	34,082	8.5
50–54	3,390,000	40,465	11.9
55–59	2,740,000	48,533	17.7
60–64	2,290,000		
65–69	1,760,000		
70–74	1,290,000	355,602	55.7
75 and above	1,040,000		

Specific death rate per 1000 population for age-group under 5 years is

$$\frac{223,946}{11,150,000} \times 1000 = 20.1 \text{ per } 1000$$

Rates computed for the entire life span invariably show high mortality operating at the two ends of life, i.e., among infants and old persons.

Specific death rate by sex. When age-specific death rates are calculated separately for each sex, the mortality is generally at a lower level among females than among males. However, in the reproductive age-groups mortality among women may be higher than among men because of lack of medical care of the mother.

Cause-of-death rate. Most death rates by cause are computed on total population; the exceptions being the infant mortality rate and some age, and sex-specific rates. The formula for cause-of-death rate is:

$$\text{Annual cause-of-death rate} = \frac{\text{Number of deaths from a specified cause which occurred among the population of a given geographic area during a given year}}{\text{Mid-year total population of the given geographic area during the same year}} \times 100,000$$

The numerator of this rate is often small. For this reason, the formula uses a multiplying factor of 100,000 rather than the usual 1000 so that the computed rates are not expressed as small decimal fractions.

Example. There were 52 deaths due to diarrhoeal diseases in Town 'A' during 1992. The estimated mid-year population for 1992 was 34080. Therefore:

$$\text{The death rate due to diarrhoeal diseases} = \frac{52 \times 100{,}000}{34080} = 152.6 \text{ per } 100{,}000.$$

We can calculate cause-of-death rates specific for age and sex as in the example below:

27 of those deaths occurred among children under 10 years, of which 15 were boys and 12 girls. The estimated mid-year populations in 1992 for these 2 groups are 6200 and 5700 respectively. Then,

Specific death rates for diarrhoeal diseases among boys under 10 years = 15 × 100,000/6200 = 241.9 per 100,000.

Specific death rates for diarrhoeal diseases among girls under 10 years = 12 × 100,000/5700 = 210.5 per 100,000.

Infant mortality rate. For a health worker, mortality rate among infants is of special significance because it is regarded as one of the most sensitive indexes of health conditions of the general population. It is a sensitive measure of health because a baby in its extrauterine life, is suddenly exposed to a multitude of new environmental factors and their reactions are reflected in this rate. Under ideal conditions of social welfare, no baby should die, except possibly the few who are born with serious handicaps such as congenital disabilities or disorders originating in uterine life. The computation of the infant mortality rate on the basis of the population of infants presents difficulty because the infants are usually under-enumerated. Sometimes, the estimates of the population by age are not available. For the calculation of this rate, the population of infants in the denominator is, therefore, substituted by the number of live births. The formula for its calculation is:

$$\text{Annual infant mortality rate} = \frac{\text{Number of deaths under one year of age which occurred among the population of a given geographic area during a given year}}{\text{Number of live births which occurred among the population of the given geographic area during the same year}} \times 1000$$

Example. The number of live births during 1990 in a rural area was 4757. The number of infant deaths during the same year was 543. The infant mortality rate for 1990 would, therefore, be

$$\frac{543 \times 1000}{4757} = 114.1 \text{ per } 1000 \text{ live births}$$

Comparison of infant mortality rates enables one to locate populations where the rate is high and thus provides indication of the need for special action. For example, in many parts of the world, where the infant mortality rate is computed separately for urban and rural areas, the unequal distribution of health facilities is brought out sharply. But in calculating the rate for urban populations, it has to be noted that frequently, owing to better hospital facilities in cities or towns, maternity cases are attracted from the adjoining rural areas. We should, therefore, exclude all

births occurring in urban maternity institutions if the normal place of residence of the mother is outside the area.

The infant mortality rate is studied separately for geographical sub-divisions or for specific population groups. As a matter of fact, it is a problem that needs study and attack on a wider front, and the health worker should consider the effect of such factors as overcrowding, unemployment, social habits and customs, food habits, topography, occupation of parents, and illegitimacy.

It is universally true that mortality in the first year of human life is high, higher than at any age except the very advanced. In many countries, it is about 20 to 50 times higher than in the 10–15 year age-group. In all the advanced countries there has been a marked reduction in the infant mortality rate during recent years.

Since the infant mortality rate is of special significance in health work, its correct computation and interpretation is important. Errors in the rate arise from wrong use of the definition or under-registration of vital events. For instance, variations in the definition of the terms 'live birth' and 'foetal death' may produce appreciable differences. Live-born infants who die at birth or very soon after, may fail to be registered as live births. On the other hand, by improving the registration of live births, a higher figure in the denominator will automatically produce a reduced infant death rate. It has, therefore, been said that (statistically speaking), it is possible to lower the infant mortality rate even without a reduction in infant deaths.

Neonatal mortality rate. Infant mortality is studied in detail by classifying deaths first by the age of the infant. In accordance with the WHO Regulations No. 1, deaths in the first year of life are classified by single days for the first week of life, by weeks up to the fourth, and then by single months up to age 1 year. For deaths occurring under 28 days, a rate called the neonatal mortality rate is calculated:

$$\text{Annual neonatal mortality rate} = \frac{\text{Number of deaths under 28 days of age which occurred among the population of a given geographic area during a given year}}{\text{Number of live births which occurred among the population of the given geographic area during the same year}} \times 1000$$

Post-neonatal mortality rate or late infant mortality rate

$$\text{Post-neonatal mortality rate (or late infant mortality rate)} = \frac{\text{Deaths of infants of age 28 days to under 1 year}}{\text{Live births minus neonatal deaths}} \times 1000$$

The infant mortality rate is then approximately the sum of the neonatal and the post-neonatal mortality rates.

Foetal death ratio. For international comparability, the WHO Expert Committee on Health Statistics has recommended that foetal death ratio be computed on the basis of live births. As the registration of all foetal deaths is incomplete, better figures are obtained if the calculation of this ratio is limited to late foetal deaths (still-births). This ratio relates the number of late foetal deaths to the number of live births.

$$\text{Annual late foetal death ratio} = \frac{\text{Number of foetal deaths of 28 or more completed weeks of gestation which occurred among the population of a given geographic area during a given year}}{\text{Number of live births which occurred among the population of the given geographic area during the same year}} \times 1000$$

Another name for this ratio is the 'still-birth ratio'.

Example. The number of still-births registered from a rural area during 1990 was 128. During the same year, there were 4757 live births.

The *still-birth ratio* for 1990 will be:

$$\frac{128 \times 1000}{4757} = 26.9$$

Occasionally another measure defined as the still-birth rate is used which is:

$$\text{Still-birth rate} = \frac{\text{Number of still-births which occurred among the population of a given geographic area during a given year}}{\text{Total number of births (live plus still) of the given geographic area during the same year}} \times 1000$$

Thus, for the example given above, we can determine the *still-birth rate* as

$$\frac{128 \times 1000}{4885} = 26.2$$

Perinatal mortality rate. Many late foetal deaths and early neonatal deaths may be attributed to similar underlying conditions and it has been suggested that a single rate called the perinatal mortality rate be calculated by combining the deaths in both categories. The idea is to include the period of perinatal existence after the foetus is considered viable (i.e., foetal deaths occurring after 20 or 28 completed weeks of gestation), the duration of labour, and the early part of the postnatal life. Some prefer to restrict the postnatal period to under 7 days, others to the neonatal period, i.e., under 28 days. No agreement has yet been reached on these points or on the denominator to be used, i.e., whether it should be only the live births or the sum of live births and foetal deaths. In setting out the rate of perinatal mortality, the terms should, therefore, be clearly defined. At its Sixth National Meeting in 1956,

the U.S. Public Health Conference on Records and Statistics proposed this definition of the perinatal mortality rate for general use in analysis of vital records:

$$\text{Perinatal mortality rate} = \frac{\text{Foetal deaths (20 weeks or more) plus post-natal deaths under 1 week}}{\text{Foetal deaths (20 weeks or more) plus live births}} \times 1000$$

More conventionally, the following formula is used:

$$\text{Perinatal mortality rate} = \frac{\text{Late foetal deaths (28 weeks or more) plus deaths under one week}}{\text{Live births plus late foetal deaths}} \times 1000$$

Maternal mortality rate. The risk of dying from causes associated with childbirth is measured by the maternal mortality rate. For this purpose the deaths used in the numerator are those arising from puerperal causes. The numbers exposed to the risk of dying from puerperal causes are women who have been pregnant during the period. Their number being unknown, the number of live births is used as the conventional base for computing comparable maternal mortality rates. The formula is:

$$\text{Annual maternal mortality rate} = \frac{\text{Number of deaths from puerperal causes which occurred among the female population of a given geographic area during a given year}}{\text{Number of live births which occurred among the population of the given geographic area during the same year}} \times 1000$$

The classification and coding of deaths as puerperal deaths vary from one country to another or even within the same country, so we must be cautious in comparing maternal mortality rates for different places.

17.6 FERTILITY RATES

17.6.1 Crude Birth Rate

As an index of the relative speed at which additions are being made to the population through childbirth, use is made of the crude birth rate. The formula is:

$$\text{Annual crude birth rate} = \frac{\text{Number of live births which occurred among the population of a given geographic area during a given year}}{\text{Mid-year total population of the given geographic area during the same year}} \times 1000$$

Example. During 1990, there were a total of 4757 live births in a rural area and 3485 in an urban settlement. The estimated mid-year population in these areas were 136,303 and 113,518 respectively.

The *crude birth rates* for the two areas will be:

$$\text{Rural} = \frac{4757 \times 1000}{136,303} = 34.9$$

$$\text{Urban} = \frac{3485 \times 1000}{113,518} = 30.7$$

This rate is a measure influenced by several factors affecting natality in the population, viz., the age and sex composition, marriage rate, economic level, the true fertility, migration, etc. In spite of this limitation, the rate is widely used because of the general availability of the data.

17.6.2 General and Specific Fertility Rates

(i) General fertility rate (GFR). The number of births in a population is largely dependent upon the proportion of women of child-bearing age, especially the married women. A more appropriate measure of the fertility of a population is, therefore, obtained if we relate live births to the total number of women of reproductive ages, viz., 15–44 years. The rate is known as the *general fertility rate* (*GFR*) or age-limited live birth rate, calculated as follows:

$$\text{GFR} = \frac{\text{Number of live births in one year}}{\text{Number of women in age group 15–44 years}} \times 1000$$

The numerator consists of all live births, whether legitimate or illegitimate, and the denominator comprises all women of ages 15–44 years, whether married or unmarried. Since women of child-bearing ages of 15–44 constitute approximately a quarter of the total population, the GFR is approximately 4 times the crude birth rate.

(ii) General marital fertility rate (GMFR). If only the married women are considered, this rate will be known as the *General marital fertility rate* (*GMFR*). Thus,

$$\text{GMFR} = \frac{\text{Number of live births in one year}}{\text{Number of married women in age group 15–44 years}} \times 1000$$

In some countries, the age-group extends from 15–49 years or 10–49 years.

(iii) Age-specific fertility rates (ASFR). These take into account the age composition of population and, therefore, provide better basis for comparisons of natality. They are useful in analysing the historical trend of the live birth rate for various ages. The formula for the age-specific birth rate is:

$$\text{Annual live birth rate specific for age of mother (or father)} = \frac{\text{Number of live births which occurred to mothers (or fathers) of a specified age group of the population of a given geographic area during a given year}}{\text{Mid-year female (or male) population of the specified age group in the given geographic area during the same year}} \times 1000$$

(iv) Total fertility rate (TFR). This is the average number of children that would be born alive to a woman during her lifetime, who would be subjected to age-specific fertility rates of a given year. Thus,

TFR = (Sum of all age-specific fertility rates) × (Width of each age-group) + 1000.

This is one of the most important measures of fertility.

(v) Gross reproduction rate (GRR). The average number of female live births that would be born to a woman during her lifetime who would be subjected to age-specific fertility rates of a given year.

(vi) Net reproduction rate (NRR). The average number of female live births that would be born to a woman who would be subjected to age-specific fertility and mortality rates of a given year.

Some vital statistical data are now presented in Tables 17.2 to 17.6.

Table 17.2 Birth and Death Rates of India and Major States—2007

India/States	Birth rate			Death rate		
	Rural	Urban	Combined	Rural	Urban	Combined
INDIA	24.7	18.6	23.1	8.0	6.0	7.4
Andhra Pradesh	19.5	16.7	18.7	8.0	5.7	7.4
Assam	25.7	15.5	24.3	9.1	5.7	8.6
Bihar	30.2	22.9	29.4	7.6	6.2	7.5
Gujarat	24.5	20.7	23.0	8.1	5.8	7.2
Haryana	24.5	20.8	23.4	7.0	5.7	6.6
Himachal Pradesh	17.9	12.6	17.4	7.3	5.0	7.1
Karnataka	21.2	17.5	19.9	8.3	5.4	6.8
Kerala	14.8	14.5	14.7	6.9	6.4	6.8
Madhya Pradesh	30.5	21.5	28.5	9.4	6.2	8.7
Maharashtra	18.7	17.3	18.1	7.3	5.7	6.6
Orissa	22.4	16.1	21.5	9.5	7.0	9.2
Punjab	18.3	16.4	17.6	7.7	5.9	7.0
Rajasthan	29.2	23.7	27.9	7.0	6.0	6.8
Tamil Nadu	16.0	15.5	15.8	8.0	6.3	7.2
Uttar Pradesh	30.5	25.5	29.5	9.0	6.5	8.5
West Bengal	19.8	12.7	17.9	6.3	6.4	6.3

Source: Health Monitor (2008), Foundation for Research in Health Systems, Ahmedabad.

Table 17.3 Infant Mortality Rates of India and Major States—2007

India/States	Infant mortality rate		
	Rural	Urban	Combined
INDIA	61	37	55
Andhra Pradesh	60	37	54
Assam	68	41	66
Bihar	59	44	58
Gujarat	60	36	52
Haryana	60	44	55
Himachal Pradesh	49	25	47
Karnataka	52	35	47
Kerala	14	10	13
Madhya Pradesh	77	50	72
Maharashtra	41	24	34
Orissa	73	52	71
Punjab	47	35	43
Rajasthan	72	40	65
Tamil Nadu	40	32	39
Uttar Pradesh	72	51	69
West Bengal	39	29	37

Source: Health Monitor (2008), Foundation for Research in Health Systems, Ahmedabad.

Table 17.4 Other Mortality Indicators of India—2007

Rate	Rural	Urban	Combined
Neonatal mortality rate	40	22	36
Perinatal mortality rate	41	24	37
Still-birth rate	9	8	9
Age specific death rate (selected age groups)			
0–4	18.0	9.6	16.0
5–9	1.5	0.7	1.3
60–64	23.6	20.0	22.6
85+	194.4	198.7	195.5

Source: Health Monitor (2008), Foundation for Research in Health Systems, Ahmedabad.

Table 17.5 Fertility Indicators of India—2007

Indicator	Rural	Urban	Combined
Total fertility rate	3.0	2.0	2.7
Total marital fertility rate	4.5	3.9	4.4
Gross reproduction rate	1.4	0.9	1.3
Female age at effective marriage	20.0	22.2	20.6

Source: Health Monitor (2008), Foundation for Research in Health Systems, Ahmedabad.

Table 17.6 Crude Birth Rate (CBR), Crude Death Rate (CDR), and Infant Mortality Rate (IMR) for some countries—2008

Country	Rates per 1000		
	CBR	CDR	IMR
Egypt	27	6	33
Nigeria	43	18	100
South Africa	23	15	45
India	24	8	57
Iran	20	5	32
Pakistan	31	8	75
Sri Lanka	19	7	15
Bangladesh	24	7	52
Nepal	29	9	48
Myanmar	19	10	70
Malaysia	21	5	9
Philippines	26	5	25
Singapore	11	5	2.4
Thailand	13	8	16
China	12	7	23
Japan	9	9	2.8
Canada	11	7	5.4
United States	14	8	6.6
Brazil	20	6	24
Argentina	19	8	13.3
Sweden	12	10	2.5
United Kingdom	13	9	4.9
Switzerland	10	8	4
Italy	9	10	4.2
Russia	12	15	9
Australia	14	7	4.7
New Zealand	15	7	5

Source: World Population Data Sheet (2008).

QUESTIONS

1. What is meant by 'vital statistics'?
2. Explain in detail the importance of infant mortality rate.
3. Mention the uses of age-specific fertility rates.
4. The number of vital events occurred during a year in a place are:

   Live births: 5400
   Deaths aged below 7 days: 50
   Deaths at age 7–27 days: 150
   Deaths at age 28 days–364 days: 300

   Find neonatal, postneonatal and infant mortality rates.

CHAPTER

18 Health Statistics

18.1 INTRODUCTION

No definition of the term 'health statistics' has yet been agreed upon. In general, the term should include all statistical information required for the administration of a health agency, and so would comprise not only vital statistics but also a good deal of other numerical information. More usually, the term is limited to health data than the term 'vital' statistics and includes data on all manner of sickness (morbidity, illness, ill health).

Since health is a function of social, economic, political and environmental factors, any data pertaining to these in relation to health and disease are useful. In addition, the statistics on health services such as health, manpower and recipients of health care also constitute health statistics.

18.2 UTILISATION OF THE BASIC DATA

Briefly stated, the *usefulness* of health statistics lies, in general, in their giving information on the following:

1. How many people suffer from particular diseases, how often and for how long?

2. What demands do these diseases place on the medical and public health resources; and what financial loss do they cause?

3. How fatal are the different diseases?

4. To what extent are people prevented by these diseases from carrying on their normal activities?

5. To what extent are diseases concentrated in particular groups of the population, e.g., according to age, sex, ethnic group, occupation or place of residence?

6. How far do the above factors vary from time to time (variation according to season or from year to year)?

7. What is the effect of medical care and health services on the control of disease incidence?

From the point of view of usefulness of information the available statistics fall into three categories, relating to

1. Health status of persons and populations in a given area, providing indices of vitality and health,

2. Physical, environmental and other conditions and factors having more or less direct bearing on the health status of the population—indices of social and environmental factors, and

3. Health services and activities directed at the improvement of health conditions—indices of health activity and facilities.

Some of these data are collected in accordance with established legal provisions; others in response to specific needs of the health agency and its stage of development. Some are required because they are largely of local concern, others because the health agencies at higher levels—state, national and international—are also interested. Some are needed to promote health plans, some to ensure their effective execution, others for the evaluation of health work.

18.3 SOURCES OF HEALTH STATISTICS

The various sources of information can be classified as follows:

1. Systems organised on a national scale to obtain information continuously from each household or institution, that is, registration and notification of disease. In recent years, national estimates of total morbidity have been obtained through nation-wide sample surveys. Similar surveys have given estimates of the prevalence of certain diseases of long duration, e.g., tuberculosis, cancer and leprosy.

2. Medical and health institutions providing service to the community, such as hospitals, clinics, and dispensaries, maternal and child health centres, local health units, diagnostic laboratories, mass immunisation or treatment programmes, social security schemes, medical care programmes, and the like. These provide statistics of admissions to hospitals and other medical institutions or services, and diagnostic returns of diseases for which treatment was provided, their duration and outcome. The data generally relate to 'selected' population groups.

3. Surveys or investigations conducted in response to the need for more detailed information on a narrower geographic basis such as nutrition surveys, epidemiological investigations in various diseases, field trials to test the therapeutic or prophylactic values of drugs, vaccines, or procedures.

4. Miscellaneous, e.g., physicians' case records, police reports on accidents, meteorological data and information on social, economic or occupational factors affecting health.

From these various sources, the following are of importance:

1. Notifiable disease records;

2. Hospital records of in-patients together with records of attendance of out-patients in hospitals, clinics and dispensaries;

3. Cases seen by the staff of infant welfare centres, clinics, school medical services, by factory medical staff and other medical staff responsible for special groups of population;

4. Medical practitioners' records based on domiciliary visits and patients attending their surgery;

5. Records collected during domiciliary visits paid by other health and welfare staff;

6. Special sickness surveys covering the whole population or samples of it;

7. Statistics on illness collected in connection with life or sickness insurance;

8. Medical care statistics, i.e., statistics collected under community medical care programmes;

9. National or local registers on cancer and tuberculosis;

10. Statistics of causes of death in relation to knowledge of duration of illness if recorded;

11. Records from industrial sickness-benefit associations;

12. Records from recruitment to the armed forces or jail records;

13. Mass diagnostic and screening surveys, e.g., tuberculosis, venereal diseases;

14. Absenteeism and sickness records in educational institutions, civil service examinations, and industrial concerns.

These numerous scattered fragments of statistical information can yield valuable data ranging from detailed information in some cases to moderately detailed knowledge of fairly broad segments of the population.

18.4 PROBLEMS IN THE COLLECTION OF SICKNESS DATA

In every community there are three classes of persons, namely, those whose level of health is so low that they are victims of disease; others who, while they manifest no definite signs of sickness are yet so devitalised that the possible range of their physical and mental achievements is considerably restricted, and those who are blessed with an abundance of health and vigour. An assessment of the state of public health in a country should, therefore, be based on information relating to all these three classes of people.

There are, however, certain aspects of morbidity measurement which pose serious obstacles to their wide applicability. For example, comparing with mortality measurement:

1. Morbidity can recur several times as the same condition or as a different one, while mortality is unique and a terminal event.

2. Morbidity has a duration varying from a few seconds to a number of years, while mortality occurs at an instant.

3. Morbidity may not be well defined and can be interpreted according to varying criteria, while mortality can be accurately defined.

Thus, in measuring morbidity, the following problems must be borne in mind:

Problem of definition. We need both a definition of what we are attempting to measure and record, and a scale of measurement. The dividing line between illness and health is often indefinite. Individuals respond differently to infection by certain

organisms and in some cases there may be no apparent departure from health, while in others infection may be accompanied by acute symptoms.

Problem of diagnosis. The accuracy of disease diagnosis depends upon the doctor's knowledge and the diagnostic facilities available. Both these factors may change with the advance of medical knowledge and the development of new techniques. Moreover, the diagnosis of the same doctor in regard to a patient may change during the course of a single illness as new symptoms appear.

Criteria of ill health. There are three principal criteria for establishing ill health:

1. *The opinion of the individual affected.* If a person believes himself to be ill, he could be considered so even though a physician may not be able to identify a disease.

2. *Clinical examination by a physician.*

3. *Diagnostic tests.* Contrary to general belief, diagnostic tests do not always yield diagnoses. A diagnostic test may be qualitative, such as an X-ray film, or it may give a reading on a quantitative scale. Criteria for relating specified scale values to various degrees of ill health must be established before the results of such tests can be interpreted in terms of illness.

Classification into healthy and sick. The following five groups based on the work of Percy Stocks and Harold F. Dorn, may be suggested:

GROUP I Free from any physical or mental defect.

GROUP II Having a congenital defect producing no appreciable disability and not requiring treatment but without sickness as defined in III, IV or V.

GROUP III Carrying scars or deformities left by past illness or injury which do not now cause appreciable disability nor require treatment but without sickness as defined in IV or V.

GROUP IV Affected by latent or incipient disease not yet causing the subject to suffer either symptoms of illness or restriction of activities imposed by its discovery, but without sickness as defined in V.

GROUP V Affected by a disease, ailment or result of injury or by a physical or mental defect, not covered by II, III or IV, including any departure from normal health of which the subject is aware or has been made aware and anything requiring treatment and embracing convalescence after illness or injury.

Groups I–IV may be designated as healthy and Group V as sick for the purpose of statistics. If it is desired to distinguish groups I–IV by descriptive terms, these might be healthy without defect, healthy with abnormality, healthy with scars and healthy with latent disorder.

To give an example, a man found to have a supernumerary rib causing no disability and without any other known defect would be in Group II, but if X-rayed and found to have a healed tuberculosis lesion, would pass to group III. If later he was found to have a raised blood pressure causing no symptoms and requiring no

immediate treatment or restriction of activity, he would pass to group IV; and the moment any symptoms of this appeared he would begin to be 'sick' (Group V).

18.5 MEASUREMENT OF SICKNESS

18.5.1 Recommendations of WHO

The Expert Committee on Health Statistics of the WHO stated as follows:

The Committee agreed that the statistical units involved in the compilation of morbidity statistics will generally be either 'persons', 'illnesses' or 'spells (or 'episodes' or 'periods') of illness'. For instance, one patient may, within a given period of observation, e.g., a year, suffer from illness *A* twice and from illness *B* three times. If the spells of illness *A* and *B* occur at different times this patient will have contributed one unit to statistics of 'persons', two units to statistics of 'illness' (i.e., 'diagnoses' or 'causes of illness') and five units to 'spells'. The Committee regarded it as essential that in the presentation of morbidity statistics it should be made clear as to which of these three concepts of morbidity the statistics actually relate. Furthermore, the items of information upon which morbidity statistics are compiled, in whichever of the above three classes they fall, will relate either to sickness beginning within the period of observation, to sickness ending during the period, to sickness current at any time during the period or to sickness current at some particular point of the time during the period. Alternatively, the statistics may relate to the duration of sickness current or ending during the period; these are widely used in the social security field. Great confusion can result from attempts to compare morbidity statistics based on sickness beginning, ending, or current within a given period of observation with morbidity statistics in which the aspect of sickness measured is its duration.

In its next report the Committee has recommended that the term 'incidence' be used to describe the measurement of frequency of illness commencing during a defined period and that this term be used with no other connotation. In utilising the term 'incidence' it should be made clear whether the term is used in respect of illness or in respect of persons ill. The term 'period prevalence' (or more specifically, annual, monthly, weekly prevalence) has been recommended for use to describe the measurement of frequency of illnesses in existence at any time during a defined period (i.e., a year, a month, a week etc.) and this term is not to be used with any other connotation. The term 'point prevalence' (or, more specifically, the 'prevalence' at a stated point in time) has been recommended for use to describe the measurement of frequency of illness in existence at a particular point in time. This term also is not to be used with any other connotation.

18.5.2 Some Morbidity Indices

(a) *Inception rate or incidence rate (spells).* The number of spells of sickness which start during a defined period divided by the average number of persons exposed to risk during that period. A 'spell of sickness' is a period during which a person is sick on one day (or shift) or on each day of a consecutive series of days (or shifts).

(b) *Inception rate or incidence rate (persons).* The number of persons who start at least one spell of sickness during a defined period divided by the average number of persons exposed to risk during that period.

(c) *Period prevalence rate (spells).* The number of spells of sickness which are current at sometime during a defined period divided by the average number of persons exposed to risk during that period.

(d) *Point prevalence rate (spells).* The number of spells of sickness which are current at a given time divided by the number of persons exposed to risk at that time.

(e) *Period prevalence rate (persons).* The number of persons who are sick sometime during a defined period divided by the average number of persons exposed to risk during that period.

(f) *Point prevalence rate (persons).* The number of persons who are sick at a given time divided by the number of persons exposed to risk at that time.

(g) *Average duration per spell.* The total of the entire durations of all spells of sickness ending during a defined period divided by the number of spells ending during that period.

(h) *Average duration per sick person.* The total of the entire durations of all spells of sickness ending during a defined period divided by the number of persons who experienced at least one spell ending during that period.

(i) *Average duration per person.* The total duration within a defined period of all spells of sickness that occurred wholly or partly within that period divided by the average number of persons exposed to risk during the period.

(j) *Proportion of time lost.* The total duration within a defined period of all spells of sickness that occurred wholly or partly within that period divided by the total duration of exposure to risk in the period.

(k) *Fatality ratio.* The number of deaths from a disease recorded during a defined period divided by the number of new cases of that disease recorded during the same period (or related period).

(l) *Notification rate.* This is the number of notifications during a defined period divided by the average number of persons exposed to risk during the period.

(m) *Carrier rate.* The proportion (per 100, per 1000, etc.) of carriers found in apparently normal individuals.

(n) *Secondary attack rate.* The number of new cases in the household is

expressed as a percentage of the total number in the households, *excluding* the index cases.

This is a measure of the risk of persons in a household exposed to a disease in another member (index case) in the same household, thus showing the increased risk of contracting the disease due to familial exposure.

18.5.3 Some Examples

(a) The estimated mid-year population for place A in 1990 was 761,335, of which 76,100 were in the age-group 1–4 years. During the year, there were 700 attacks of measles in the age group 1–4 years. The total new cases of tuberculosis were 912; the total cases of tuberculosis present in the year were 23,000. From this we calculate:

$$\text{Incidence rate for tuberculosis} = \frac{912}{761{,}335} \times 1000 = 1.2$$

$$\text{Age-specific incidence rate for measles in the age group 1–4 years} = \frac{700}{76{,}100} \times 1000 = 9.2$$

$$\text{Prevalence rate for tuberculosis} = \frac{23{,}000}{761{,}335} \times 1000 = 30.2$$

(b) In a certain school, the total population at risk was 530. After an epidemic, there were 3 paralytic polio cases and 130 with inapparent infections (those with complement fixating antibodies in the blood). From this, we can calculate:

$$\text{Infection rate} = \frac{3+130}{530} \times 100 = \frac{133}{530} \times 100 = 25.1$$

18.6 HOSPITAL STATISTICS

Even if information through household interview survey methods is not obtainable, statistics from hospitals and dispensaries can provide valuable information provided they are interpreted with due regard to the selective nature of these data. In any case, these statistics do show the kind of diseases or injuries for which medical facilities are being sought, as well as seasonal and annual variation in the diseases as reflected by changes in admissions or attendances.

18.6.1 Terminology

Serious consideration has been given in the United States, the United Kingdom, France and other countries to defining various terms which frequently appear in hospital statistics. The subject has also been discussed by the WHO Expert Committee on Health Statistics. In addition, the United Nations Trusteeship Council has attempted to standardise some terms for use in non-self governing territories. Although no single definition of many of the terms appears to have been agreed upon, those so far used or formulated are in close accord. The difference in the quality or progress of medical services is so marked that sometimes even the best hospital in one area may not compare favourably with one regarded as poor in an advanced

country, and so mere adherence to terminology cannot provide strictly comparable data. Many figures published by UN and WHO contain clear reservations to this effect. The important point is that the terminology used in publishing statistics should mean the same thing to the reader as it does to the writer of the report.

18.6.2 Some Indices

Several indices to describe hospital services are used. Some in common use are defined below. Many rates calculated from hospital statistics also are based on the concept of incidence or prevalence. For instance:

'Admission rate' corresponds with incidence rate. If possible, the admission rate should be subdivided into first admissions and readmissions. Figures for discharges rather than admissions are often used as a basis for tabulation, the admission rate and the discharge rate being largely interchangeable.

The 'average daily bed occupancy' or 'average daily census' is the average number of patients in a hospital and corresponds with the point prevalence rate. It is calculated by dividing the number of patient-days during a period by the number of days in the period.

The 'average duration of stay' and the frequency distribution of durations of stay are usually tabulated in relation to individual periods of hospitalisation. They could equally well be tabulated in relation to the total hospitalisation time of persons during a defined period.

Average duration of stay. Total duration of stay of discharged patients divided by the number discharged during a period.

Percentage of occupancy. Ratio of actual patient-days to the maximum patient-days (based on bed-complement) during any given period of time.

Fatality ratio. The number of deaths from a disease occurring in a hospital during a defined period, divided by the number of new cases of that disease admitted to that hospital during the same period.

Hospital death rate. The total number of hospital deaths in a given period divided by the number of discharges during the same period.

Autopsy rate. The ratio during any given period of all hospital autopsies to all hospital deaths.

18.7 INTERNATIONAL CLASSIFICATION OF DISEASES

18.7.1 General Principles

Classification is fundamental to the quantitative study of any phenomenon. It is recognised as the basis of all scientific generalisation and is, therefore, an essential element in statistical methodology. Uniform definitions and uniform systems of classification are prerequisites in the advancement of scientific knowledge. In the

study of illness and death, therefore, a standard classification of disease and injury for statistical purposes is essential.

There are many approaches to the classification of disease. The anatomist, for example, may desire a classification based on the part of the body affected. The pathologist, on the other hand, is primarily interested in the nature of the disease process. The clinician must consider disease from these two angles, but needs further knowledge of etiology. In other words, there are many axes of classification and the particular axis selected will be determined by the interests of the investigator. A statistical classification of disease and injury will depend, therefore, upon the use to be made of the statistics to be compiled.

The purpose of a statistical classification is often confused with that of nomenclature. Basically, a medical nomenclature is a list or catalogue of approved terms for describing and recording clinical and pathological observations. To serve its full function, it should be extensive, so that any pathological condition can be accurately recorded. As medical science advances, a nomenclature must expand to include new terms necessary to record new observations. Any morbid condition that can be specifically described will need a specific designation in a nomenclature.

This complete specificity of a nomenclature prevents it from serving satisfactorily as a statistical classification. When one speaks of statistics, it is at once inferred that the interest is in a group of cases and not in individual occurrences. The purpose of a statistical compilation of disease data is primarily to furnish quantitative data that will answer questions about groups of cases.

International conferences have been held every 10 years since the beginning of this century to arrive at satisfactory systems of classifications of diseases. The most recent meeting was the International Conference for the Tenth Revision of the International Classification of Diseases convened by WHO in Geneva in 1990. Its report was adopted by the forty-third World Health Assembly. The title is now amended as the "Tenth Revision of the International Statistical Classification of Diseases and related Health Problems" to make clearer the content and purpose and to reflect the progressive extension of the scope of classification beyond diseases and injuries. However, the familiar abbreviation 'ICD' is retained. The next revision ICD-11 is planning for 2015.

Experience has shown that it is impossible to maintain a consistent axis of classification for all diseases. Compromises have been made in the current edition of ICD between aetiology, pathologic conditions, anatomic site, age and circumstances of onset of the disease.

Upto the Ninth Revision, the ICD was published in two Volumes, 1. Tabular list and 2. Alphabetical index. In the Tenth Revision there are three volumes:

Volume 1: Tabular list. This contains the report of the International Conference for the Tenth Revision, the classification itself at the three- and four-character levels, the classification of the morphology of neoplasms, special tabulation lists for mortality and morbidity, definitions, and the nomenclature regulations.

Volume 2: Instruction manual. This brings together the notes on certification and classification formerly included in Volume 1 with a good deal of new background and instructional matter and guidance on the use of Volume 1, on tabulations, and on

planning for the use of ICD, which was seen as lacking in earlier revisions. It also includes the historical material formerly presented in the introduction to Volume 1.

Volume 3: Alphabetical index. This presents the index itself with an introduction and expanded instructions on its use.

Volume 3 is used as the primary coding tool, but one must always refer back to Volume 1 to ensure that the code selected in Volume 3 fits the classification assignment of a particular case. It is necessary to read and be guided by the inclusion and exclusion notes that appear at the beginning of each chapter in Volume 1 as well as below the three- and four-digit code titles. Volume 3 also gives specific directions to refer elsewhere for further information.

18.7.2 Numbering System of the Classification

Three-digit categories. The classification, arranged in 21 main sections in the Tenth Revision deals first with diseases caused by well-defined infective agents; these are followed by categories for neoplasms, blood and immunity related disorders, endocrine, metabolic, and nutritional diseases. Most of the remaining diseases are arranged according to their principal anatomical site, with special sections for mental diseases, complications of pregnancy and childbirth, certain diseases peculiar to the perinatal period, and ill-defined conditions including symptoms. One section, S00 to T98 in the Tenth Revision, (800–999 in the Ninth), provides a classification of injuries, according to the nature of injury, such as puncture, open wound, or burn. These numbers also have an alternate classification based on external cause of injury and they are distinguished by the prefix 'E' in the classification in the Ninth and allotted separate numbers V01 to Y98 in the Tenth.

In many instances, the first two digits of the three-digit number designate important or summary groups that are significant. The third and fourth digits divide each group into categories and subcategories which represent specific disease entities or a classification of the disease or condition according to some significant axis such as anatomical site. Further, the detailed or three-digit categories have not been numbered consecutively, but numbers have been omitted in order that the summary character of the first two digits could be preserved wherever they are meaningful. No additional three-digit categories should be introduced in the classification except when the list is revised by international agreement.

The main sections in the Tenth and Ninth revisions are as follows:

		X (1985)	IX (1975)
I	Certain infectious and parasitic diseases	A00–B99	001–139
II	Neoplasms	C00–D48	140–239
III	Diseases of blood and blood-forming organs and certain disorders involving the immune mechanism	D50–D89	279–289
IV	Endocrine, nutritional and metabolic diseases	E00–E90	240–278

V	Mental and behavioural disorders	F00–F99	290–319
VI	Diseases of the nervous system	G00–G99	320–359
VII	Diseases of the eye and adnexa	H00–H59	360–379
VIII	Diseases of the ear and mastoid process	H60–H95	380–389
IX	Diseases of the circulatory system	I00–I99	390–459
X	Diseases of the respiratory system	J00–J99	460–519
XI	Diseases of the digestive system	K00–K93	520–579
XII	Diseases of the skin and subcutaneous tissue	L00–L99	680–709
XIII	Diseases of the musculoskeletal system and connective tissue	M00–M99	710–739
XIV	Diseases of the genitourinary system	N00–N99	580–629
XV	Pregnancy, childbirth and the puerperium	O00–O99	630–679
XVI	Certain conditions originating in the perinatal period	P00–P96	760–799
XVII	Congenital malformations, deformations and chromosomal abnormalities	Q00–Q99	740–759
XVIII	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	R00–R99	780–799
XIX	Injury, poisoning and certain other consequences of external causes	S00–T98	800–999
XX	External causes of morbidity and mortality	V01–Y98	E800–E999
XXI	Factors influencing health status and contact with health services	Z00–Z99	V01–V82

The last category replaces the earlier supplementary classification of factors influencing health status and contact with health services known as the 'V' code in the Ninth Revision and 'Y' code in the Eighth Revision (1965). This can arise in two main ways:

1. When a person who is not currently sick encounters the health services for some specific purpose, such as to act as a donor of an organ or tissue, to receive prophylactic vaccination or to discuss a problem which is in itself not a disease or injury. This will be a fairly rare occurrence among hospital in-patients but will be relatively more common among hospital outpatients and patients of family practitioners, health clinics, etc.

2. When some circumstance or problem is present which influences a person's health status but is not in itself a current illness or injury. Such factors may be elicited during population surveys, when the person may or may not be currently sick, or be recorded as an additional factor to be borne in mind when the person is receiving care for some illness or injury. In the latter circumstances, the Z code (Z00–Z99) should be used only as a supplementary code and should not be the one selected for use in primary, single-cause tabulations. Examples of these circumstances

are a personal history of certain disease, or a person with an artificial heart valve in situ. This code is broadly subdivided as follows:

Persons encountering health services for examination and investigation	Z00–Z13
Persons with potential health hazards related to communicable diseases	Z20–Z29
Persons encountering health services in circumstances related to reproduction	Z30–Z39
Persons encountering health services for specific procedures and health care	Z40–Z54
Persons with potential health hazards related to socioeconomic and psychosocial circumstances	Z55–Z65
Persons encountering health services in other circumstances	Z70–Z76
Persons with potential health hazards related to family and personal history and certain conditions influencing health status	Z80–Z89

18.7.3 Special Lists for Tabulation

In order to assist in uniformity of returns on mortality and morbidity data, the following lists have been prepared and included in the Eighth Revision:

List A: List of 150 causes for tabulation of morbidity and mortality.

List B: List of 50 causes for tabulation of mortality.

List C: List of 70 causes for tabulation of morbidity.

List D: List of 300 causes for tabulation of hospital morbidity.

List P: List of 100 causes for tabulation of perinatal morbidity and mortality.

In the Ninth Revision (1975), these lists have been replaced by a more flexible system allowing users to construct their own tabulation lists from the rubrics of a basic list. The basic tabulation of digit rubrics contains 307 two- and three-digit rubrics and is intended to replace the previous lists A and D. For the national display of mortality and morbidity data, countries are free to use any lists constructed from the items of the basic list, but to ensure a minimum of international comparability. Any tabulation lists used for these purposes must contain the 50 rubrics of the attached Mortality and Morbidity Lists, which replace the previous Lists B and C.

In the Tenth Revision, five special tabulation lists—four for mortality and one for morbidity were suggested as follows:

List 1: General mortality–condensed list (103 causes)

List 2: General mortality–selected list (80 causes)

List 3: Infant and child mortality–condensed list (67 causes)

List 4: Infant and child mortality–selected list (51 causes)

List 5: Morbidity list (298 causes).

18.7.4 Problems in Application of ICD

The ICD represents only a first step in the actual compilation of statistics of causes of illness and death. Before applying the classification to actual data, it is necessary to decide what purpose is to be served by the statistics. The problem is complex because frequently in both sickness and death more than one morbid process is involved. As a consequence, in many countries more than one-half of the medical certificates of death and an even larger proportion of hospital and other records of illness and disability contain mention of more than one cause. The first problem which arises, therefore, is whether the statistics are to deal with the count of persons who are sick or have died, or of conditions that produced the illness or death. Both kinds of statistics are important attributes of morbidity and mortality and warrant study.

The early records of death contained only a single cause and so, a few simple rules sufficed to secure uniform selection of the cause of death. The usual mortality table, therefore, is based on individuals who have died, assigning a single cause to each death. As an increasingly larger proportion of the certificates contained multiple causes, the problem of selection became more important in securing comparable statistics.

Morbidity is far less definite than mortality and represents a dynamic rather than a static event. The basic problem of what is to be counted is complex, and the application of the classification to the morbidity statistic cannot be laid down as precisely and relatively simply as in mortality. The application will vary, depending on the kind of morbidity experiences to be studied and on the purposes to be served by the statistics.

QUESTIONS

1. What constitutes 'health statistics'? Write a note on its uses.
2. Describe the problems in collection of morbidity data.
3. Briefly explain:
   (a) Hospital statistics
   (b) ICD.
4. The estimated mid-year population of children in a town during 2007 was 75,000. There were 9,500 total cases of polio, including 50 occurred during the year 2007. Find the incidence and prevalence of polio.

CHAPTER

19 Standardised Death Rates and Life Tables

19.1 INTRODUCTION

Several commonly used measures of mortality were discussed in Chapter 18. However, in certain situations especially for comparative studies, comprehensive methods of describing mortality are needed. Two such methods are 'Standardisation' and 'Life (mortality) Table Techniques'. These are explained in this chapter.

19.2 ADJUSTED OR STANDARDISED RATES

The crude death rates of two or more places cannot be compared when their populations have different age compositions. This point is illustrated in Table 19.1 by a hypothetical example of three areas, *A*, *B* and *C*, for which the specific death rates by age and the corresponding populations are given.

Table 19.1 Death Rates in Three Areas: *A*, *B*, and *C*

Age	Death rate per 1000 per year			Population (in thousands)			Deaths* at each age		
(years)	*A*	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0–14	15	16	17	4000	3500	3000	60	56	51
15–29	10	11	12	2000	3000	3500	20	33	42
30–44	14	15	16	1500	2000	2500	21	30	40
45 +	40	45	50	2500	1500	1000	100	68	50
All Ages	20.1	18.7	18.3	10000	10000	10000	201	187	183

*Number of deaths based on the population and death rate.

It is clear from columns (2) to (4) that at each age-group area *A* had the lowest death rate and area *C* had the highest. This led us to infer that *A* enjoyed better health status than *C*. The crude death rates shown at the bottom of columns (2) to (4) on the other hand, lead to contradictory conclusions, as the crude death rate of area *A* is the highest. This fallacious conclusion arises because of marked differences in the age composition, area *A* having the largest proportion of old people. To compare the overall death rate of different populations, we should ascertain how much of the difference in the rates is due to the unfavourable age distribution of the populations. One method that has been adopted to eliminate the effect of varying age

or sex composition is to compute, for each place, a rate which would have been the death rate had the age compositions been the same. This is known as the *direct method* of standardisation and is illustrated below by comparing the death rate for Vellore with that of India as a whole.

19.2.1 Direct Method

The specific death rates by age and sex for Vellore are given in Table 19.2.

Table 19.2 Specific Death Rates Per 1000 by Age and Sex for Vellore

Age (years)	Specific death rates per 1000 (Vellore)	
	Males	Females
Below 1 yr.	120.5	111.7
1–4	21.6	25.1
5–9	3.7	3.6
10–14	1.9	2.5
15–19	2.5	3.3
20–29	3.5	4.3
30–39	6.4	6.2
40–49	8.2	5.3
50–59	22.7	14.5
60 +	61.0	64.2
All ages	15.1	14.9
	15.0	

India's population distribution was calculated for each age and sex group in such a manner as to give a total of 1 million persons, preserving the same proportional age and sex distribution as in the total population. This may be called the *standard million* population for India and is shown in Table 19.3.

Table 19.3 Standard Million Population: India, 1971

Age (years)	Standard million population for India, 1971*	
	Males	Females
Below 1 yr.	14,217	13,822
1–4	60,609	58,926
5–9	79,895	74,741
10–14	65,865	58,268
15–19	44,282	39,690
20–29	70,873	76,250
30–39	62,473	62,328
40–49	49,270	44,393
50–59	33,076	28,889
60 +	32,096	30,037
All ages	512,656	487,344

*Age-sex distribution of India's population assuming the total population of one million.

Each specific death rate for Vellore was multiplied with the corresponding population in the standard million to obtain the expected number of deaths. The total deaths thus obtained are divided by 1 million and converted to rate per 1000. The single rate resulting from this adjustment is called the *standardised, adjusted* or *corrected* death rate. The calculations are shown in Table 19.4.

Table 19.4 Expected Deaths in Standard Population

Age (years)	Deaths that would have occurred in the standard population from specific death rates of Vellore	
	Males	Females
Below 1 yr.	1713	1544
1–4	1309	1479
5–9	296	269
10–14	125	146
15–19	111	131
20–29	248	328
30–39	400	386
40–49	404	235
50–59	750	419
60+	1958	1928
Total for all ages	7314	6865
Grand total	14,179	

The numbers in the last two columns are obtained as follows: the first number, viz. 1,713 is obtained by applying the rate 120.5 to the number of male children under 1 year of age, i.e. 14,217

$$\left(\frac{14,217 \times 120.5}{1000}\right) = 1,713\ldots$$

Adjusted or standardised death rate is equal to

$$\frac{14,179 \times 1,000}{1,000,000} = 14.2 \text{ per thousand}$$

The crude death rate of Vellore was 15.0 but after correction for age and sex composition, the death rate becomes 14.2. This would have been the death rate of Vellore if the area had the same age/sex composition as the Indian population of 1971. If we want to compare the death rate of Mumbai City with Vellore's death rate, we must carry out the same computation using Mumbai's age and sex specific death rates and applying to them the standard million population of India.

This method requires information on the specific death rates by age and sex. Sometimes this information is not available and even if it is, the specific death rates calculated for some age-groups are subject to large fluctuations due to small numbers. In such instances, the 'indirect method' of adjustment or standardisation is used.

19.2.2 Indirect Method

The indirect method as shown in Table 19.5 requires knowledge only of the age and sex compositions of the populations to be compared. The specific death rates of the standard population are applied to the respective age-groups of the populations of the areas to be compared. The deaths expected in each age-group are added and divided by the total population of the area to first obtain what is known as the *index death rate*, indicating what would have been the death rate if the area had experienced the mortality rates of the standard population.

Table 19.5 Calculations for Indirect Method of Standardisation

Age (years)	Population of Vellore		Standard specific death rates of India, 1971		Deaths that would have occurred in Vellore population if standard specific death rates of India were operative	
	Males	Females	Males	Females	Males	Females
Below 1 yr.	887	865	73.9	70.8	66	61
1–4	2628	2576	12.0	12.0	32	31
5–9	2966	3076	3.1	2.9	9	9
10–14	2760	2797	2.1	2.2	6	6
15–19	2049	1576	3.0	3.0	6	5
20–29	3319	3315	2.5	2.7	8	9
30–39	2643	2593	3.7	3.9	10	10
40–49	2185	2154	6.5	5.5	14	12
50–59	1537	1611	12.7	10.6	20	17
60+	1469	1413	39.7	34.4	58	49
All ages	22,443	21,976	8.5	7.9	229	209
All ages both sexes	44,419		8.1		438	

The ratio of the crude death rate of the standard population to the index death rate is known as the *standardising factor* because it gives the correction necessary for the crude death rate of the area owing to its difference in age composition from that of the standard population. Here is an example of standardisation by the indirect method:

Age and sex distribution of the population of Vellore and the specific death rates for India are given in Table 19.5. The crude death rates of Vellore and of India were 15.0 and 8.1 respectively.

The numbers in the last two columns are obtained by applying the specific death rates of India to the appropriate age-groups. For example, for the age-group 1–4 years the expected male deaths are:

$$\frac{2628 \times 12.0}{1000} = 32, \ldots$$

$$\text{Index death rate for Vellore} = \frac{438 \times 1000}{44419} = 9.9/1000$$

$$\text{Crude death rate for India} = 8.1/1000$$

$$\text{Standardising factor} = \frac{8.1}{9.9} = 0.82$$

Standardised death rate for Vellore = Crude death rate for Vellore multiplied by the standardising factor

$$= 15.0 \times 0.82 = 12.3/1000$$

From the above calculation, it follows that the death rate of 15.0 per 1000 per Vellore cannot be compared with the death rate of 8.1 per 1000 for India because of differences in age composition.

Both the direct and indirect methods of age adjustments have been criticised on the ground that the rates obtained are also dependent on the age and sex structure of the standard population used and that greater gains in mortality reduction obtained at younger ages are not adequately accounted for.

19.2.3 Comparative Mortality Figure and Standardised Mortality Ratio

An alternative, although equivalent, mortality measure is the comparative mortality figure (CMF) to denote the actual or observed total deaths in the standard population as D. Then, for Region 1, the CMF is $(D_1 \times 100)/D$ and for Region 2 it is $(D_2 \times 100)/D$. The CMF is thus the ratio of 'Expected' (i.e. the number which would occur if the standard population were subject to the mortality conditions of Regions 1 and 2 respectively) to 'Observed' deaths, multiplied by 100 to express it as a percentage. A CMF less than 100 indicates that mortality conditions are 'better' in a particular region compared with the standard population, while a CMF greater than 100 indicates they are 'worse'. Moreover, the CMFs for different regions can be compared.

The direct age-standardised death rate and the CMF are related. Defining the observed death rate for the standard population as $(D \times 1000)/P$, it is easily shown that

$$\text{CMF} = \frac{\text{Direct age-standardised death rate for the region}}{\text{Observed death rate for standard population}} \times 100$$

The reader may note that while the measures explained above have been standardised with respect to age, the numerical values of the measures will depend upon the standard population selected. In the example above, and in similar cases, the normal practice is to select as standard the national population, of which the populations of the different regions (or occupations, or socioeconomic groups) form a part. For comparisons over time, however, there is a choice of standard populations and the numerical values of the CMFs will vary, depending on which year is selected as the population standard. If the age structure of the standard population remains relatively stable over time, which year is chosen as standard will not make a great deal of difference, particularly since it is the relative rather than the absolute values of the CMF which are of interest. Nevertheless, this is a potential limitation of the CMF. Another feature of direct standardisation which may cause problems is where the population of interest (of a region, or occupational group, or whatever) is small,

or contains small numbers in particular age groups. In these circumstances, the chance occurrence of one or two individual deaths in a particular year may give rise to a much greater than normal age-specific mortality rate and hence distort the CMF. The coverage (i.e., an atypically low mortality rate in a particular age group) may also occur.

In the indirect method of standardisation, the age-specific mortality rates of a standard population are applied to the corresponding age-groups of the populations of interest (in the example, Regions 1 and 2), to yield the number of deaths 'expected' if each region had experienced the mortality conditions of the standard population. Denoting the number of expected deaths in each region as D_1^E and D_2^E respectively, and the observed or actual number of deaths in each region as D_1^O and D_2^O, the ratio of observed to expected deaths multiplied by 100 is called the standardised mortality ratio (SMR); i.e. for Region 1,

$$\text{SMR} = \frac{\text{Observed deaths in Region 1}}{\text{Expected deaths in Region 1}} \times 100 = \frac{D_1^O}{D_1^E} \times 100$$

and a similar equation holds for Region 2.

Like the CMF, the SMR is based on a comparison of observed and expected deaths and is similarly interpreted. The former compares observed deaths in the standard population with the number which would have occurred if that population had been subject to the age-specific mortality rates of a particular region (or occupation). The latter compares the observed deaths in a particular region (or occupation) with the number which would have occurred if that region had been subject to the age-specific mortality rates of the standard population. As with the method of direct standardisation, the calculations involved in indirect standardisation also permit the calculation of age-standardised death rates, but this is not elaborated here.

In many applications, the CMF and SMR will be equal or very close in value, but this is not always the case. Like the CMF, the SMR has a number of limitations, but it requires less information to calculate. In particular, it is not necessary to know the age distribution of deaths in the regions or occupations of interest.

The most common applications of SMRs are comparisons of mortality between different occupations or socio-economic groups, and the analysis of trends in mortality.

In concluding this discussion of standardisation, it is worth repeating that some caution should be exercised in the interpretation of standardised mortality ratios for particular occupations or social groups. Firstly, in some cases quite small populations are being considered, and SMRs can change significantly from one period to the next, just through chance. Secondly, SMRs do not necessarily indicate the mortality risk of a particular occupation. In a pure experiment, individuals would be randomly allocated to particular occupations and their mortality experience then examined, but of course, in reality the choice of occupation is not a random process. Professional athletes, for example, need talent and specific physical characteristics to make a success of such a career, while fishermen will usually be found amongst people who live on or near a sea-coast with an established fishing

industry. Occupational mortality may, therefore, be influenced not so much by the nature of the occupation as by other factors (education, area of work or residence, specific employment requirements and so on) which are associated with individuals who work in that occupation. In principle, these factors, like age and sex, should also be standardised before valid comparisons of occupational mortality risk can be made, and more complex standardisation techniques do take account of this.

19.3 LIFE TABLES

Another effective and at the same time comprehensive method of describing mortality in a population is by means of life tables. A life table is composed of several sets of values showing how a group of infants, all supposed to be born at the same time and experiencing specified mortality conditions, would gradually die out. The original purpose of a life table was mainly to provide a basis for actuarial computations, but its usefulness in many other fields of study was soon recognised. William Farr called the life table the 'Biometer' of the population. Today it is widely accepted as an important basic material in demographic and public health studies.

Life tables are usually worked out after each decennial census to represent mortality conditions either during the previous decennium or during shorter periods covering the date of the census. Separate tables for males and females are usually prepared as the mortality experienced by males is different from females. For detailed study it is not uncommon to construct tables for each geographical subdivision of a country or different population segments such as ethnic or occupational groups. A life table can also be constructed to show how a group of babies would die if, hypothetically, one or more causes of death were eliminated. In recent years life table techniques are being increasingly applied to follow-up studies of chronic diseases of hospital patients.

19.3.1 Construction of a Life Table

It is usual to construct a life table showing survival and death occurring in a generation of 100,000 babies. We ask what would happen if these 100,000 babies born at the same time were subjected to these mortality influences at various ages that are found to be influencing the population at a certain period of time. How many of these 100,000 would live to see their first birthday; how many their second birthday; how many their third birthday and so on, till the last one? On the basis of the mortality rate operating at the time under study, we can estimate what number would be alive at the first birthday by applying the mortality rate during the first year on the 100,000 children. By the application of the mortality rate of the second year on the number surviving at the end of the first year, we estimate the number who would survive to the end of the second year of life. Similarly for other ages. These numbers of survival to various ages form the basic data set out in a life table. From these we can also calculate the average time a person can expect to live after any particular age.

19.3.2 Structure of a Life Table

Two indices included in the life table and most frequently used in public health studies are the life table mortality rate, usually denoted by the symbol q_x, and the expectation of life, usually denoted by the symbol e_x. The mortality rate q_x describes the probability of death within a designated interval of time for persons of any exact age x. The expectation of life e_x measures the average number of years a person of a given age x is expected to live under the prevailing mortality conditions. It is an index showing in summary form the effect of mortality on survival. The other columns and symbols used in a life-table are as follows:

1. x = exact age in years

2. P_x = midyear population between ages x and $x + 1$

3. d_x = number of deaths between ages x and $x + 1$

4. m_x = age-specific death rate d_x/P_x

5. a_x = average fraction of the last year of life lived by those dying between ages x and $x + 1$; usually after 5, it is taken as 1/2

6. q_x = proportion of those alive at age x dying before age $x + 1$

$$q_x = \frac{m_x}{1 + (1 - a_x)\, m_x}$$

p_x = proportion of those alive at age x surviving to age $x + 1$

$$p_x = 1 - q_x$$

7. l_x = number of people surviving to age x out of life table cohort $l_0 = 100{,}000$

$$l_{x+1} = l_x\, p_x$$

8. d_x = number of people dying between ages x and $x + 1$ out of the life table cohort

$$d_x = l_x - l_{x+1}$$

9. L_x = number of years lived by the life table cohort between the ages x and $x + 1$

$$L_x = l_x - (1 - a_x)d_x$$

10. T_x = number of years lived by the life table cohort beyond age x

$$T_x = L_x + L_{x+1} + \ldots$$

11. e_x^0 = expectation of life at age x (or the average number of years of life remaining to persons aged x)

$$e_x^0 = T_x / l_x$$

Table 19.6 gives an extract from life table for females in India during 1961–70.

Table 19.6 Life Table (1961–70) for Females in India

Age	Survivors to exact age x	Deaths between ages x and $x+1$	Probability of death between ages x and $x+1$	Number of those living between ages x and $x+1$	Number of those living above age x	Expectation of life at age x
x	l_x	d_x	q_x	L_x	T_x	e_x^0
0	100,000	12,837	.12837	91,014	4,465,447	44.7
1	87,163	2969	.03406	85,382	4,374,437	50.2
2	84,194	1348	.01601	83,520	4,289,051	50.9
3	82,846	1198	.01446	82,367	4,205,531	50.8
4	81,648	1072	.01313	81,326	4,123,164	50.5
5	80,576	971	.01205	80,091	4,041,838	50.2
6	79,605	893	.01122	79,159	3,961,747	49.8
7	78,712	832	.01057	78,296	3,882,588	49.3
8	77,880	728	.00935	77,516	3,804,292	48.8
9	77,152	658	.00852	76,823	3,726,776	48.3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
75	16,024	1333	.08320	15,358	125,169	7.8
76	14,691	1267	.08623	14,058	109,811	7.5
77	13,424	1198	.08926	12,825	95,753	7.1
78	12,226	1128	.09227	11,662	82,928	6.8
79	11,098	1057	.09528	10,570	71,266	6.4
80	10,041	987	.09827	9548	60,696	6.0
81	9054	917	.10126	8596	51,148	5.6
82	8137	848	.10423	7713	42,552	5.2
83	7289	781	.10720	6899	34,839	4.8
84	6508	717	.11015	6150	27,940	4.3
85+	5791			21,790	21,790	3.8

The column of deaths in the life table may be split into component causes of death on the basis of ratios of specific causes to all causes as observed in vital statistics tabulations. Each age or narrow age group would be treated separately. With this arrangement, the decrements to the column of the living may now be described in terms of a number of causes instead of a single cause. This produces one form of a multiple decrement table.

In the more general type of multiple decrement table, the column of the living traced from birth is split into components on the basis of newly acquired characteristics or of changes in status. For example, in the cohort of white females traced from birth, it is possible to distinguish, after the childhood period, between those who enter marriage and those who remain single. In the same way, a distinction may be made between those who become permanently disabled and those who remain active. Thus, the column of the living in the life table may be split into components on the basis of rates, specific according to age, for the acquisition of new characteristics. The split may also be on the basis of more than one new characteristic.

Table 19.7 Example of a Double Decrement Table with Decrements by Death and by Disability

	Rate of mortality per 1000			Of 100,000 Born Alive						
				Number surviving to exact age x				Number dying between ages x and $x+1$		
Year of age x	Among active lives	Among disabled lives	Rate of disability within year of age per 1,000 active lives at exact age x	Total	As actives lives	As disabled lives	Number of active lives becoming disabled between ages x and $x+1$	Total	Among active lives	Among disabled lives
15	7.55	267	0.587	66,949	66,949	0	40	511	506	5
16	7.47	254	0.584	66,438	66,404	34	39	509	496	13
17	7.40	241	0.581	65,229	65,289	60	38	507	487	20
18	7.40	229	0.578	65,422	65,343	79	38	505	483	22
19	7.40	217	0.575	64,916	64,822	94	38	504	480	24
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The split of the column of the living implies that the column of deaths will also be divided on the basis of the same characteristics. In the case of the first example in the preceding paragraph, there will be life table deaths among the single and also among those who had married, each on the basis of their own mortality experience. Similarly, in the second example, life table deaths will be split on the basis of the mortality experienced by permanently disabled lives and by active lives. An example of a multiple decrement table is shown in Table 19.7.

19.3.3 Applications to Health Problems

The life table has been used extensively in public health problems and has also found an important place in clinical medicine.

(i) Comparisons of communities. The expectation of life at birth and at older ages is frequently used to compare health conditions of communities. Actually, there is no single index by which to measure the health conditions of a community and the use of the expectation of life or any other life table statistic for this purpose places undue emphasis upon mortality. In fact, the community with the higher expectation of life at birth may actually have the greater proportion of the chronically ill simply because such persons may be kept alive for a longer period. Also, comparisons of localities in terms of expectation of life may be influenced by selective factors in migration. Thus, some areas may attract impaired lives who migrate with the hope of improvement or recovery. In the same way, the experience of the havens for the retired will be influenced by their migrations. Even as an index of mortality, the life table has a fundamental weakness: the relative differences in expectation of life between two communities are by no means proportionate to

those in death rates. On the other hand, satisfactory comparisons may be made on the basis of life table death rates for broad age ranges.

(ii) Analysis by causes of death. The distribution of life table deaths according to specific causes, as in the multiple decrement table, is the first step in computing the chances of ultimate death from such causes. The desired result is obtained by summing the life table deaths from a specified cause after a stated age and dividing the total by the life table survivors to that age. Also of interest is the gain in expectation of the life that may be achieved by eliminating any cause of death. For this computation, the life table is based upon mortality rates excluding deaths from the specified cause, but with an allowance for a small increase in rates for the remaining causes. The allowance is necessitated by the fact that those who would have died from the cause eliminated are exposed to the causes remaining.

(iii) Population problems. In addition to employing the life table for mortality analysis, the student of population has found it a useful tool for studies of marriage, population growth, the labour force and the money value of a man to his dependent. The life table technique has been successfully used to evaluate the effectiveness of various population control programmes operated in a population during a specified duration of time.

(iv) Morbidity analyses. Reference has already been made to the use of multiple decrement tables and of select tables in disability studies. By means of the multiple decrement table, the incidence of certain morbid conditions may be estimated in the absence of direct observation. The requisite data for the computation are a life table for the general population, the prevalence rates for the morbid condition, and the mortality among those so afflicted. Such computations have been carried out for the chances of their eventually becoming diabetic and of eventually acquiring some cardiovascular-renal condition.

(v) Hospital studies. Life table techniques have also been adapted for the study of certain hospital data. Thus, on the basis of data regarding first admissions to New York State hospitals and life tables for the general population, computations have been made of the expectation of mental disease. A Canadian report describes the application of the multiple decrement table to patients in mental hospitals who are separated by discharge, death, or withdrawal from study. The expected length of hospital stay for specific morbid conditions has been estimated by constructing 'hospital stay' tables following the pattern of life tables.

(vi) Clinical medicine. The life-table technique was adapted to many areas in clinical and preventive medicine to compute mean times from entry to a study to the occurrence of an event such as a cardiac arrest, the occurrence of metastasis in cancer, the discontinuation of a contraceptive, healing of a foot ulcer, etc. Originally, the term "survival analysis" was concerned with time from treatment until death, hence the name, but now the term 'survival analysis" is applicable to many areas. When the outcome of a study is "time" between one event and another, a number of

issues arise: one, the times may not be normally distributed. Secondly, we cannot afford to wait until events have occurred to all the subjects, for example, until all are dead. There are likely to be dropouts and loss to follow-up. Thirdly, because the study has to be terminated, a few patients who are alive must be included in the calculation. These are usually called "censored" observations. Thus, the standard statistical techniques cannot usually be applied. The best method will then be the life-table technique modified and adapted by Kaplan and Meier, and others.

For example, if the event is the death of the patient, and so the censored data are those where the outcome is survived or unknown in analysing survival data, two functions that are dependent on time are of particular interest: the survival function and the hazard function. The survival function $S(t)$ is defined as the probability of surviving at least to time t. The hazard function $h(t)$ is the conditional probability of dying at time t having survived to that time.

The graph of $S(t)$ against t is called the *survival curve*. The Kaplan–Meier method can be used to estimate this curve from the observed survival times without the assumption of an underlying probability distribution. The method is based on the basic idea that the probability of surviving k or more periods from entering the study is a product of the k observed survival rates for each period (i.e., the cumulative proportion surviving) given by the following.

$$S(k) = P_1 \times P_2 \times P_3 \times \ldots \times P_k$$

Here, P_1 is the proportion surviving the first period. P_2 is the proportion surviving beyond the second period conditional on having survived up to the second period, and so on. The proportion surviving period 1 having survived up to period 1 is given by

$$\rho_1 = \frac{r_1 - d_1}{r_1}$$

where r_1 is the number alive at the beginning of the period and d_1 the number of deaths within the period.

Example. Let's start with 10 patients, and the first dies after 3 weeks and 6 days.

The proportion surviving to 4 weeks is therefore 9/10, or 0.9. Of the 9 patients surviving at 4 weeks, all survive for another 4 weeks. The proportion surviving to 8 weeks is, therefore, also 0.9. At this point a patient is lost to follow-up. Of the 8 patients still being observed after 8 weeks, a further one dies after 6 more weeks, so the proportion who survive from 8 to 14 weeks (or, in other words, the probability of survival to 14 weeks) is calculated as $0.9 \times 0.875 = 0.788$. Note that we have assumed that the patient lost to follow-up would have had the same survival experience as the patients who remained under observation. The results are displayed in the diagram, referred to as a 'survival curve'. In fact, it is a series of steps, or 'step function'. The median survival time can be read from the diagram; it is the time to which half (0.5) of the patients survive, in this example 26 weeks.

Confidence intervals for each survival proportion can be easily calculated.

Many examples of such applications are found in medical literature.

When we wish to compare two survival curves, the most popular statistical test is the use of log-rank test or the Cox proportional hazard analysis.

The log-rank test is used to test the null hypothesis that there is no difference between the population survival curves (i.e., the probability of an event occurring at any time point is the same for each population). The test statistic is calculated as follows:

$$\chi^2(\text{log rank}) = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

where the O_1 and O_2 are the total numbers of observed events in groups 1 and 2, respectively, and E_1, and E_2 the total number of expected events.

The total expected number of events for a group is the sum of the expected number of events at the time of each event. The expected number of events at the time of an event can be calculated as the risk for death at that time multiplied by the number alive in the group. Under the null hypothesis, the risk of death (number of deaths/number alive) can be calculated from the combined data for both groups.

19.4 COX PROPORTIONAL HAZARDS MODEL (COX REGRESSION)

The log-rank test is used to test whether there is a difference between the survival times of different groups but it does not allow other explanatory variables to be taken into account.

The Cox proportional hazards model is analogous to a multiple regression model and enables the difference between survival times of particular groups of patients to be rested while allowing for other factors. In this model, the response (dependent) variable is the 'hazard'. The hazard is the probability of dying (or experiencing the event in question) given that the patients have survived up to a point in time, or the risk for death at that moment.

In Cox's model no assumption is made about the probability distribution of the hazard. However, it is assumed that if the risk for dying at a particular point in time in one group is, say, twice that in the other group, then at any other time it will still be twice that in the other group. In other words, the hazard ratio does not depend on time.

Many excellent examples can be seen in published papers. The statistical software such as SPSS or Stata have the facility to carry out survival analysis and the various tests of hypothesis.

QUESTIONS

1. (a) When standardised death rates are used?
   (b) Explain the two methods of standardisation.
2. Explain the construction and structure of a life table.
3. Write a note on the applications of life table in medicine.

CHAPTER

20 Demography

20.1 INTRODUCTION

Demography is the study of population. More explicitly, it is the study of the growth, composition and distribution of human population and their interrelationship with social, economic and behavioural factors. The main source of data for demographic study is the census. In India census is taken once in 10 years. Surveys and registration statistics are also used as valuable sources of demography.

20.2 POPULATION AND GROWTH

The world population (mid-2011) was estimated to be 6,928 millions, with an annual growth rate of 1.088%. Developing countries account for most of the growth. India's population was 1,210.2 million on March 31, 2011 (provisional) and has increased by 181 million during 2001–2011. India's population was 17.31% of world population. The annual growth of India's population was 1.64% during the decade 2001–2011. The top five populous states in India were Uttar Pradesh (199.6 million), Maharashtra (112.4 million), Bihar (103.8 million), West Bengal (91.3 million) and Andhra Pradesh (84.7 million).

Although absolute population growth has increased, the growth rate has decreased during the decade 2001–2011. Since the beginning of last century India's population has increased more than four-fold. The growth rate of the population can be estimated in many ways, some methods are given in Section 20.8.

Table 20.1 Growth of Population of India: 1901–2011

Year	Population	Decadal growth		Average annual exponential growth rate (per cent)	Progressive growth rate over 1901 (per cent)
		Absolute	per cent		
1901	238,396,327				
1911	252,093,390	+ 13,697,063	+ 5.75	+ 0.56	+ 5.75
1921	251,321,213	– 772,177	– 0.31	– 0.03	+ 5.42
1931	278,977,238	+ 27,656,025	+ 11.00	1.04	+ 17.02
1941	318,660,580	+ 39,683,342	+ 14.22	1.33	+ 33.67
1951	361,088,090	+ 42,420,485	+ 13.31	1.25	+ 51.47
1961	439,234,771	+ 77,682,873	+ 21.51	1.96	+ 84.25
1971	548,159,652	+ 108,924,881	+ 24.80	2.20	+ 129.94
1981	683,329,097	+ 135,169,445	+ 24.66	2.22	+ 186.64
1991	843,930,861	+ 160,601,764	+ 23.50	2.11	+ 254.00
2001	1028,610,328	+ 184,679,467	+ 21.34	1.93	+ 331.50
2011	1210,193,422	+ 181,455,986	+ 17.64	1.64	+ 407.64

20.3 AGE AND SEX COMPOSITION

Age and sex are the primary variables studied in any population. The age-sex composition of the population at any given time is the result of past trends in fertility, mortality and migration. It in turn influences the current levels of birth and death rates and the rate of future population growth. Populations in specified age groups denote the quantity of a nation's human resources, size of school-aged population, basis for providing maternal and child health services and so on. Age-sex composition is an essential consideration for planning for economic development, social and health services, and human resources development.

Age structure can be used for estimating vital rates, particularly where registration system either does not exist or is incomplete or inefficient. The age distribution of successive censuses can be used for estimating net migration using survival ratios. These survival ratios are used for projecting population and also for determining probability of living. An examination of age and sex composition will help determine the quality of data collected. Age and sex ratios for each age group provide a clue for diagnostics of age distortion.

Sex ratio is an important demographic as well as social indicator. It is defined as the number of females per 1000 males—in some countries it is defined as the number of males per 1000 females. Sex ratio is determined by sex ratio at birth, sex differentials in mortality, and sex selective migration. The sex ratio of India was 933 females per 1000 males according to the 2001 census, and it has shown a declining trend from the beginning of this century. The low sex ratio indicates the low status of women and also excess of mortality of females compared to males.

According to the 2011 India census, males were 625.7 million and females were 586.5 million. The sex ratio of the population was 940 females per 1000 males indicating shortage of females; literacy was 82.4% among males, 65.46% among females and 74.04% among total population. These literacy rates are higher than that of previous census years but the difference between the sexes persists. Population aged 0 to 25 years account for 50% of total population. With declining birth rate coupled with very low death rates the population of the elderly increases.

Table 20.2 Selected Demographic Characteristics of India: 1901–2001

Year	Sex ratio	Population by age group (per cent)			Urban population (per cent)	Literates to total population (per cent)		
		0–14	15–59	60 +		Male	Female	Total
1901	972	38.2	56.7	5.1	10.8	9.8	0.7	5.4
1911	964	38.5	56.3	5.2	10.3	10.6	1.1	5.9
1921	955	39.0	55.8	5.2	11.2	12.2	1.8	7.2
1931	950	39.8	55.9	4.3	12.0	15.6	2.9	9.5
1941	945	37.9	57.1	5.0	13.9			
1951	946	37.6	56.8	5.6	17.3	27.2	8.9	18.3
1961	941	41.0	53.5	5.5	18.9	40.4	15.3	28.3
1971	930	42.0	52.0	6.0	19.9	45.9	22.0	34.5
1981	934	39.2	55.3	5.5	23.3	56.4	29.8	43.6
1991	927	37.2	55.4	7.3	25.7	63.9	32.4	52.1
2001	933				27.8	75.3	53.7	64.8

Note: The literacy rates prior to 1951 are based on the total population. The literacy rates for 1951, 1961 and 1971 are based on the population aged 5 years and above.
The literacy rates for 1981 to 2001 are based on the population aged 7 years and above.

The age-sex structure is depicted by the population pyramid (Fig. 20.1). A comparison of India and the USA indicates that India has a broad base implying

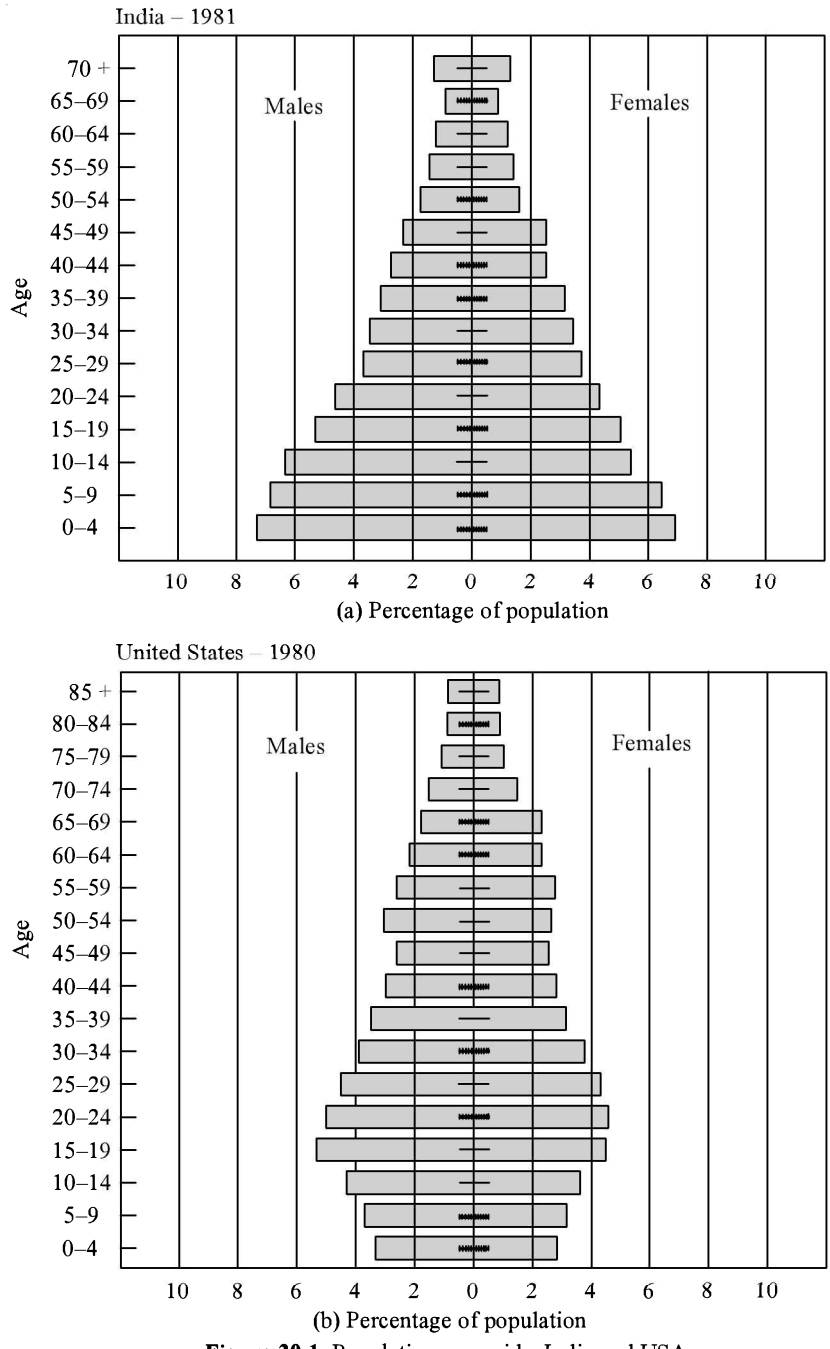


Figure 20.1 Population pyramids: India and USA.

high fertility and a large proportion of young people with a sharp tapering apex, suggesting high mortality at older ages whereas the USA has a narrow base and conspicuously broad top. In India, it is observed that nearly 37 per cent of the population are aged 15 or lower, and about 7 per cent are 60 and higher. This indicated effect of a small reduction in fertility over the recent decades.

20.4 DEPENDENCY RATIO

Total dependency ratio is defined as

$$\frac{\text{(Population aged 0 to 14) + (Population aged 60 and above)}}{\text{Population aged 15 to 59}} \times 100$$

Child dependency ratio is defined as

$$\frac{\text{Population aged 0 to 14}}{\text{Population aged 15 to 59}} \times 100$$

Old age dependency ratio is defined as

$$\frac{\text{Population aged 60 and above}}{\text{Population aged 15 to 59}} \times 100$$

The age structure of India reveals heavy dependency. A large dependent population tends to reduce savings and investments and thereby inhibits rate of economic and social advancement of a country as a large quantity of scarce resources are diverted towards current consumption. An increasingly large number of persons continue to enter the working ages swelling the ranks of the unemployed.

20.5 OTHER INDICES

Urban population was 27.8% in India in 2001 and it has more than doubled since the beginning of last century. The annual growth of urban population was 3.2% (during 1991–2001), which is considerably higher than the growth rate of the rural counterpart.

The main workers comprise 30.5% (males 45.2% and females 14.6%) of the population. The Total Workers (Main and Marginal Workers) were 39.1% of the population. It was 51.7% for males and 25.6% for females. It was higher in the rural (41.8%) than in the urban (32.3%) segment. This is true for males as well as females, the difference being considerably more among females.

One of the indices giving the quality of population is the literacy rate. In 2001, it was defined as the number of literates based on population aged 7 and above. Not only the overall literacy but also literacy of males and females was also increasing. There used to be wide gap between the rates of males and females but in 2001, the gap declined considerably.

20.6 FERTILITY AND MORTALITY

The major components of population change are fertility and mortality. Since the beginning of the 20th century, the birth rates and death rates have been declining steadily. Death rate, however, was declining faster than birth rate, recording ever-increasing population growth. This is the main reason for the population explosion.

The birth rate was 49.2 (per 1000 population) in 1901–11 and it came down to 25.4 in 2001. Kerala recorded the lowest birth rate of 17.3 and Uttar Pradesh, the highest rate of 32.1 in 2001. India's death rate was 47.2 in 1911–21 and it has dropped to 8.4 in 2001. Both the birth and death rates of rural areas have been always higher than those of the urban areas. Orissa had the highest death rate of 10.4 and Kerala the lowest (6.6) in 2001. There is a wide variation in the infant mortality rate over the states. Kerala has the lowest (11) and Orissa the highest (91) infant mortality rate. The expectation of life at birth of both males and females has also been showing an increasing trend.

The main factors influencing higher fertility were the traditional social norms and values, low literacy especially of women, low status of women, low urbanisation, certain religious beliefs against contraception, desire for large number of sons, especially for old-age security, low age at marriage of female, near universal marriage, low level of contraception of all kinds, and high infant mortality necessitating bearing of more children in order to have a few survive to adulthood. Fertility differentials do exist in all these factors.

Table 20.3(a) Fertility and Mortality Indicators: India (1901–2001)

Period	Crude birth rate	Crude death rate	Life expectancy at birth	
			Male	Female
1901–11	49.2	42.6	22.6	22.3
1911–21	48.1	47.2	19.4	20.9
1921–31	46.4	36.3	26.9	26.6
1931–41	45.2	31.2	32.1	31.4
1941–51	39.9	27.4	32.4	31.7
1951–61	41.7	22.8	41.9	40.6
1961–71	41.2	19.0	46.4	44.7
1971–81	37.2	15.0	50.9	50.0
1981–91	32.5	11.4	58.1*	58.6*
1991–2001	24.8	8.9		

* For 1987–1991.

Sources: 1. Government of India, Family Welfare Programme in India Year Book: 1986–87.
2. Sample Registration Bulletin, December 1987, December 1992 and April 2003.

Table 20.3(b) Fertility and Mortality Indicators—Rural, Urban: India (1991–2001)

Year	Crude birth rate			Crude death rate		
	Rural	Urban	Combined	Rural	Urban	Combined
1991	30.9	24.3	29.5	10.6	7.1	9.8
1992	30.9	23.1	29.2	10.9	7.0	10.1
1993	30.4	23.7	28.7	10.6	5.8	9.3
1994	30.5	23.1	28.7	10.1	6.7	9.3
1995	30.0	22.7	28.3	9.8	6.6	9.0
1996	29.3	21.6	27.5	9.7	6.5	9.0
1997	28.9	21.5	27.2	9.6	6.5	8.9
1998	28.0	21.1	26.5	9.7	6.6	9.0
1999	27.6	20.8	26.1	9.4	6.3	8.7
2000	27.5	20.7	25.8	9.3	6.3	8.5
2001	27.1	20.3	25.4	9.1	6.3	8.4

Source: Sample Registration Bulletins

There are variations in mortality due to biological aspects such as differences in age, sex; social and cultural aspects such as place of residence, occupational groups, education; environmental aspects such as air, water, weather conditions and physiological conditions including diet, smoking, blood pressure, weight or biochemical constituents of the body.

20.7 DEMOGRAPHIC TRANSITION

Most of the developed countries have undergone demographic transition characterised mainly by three stages. The pretransition stage was the era of high fertility and high mortality with low population growth, then the transition stage had declining mortality and fertility, mortality decreasing faster than fertility, with the consequence of big demographic gap or high population growth, followed by post-transition stage of low fertility and low mortality with the resulting slow population growth (Figure 20.2).

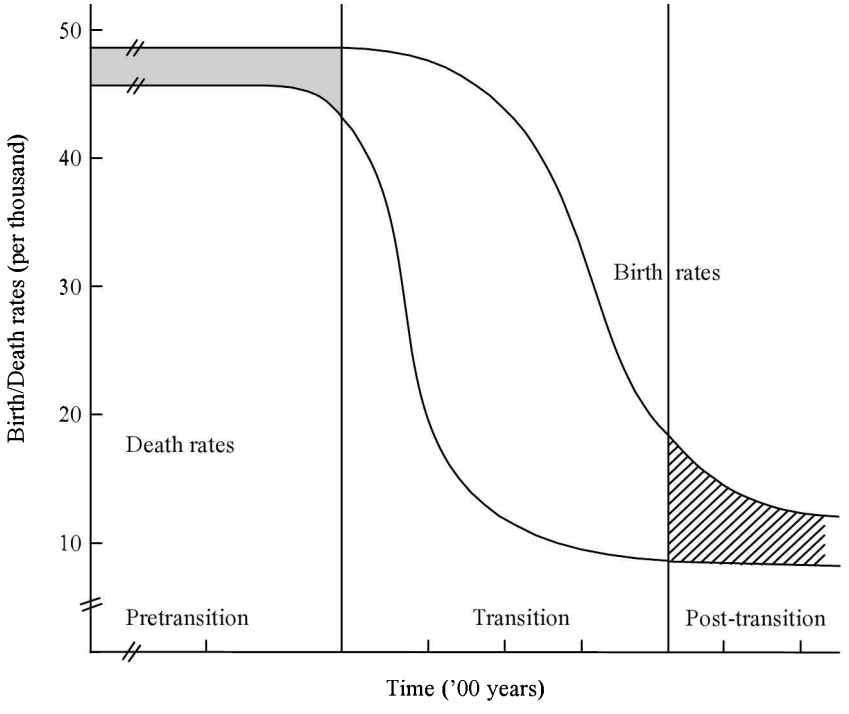


Figure 20.2 Demographic transition.

All countries of the world can be classified into any one of the three stages of transition. All the more developed countries are in the post-transition stage. Some of the countries took two hundred years to pass through the transitions stage (the middle stage). India is now in the demographic transition stage, with a fast declining mortality (death rate) and slowly decreasing fertility (birth rate) with the huge demographic gap resulting in the population explosion (Figure 20.3). All

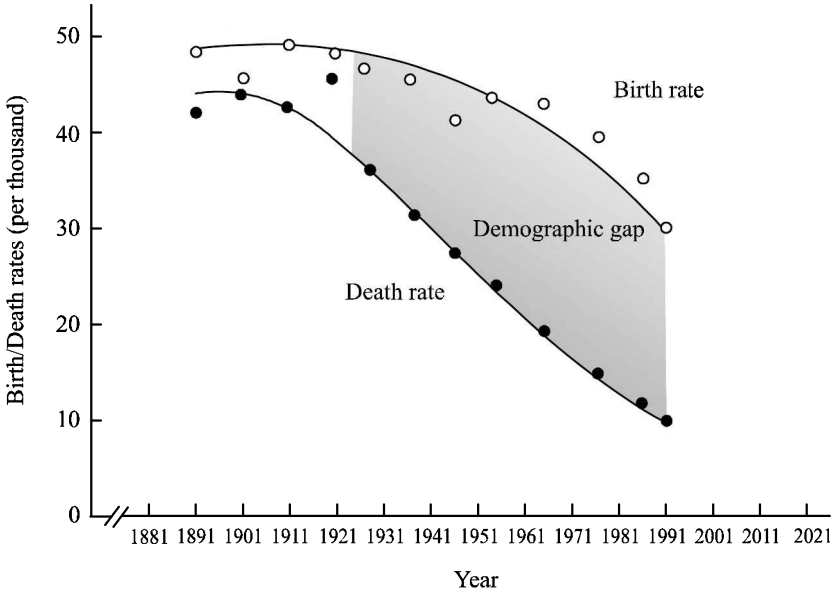


Figure 20.3 Demographic transition: India.

states in India are in different levels of transition and Kerala is nearing the post-transition stage of low mortality and reasonably low fertility with slow population growth.

20.8 POPULATION ESTIMATION

As population is liable to change every instant due to births, deaths, immigration and emigration, the population at any given time can only be obtained as an estimate. Estimates of population are required for the years falling between two census dates (intercensal estimates) or following the last census (post-censal estimates). Most annual rates are computed by using the population estimate referring to 1 July (mid-year population). Sometimes, as the result of a new census or some other fresh information, there is an abrupt change of considerable magnitude in the population estimates used for calculating vital rates. When such discontinuity occurs, reference to the change should be indicated, separating the years before and after the change.

A number of methods have been suggested for making population estimates of which the three most commonly used are described below.

20.8.1 The Method of Natural Increase

This method can be applied for an area where the registration of births and deaths is satisfactory and, further, if satisfactory statistics of migration are also available. The method consists of adding to the last census population the excess of births over deaths and the balance of immigration over emigration, during the period from the last census to the time when the estimate is to be made. The method does not

provide estimates of the population for any future date. In countries, where vital statistics and data on migration are of a high level of accuracy, the method provides good estimates.

Example. Using the statistics given in Table 20.4, estimate the mid-year population of town 'A' for the year 1993. (Mid-year population of town 'A' in 1990 = 134,000).

Table 20.4 Statistics of Town 'A' during 1990–1993

Year	Births	Deaths	Immigrants	Emigrants
1990	3705	1164	8220	8433
1991	3813	1114	9135	9895
1992	3856	1108	11,688	11,606
1993	4037	1127	8695	9542

We estimate the mid-year population of Town 'A' for 1993, by first calculating the net increase in each year as given in Table 20.5.

Table 20.5 Annual Net Increase in Population of Town 'A'

Year	Excess of births over deaths	Excess of immigrants over emigrants	Annual net increase
1990	2541	– 213	2328
1991	2699	– 760	1939
1992	2748	+ 82	2830
1993	2910	– 847	2063

Since the population for the year 1990 is known at the middle of the year, we calculate the net increase in the second half of 1990 by taking 1/2 of 2328, i.e., 1164. To this we add the annual increase in 1991 and 1992, i.e., 1939 and 2830, respectively. The addition of these three figures to the 1990 mid-year population gives the population estimate at the end of 1992. Since we want the estimate at the middle of 1993, we add to this population half the net increase in 1993. The mid-year population of 1993 is thus 134,000 + 1164 + 1939 + 2830 + 1032 = 140,965.

20.8.2 The Arithmetic Progression (AP) Method

In this method, it is assumed that the population increases or decreases by a constant figure from year to year between any two census years. The difference between the population figures of the two censuses gives the total increase of the population in 10 years. Dividing the number by 10 (the period between two censuses) we obtain the growth for one year. Assuming that this annual change occurs uniformly, we calculate the growth of the population from the last census to the time when the estimate is required.

Example. Population of India as enumerated on 1 March 1981 and 1 March 1991 was 683,329,097 and 844,324,222 respectively.

$$\text{Increase in 10 years} = 844{,}324{,}222 - 683{,}329{,}097$$
$$= 160{,}995{,}125$$
$$\text{Average increase in one year} = 16{,}099{,}512.5$$

To calculate the population of India in the middle of the year (1 July) 1982, since the census of 1981 was taken on 1 March 1981, we should add the increase for 1 year and 4 months. The mid-year population for 1982 is

$$683,329,097 + 1\ (4/12)\ (16,099,512.5) = 704,795,113.7 = 704,795,114$$

The A.P. method can also be used to estimate the population for any year after the last census. Such population estimates are called the postcensal mid-year estimates. The postcensal mid-year estimate for 1993 would be found by adding to the 1991 census figure the increase in 2 years and 4 months (since the 1991 census was taken as on 1 March 1991).

Thus, estimated mid-year population in 1993 is $844,324,222 + 2\frac{4}{12}$ $(160,995,125) = 881,889,774$.

20.8.3 The Geometric Progression (GP) Method

In the A.P. method described above, we have assumed that the increase in population is constant throughout the period. An alternative assumption can be made that it is the percentage rate of growth that remains constant. If a population has grown from one million to three millions in a year, then in the next year, the population would grow to nine million, at a constant rate (or according to the geometric law of growth). On the other hand, according to the arithmetic progression method, the population at the end of the second year would be only five million.

If P_1 and P_2 are the populations at the two census years, and r is the annual rate at which the population is growing in geometric progression, P_1 grows to $(P + P_1r)$ or $P_1(1 + r)$ at the end of one year. It grows $P_1(1 + r)^2$ at the end of two years, and so on to $P_1(1 + r)^{10}$ at the end of ten years. From this relationship, we can calculate the numerical value of the rate of growth and then estimate the population at any time as shown in the example below.

Example. From the data given in the previous example, estimate the mid-year population of India for 1982 and 1993 by the G.P. method. If 'r' denotes the rate at which population increases by geometric progression per year, then:

$$\text{Population on 1 March 1991} = \text{Population on 1 March 1981} \times (1 + r)^{10}$$

or

$$844,324,222 = 683,329,097 \times (1 + r)^{10}$$

Taking logarithms, we get

$$\log 844,324,222 = \log 683,329,097 + 10 \log (1 + r)$$

Therefore,

$$\log (1 + r) = \frac{\log 844,324,222 - \log 683,329,097}{10}$$

$$= 0.00919$$

The estimated mid-year population in 1982 = $683,329,097\ (1+r)^{1\frac{1}{3}}$

Log of mid-year estimated population, in 1982

$$= \log 683{,}329{,}097 + (1 + 1/3) \times \log (1 + r)$$

$$= 8.8346 + 1\frac{1}{3} (0.00919)$$

$$= 8.8469$$

Taking antilogarithms, we get the estimated mid-year population of 1982 = 702,883,336. Similarly, the estimated mid-year population for 1993 is

$$\text{Population on 1 March 1991} \times (1+r)^{2\frac{4}{12}}$$

Taking logarithms:

Log of estimated mid-year population of 1993

$$= \log 844{,}324{,}222 + 2\frac{4}{12} \times \log (1 + r)$$

Taking antilogarithms; the estimated mid-year population of 1993 = 891,250,938.

20.8.4 Comparison of the Three Methods

The natural increase method is suitable only when information on birth, death and migration is correctly known. While births and deaths may be fully recorded, migration details are hard to maintain and thus this method is not practicable for many countries. The A.P. method assumes that the population changes by a constant amount each year. This assumption, however, need not be correct. Nevertheless, the method has simplicity and the estimations have been checked by actual surveys to be reasonably correct. Thus, this method enjoys a wide applicability. The estimate by the G.P. method is considered more accurate than the one obtained by Arithmetic Progression Method. Since the calculation is slightly more difficult, this method is not so frequently employed.

It should be noted that the G.P. method gives a *higher* estimate than the A.P. method for the postcensal period, but a *smaller* estimate for the intercensal period.

QUESTIONS

1. (a) Define 'Demography'.
   (b) Explain the uses of the age-sex composition of a population.

2. Write brief notes on:
   (a) Population growth rate
   (b) Literacy rate
   (c) Dependency ratio
   (d) Demographic transition

3. Describe the various methods of population estimation.

CHAPTER

21 Introduction to Research Methods

21.1 INTRODUCTION

In very simple terms, research may be summarised as finding an answer to an unanswered question. If the answer is to be acceptable to the scientific community, the method employed for finding the answer should be scientific and flawless. The research method suitable for a study depends on the type of study and the design. Certain important concepts and steps commonly involved in research studies are discussed in this chapter.

21.2 RESEARCH QUESTION

The first requisite for research is selection of a problem, or topic or question. The aspects to be considered in selecting a question are, primarily, significance and relevance, researchability, feasibility and personal curiosity.

21.2.1 Significance and Relevance

The answer to the question should have the potential of contribution to existing body of knowledge and it should benefit the health professionals or the community. It will be good if the answers to the questions can help in formulation of health policies.

21.2.2 Researchability

All research problems may not be amenable to scientific study. To be researchable, the question should involve variables, which can be precisely defined and measured. Questions involving ethical or moral issues may not be researchable.

21.2.3 Feasibility

Feasibility in terms of physical settings and resources should be carefully examined before the question is taken up for study. The important considerations are, duration of the study and the seasonality (timing), accessibility and availability of subjects, extent of involvement of others and their cooperation, facilities and equipment that are available for research work (and time for procurement if not available), amount of financial resources needed and ethical aspects. The researcher may bank on his own experience or consult others working in the subject field to examine these aspects.

21.2.4 Personal Curiosity

Because a lot of time and effort have to be expended and full involvement of the researcher is absolutely necessary, genuine curiosity and personal interest of the researcher is essential for the success of the study. Personal interest will also maximise utilisation of opportunity for creativity and academic excellence.

21.3 LITERATURE REVIEW

Review of relevant literature serves as an essential background for any research. Critical examination of previous studies will help the researcher to formulate and delimit the problem, to minimise the possibility of duplication of research, to suggest a theoretical framework for the study, to learn from the reported experience of others about its feasibility to critically evaluate the various methods used by others and choose the most appropriate design for the investigation and so on. Such a review also helps to discuss the results and draw conclusions.

The review of literature is made easy by various publications of index such as Index Medicus and also availability of these in electronic media, such as Medlars, Medline, PubMed, Embase, etc. The researcher can locate the needed references to literature cited with reasonable ease with the help of these. One could also identify agencies or organisations such as Cochrane Library, Inforep, etc., which provide abstracts and references on selected topics.

In any publication, the references to literature cited are placed at the end as Bibliography. Instructions for citing references vary from one journal to another and also for articles in a journal or book and unpublished data. In general, the references are listed either alphabetically by first author's surname or numbered consecutively as they appear in the text. The latter, known as the Vancouver style, is considered more helpful and followed by most accredited journals. In citing, the following order is used: Names and initials of all authors, Title of article, Journal, Volume number, Year and Page numbers.

21.4 THEORETICAL FRAMEWORK OR MODEL

A theory is an abstract generalisation of an observed relationship, which may be simple or complex. With the help of previous findings culled out from literature and with the imagination of expected results of the research, it is desirable to develop a conceptual framework, which leads to a theory generalising the phenomena studied. This has, within itself, speculative or predictive elements that go far beyond the evidence available from the research data. A model is a symbolic representation of a relationship, observed or expected. Models use symbols to depict a theory or conceptual framework.

21.5 RESEARCH PROTOCOL

The practical problems encountered in an investigation regarding objectivity and precision of the data collected and the specification and standardisation of the

methods and procedures vary greatly with different studies. If prior knowledge in these matters is not available, a pilot or exploratory study may be essential. Pilot surveys are particularly useful in large-scale surveys in order to test and improve field procedures and schedules and to train the personnel involved in the collection of data. Suppose we want to investigate the incidence of, and factors contributing to, rheumatoid arthritis in a particular area, there would be the initial problem of definition, the difficulty of obtaining reliable data on past attacks, the complication of variation in judgment between different observers, the need to produce results in a form comparable with other surveys, etc. We would want to collect not only clinical data but a great deal of socio-medical information. This brings in the question of the type of interviewers to be employed and their training. Then, there is the problem of actually selecting sample of individuals for questioning or examination. This must be done so as to avoid many kinds of bias that may arise. To mention only one kind, if visits are made only during the day we are likely to obtain access to an excessively high proportion of houses where there are young children. Again decisions have to be taken about the failure of individuals to cooperate or their refusal even to be contacted, and about whether all the work must be done by medically qualified persons or whether some duties can be undertaken by a social worker, etc.

For almost all studies it is a great advantage to draw up a protocol for the study. In this protocol, the decision taken regarding the items requiring consideration are described clearly and fully. This protocol serves to introduce to every worker involved in a project the objectives of the projects and other details and procedures proposed to be followed. Every worker is guided properly to perform his responsibilities in relation to the project to achieve the objectives in an efficient manner. In the process of drawing up of such a protocol during the planning stage, the investigator is led to think more clearly and fully on various aspects of the study.

The following items generally require careful consideration and should be spelt out as clearly and in as much detail as possible:

1. Aims and objectives; Hypotheses to be tested
2. Population or categories of material to be covered; criteria for inclusion or exclusion
3. Sampling method; number to be studied; controls; method of allocation
4. Items of information to be collected and method of collection of data, Pilot Study
5. The problem of non-response
6. Processing and analysis of collected data
7. Time required; facilities available
8. Experience in other studies.

These items cannot be considered in isolation from one another. To a greater or lesser extent, any decision taken on one question will influence decisions to be taken on the others. Correct decisions cannot be taken without sufficient knowledge of the nature of material to be covered.

21.6 FORMULATION OF OBJECTIVES AND RESEARCH HYPOTHESIS

At the outset, it is essential to be clear what an 'objective' is, as compared to more general terms such as aim, purpose, goal, etc. The objective of a research project should state what is to be achieved by the research. It helps to focus on the study and avoid collection of data that are not strictly necessary for understanding or solving the identified problem.

Objective can be one only or several, but it is best to limit the objective to not more than two broad (or general) objectives, which could then, if necessary, be made as specific as required to organise the study into clearly defined parts or phases. However, care should be taken to list the objectives in a logical sequence, and in a coherent manner.

Properly formulated objectives will help developing the rest of the protocol in choosing the appropriate study design, data collection mechanisms, type of statistical or other analyses interpretation and utilisation of the results.

Objectives must be phrased in operational terms, and use action verbs that are specific enough to be evaluated. Examples of action verbs are: to determine, to compare, to verify, etc. It is best to avoid non-action verbs such as, to appreciate, to understand, to study, to know, etc., which are difficult, if not impossible to measure. It is sometimes said that objectives should be "SMART": Specific, Measurable, Achievable, Relevant, Time-bound.

A hypothesis is a prediction of a relationship between one or more factors underlying the problem under study, which can be tested. Objectives can be stated in a question form or as a hypothesis to be tested. For example, the objective could be to determine the incidence rate (or prevalence, frequency, relative risk, etc.) of a disease, and further to examine if this varies by age or sex or area and so on. The statistical handling of these two types of inferences, viz., estimation and testing hypothesis, was discussed in Chapter 12.

21.7 POPULATION UNDER STUDY

Almost all studies are based on samples but the inference should apply to the 'population' from which the sample was chosen.

The categories, or types of material, which require to be included in a study, and its geographical scope, are conditioned in broad outline by the purposes of the study, and by administrative and research requirements related to these purposes. Within this broad outline, however, there is often a certain amount of latitude, and careful consideration should, therefore, be given to the inclusion or omission of various categories in the population. For example, the question of whether or not minor categories should be included depends mainly on the purposes for which the information is required. A case is sometimes made for the inclusion of certain categories on which the information is intrinsically of little interest, in order to ensure comparability with the results of previous studies, or with the results of parallel studies in other countries. Categories, which are of very minor importance in one country, may be of great importance in another. By excluding unimportant marginal categories, particularly those on which the collection of information is likely to be

especially difficult, the task of collecting the information may often be very materially simplified, without seriously reducing the value of the results. Decisions as to their inclusion or omission should be taken primarily on the grounds of their importance in the country in which the study is being done, without undue regard to definitions designed to insure formal uniformity of world statistics.

21.8 STATISTICAL DESIGN

The validity of inference based on statistical analysis of data collected for an investigation will depend in a large measure on the way in which the investigation was planned and executed. Appropriate design and careful execution are necessary to provide definite, unambiguous answers to the questions; and to allow the fullest use to be made of the materials and facilities.

21.8.1 Risk of Bias

In many biological investigations, the problem of bias in the selection of cases for study, in the adoption of therapeutic procedures and in making observations, is likely to vitiate the ultimate results of these investigations. In order to safeguard against the risk of bias spoiling the study, the investigator must carefully follow appropriate sampling techniques and suitable randomising procedures. More details on these procedures are given in another chapter.

21.8.2 Precision of Results

The variability observed in the material or subjects under study influences the precision of the results of the study. This variation, if large, results in so much imprecision that a reliable conclusion from the data may not be possible. One way of reducing such imprecision would be to collect data on a large number of subjects. This may be difficult because the process of increasing the number may be expensive and in certain cases it may not be possible due to a restricted population being available for study.

The following refinements may be attempted to reduce the imprecision (or to increase the sensitiveness as it is called), of the experiment, which will eliminate from the experimental material certain causes of disturbances. How far we can do this is a matter of judgment and will more often depend on practical considerations.

21.8.3 Common Designs

In clinical experimentation, the widespread acceptance of useless or even harmful treatment and subjecting many patients to unnecessary experimentation can be avoided if modern statistical methods of design and analysis of experiments are appropriately used by clinical investigators. An investigation can be planned incorporating one or more of the refinements mentioned below:

1. The experiment may be conducted on a population in which the variation is known to be small.

2. The population may be divided into sub-populations or strata, of similar individuals, each sub-population to be studied separately.

3. The information contained in related variables may be utilised.

Many designs of experiments and methods of analysis have been developed. Even though most of the designs have been developed to satisfy the needs of workers in other branches of scientific research, their use has been on the increase in clinical research. Let us consider a few simple designs, which are widely used, in clinical research:

(i) Group comparison. A simple and a very common pattern of experimentation is to allot patients at random to the various treatments—drugs, therapeutic procedures, etc., and study them. In the successive stages of giving the treatments and of recording the results of the treatments, safeguards must be provided against the risk of introducing bias. If these stages of experimentation have been properly carried out, the results obtained for the different groups can be subjected to statistical analysis and valid conclusions can be drawn of the responses to the different treatments. This simple design has tremendous possibilities in medicine because of its extreme simplicity of application within the condition of ordinary clinical work.

(ii) Individual comparisons. For the type of experiment mentioned in group comparison, refinement may be possible when the experimenter has a more exact knowledge of the behaviour of the experimental material. Individuals who are expected to behave alike, apart from random variation, when they are given the same treatment, can be grouped and allotted equally to the different treatments. Identical twins and littermates of the same sex are examples of such groups. The responses of the members receiving the different treatments in each group are compared; and the conclusion is based on the differences observed in all such groups, considerable increase in the precision of the experiment will be possible by following the procedure. The number of individuals to be selected for each group will depend on the number of treatments to be compared.

(iii) The concurrent assessment of several treatments. When several treatments are offered for testing in quick succession and the required number of subjects is not available during a reasonable period, it is of considerable importance to use the available material as efficiently as possible. One simple way of economising is to test the effect of several treatments simultaneously in a single trial. When, for example, it is desired to test the effect of two treatments, the treatments given and the number of patients allotted for each treatment are shown in Table 21.1 under two experimental designs:

Table 21.1 Allocation of Subjects in Simple and Factorial Design

(A) Simple design		(B) Factorial design	
Treatment	No. of subjects	Treatment	No. of subjects
I	40	I	30
II	40	II	30
Control	40	I and II	30
		Control	30
Total	120	Total	120

When we use 120 subjects, we notice that in the case of design A, the responses of 40 patients will be available for each treatment. If treatments I and II are independent in their effects, design B results in considerable saving. For example, to find out the effect of treatment I, the following comparisons are possible, given 60 patients for each treatment:

Treatment	I	(30)	vs.	Control	(30)
Treatments	I and II	(30)	vs.	Treatment II	(30)

The simultaneous testing of treatments is important in that it allows observations to be made on the interaction between the treatments. The suitability of design B will depend upon the practical consideration of administering the treatments simultaneously.

Apart from the designs mentioned above, there are many other designs, which will be of importance in the medical field. For further details of these and other designs, the student is referred to important textbooks mentioned in the bibliography.

In an effort to make a contribution to knowledge, the clinical research worker's immediate objective would be to get as clear results as possible, and as speedily as possible. It is desirable to know the answer to questions as quickly as possible using the minimum number of patients for the minimum amount of time.

21.9 SAMPLE SIZE

An important item which needs consideration while planning any investigation is the number of cases which will be studied by the investigator. Decision on this assumes previous knowledge by the investigator of the variability of the characteristic in the population studied. Taking into account the nature of the measurement of the observation and the design of the investigation, the number of cases to be studied is calculated for the following specifications given by the investigator:

1. The precision with which the parameters should be estimated, or differences identified.

2. The degree of safeguard allowed against the risk of drawing wrong conclusions.

3. An approximate idea of the population value.

In general, the greater the precision and safeguards needed, the larger would be the number of cases to be studied in specific circumstances.

Example. The prevalence rate of some disease in a place is expected to be close to 25%. How many cases do we need to select randomly to estimate the prevalence rate with an error not exceeding 6% (± 3%) with the risk of a wrong conclusion not exceeding 5% (Type I error)?

The answer can be worked out from the binomial probability distribution tables. It is found that 800 is the number to be studied to satisfy the specifications given above. More details on calculation of sample size are given in Chapter 23.

21.10 METHODS OF DATA COLLECTION

One of the first and most decisive steps in any enquiry is the construction of a questionnaire or schedule containing the questions to be asked and the items to be recorded. If the questions are incomplete, ill-conceived or inadequately answered, no statistical analysis can compensate for the resulting defects and produce the answers that the investigator had hoped to get.

We must remember that many persons find the completion of any form a difficult task. Drafting of clear and unambiguous question is, therefore, very essential. In formulating questions or headings for inclusion on a form, the following basic principles will have to be borne in mind:

1. One should consider closely whether there is any ambiguity in the question and, consequently, in the answers likely to be received.

2. As far as possible, every question should be self-explanatory or require the respondent to consult a separate sheet of instructions. The importance of this principle will, of course, vary with the circumstances. If a few highly trained persons are making the observation, conducting the clinical examinations, or whatever it may be, then clearly they can be relied upon to turn to, and follow, detailed instructions. But if the form is to be completed by large numbers of less trained and less interested persons, experience shows that they cannot be relied upon to read and remember extensive footnotes or instructions.

3. Almost invariably, every question should require some answer. Without that precaution it is often impossible to know whether a person did not, in fact, possess some characteristic or whether no information was sought or obtained. Every question, therefore, needs some final category to make certain that some answer is given—whether the final category should be 'not known', 'no information', 'unspecified', or 'others'.

4. The degree of accuracy to which measurements are required should be specified. For example, should blood pressure be recorded to the nearest millimetre, to the nearest 5 mm, or to the nearest 10 mm? If no specification is given, the recorders will inevitably vary amongst themselves in the accuracy to which they work.

5. It would seem needless to say that much thought should be given to whether it is likely that a particular question can be answered adequately by anyone at all; yet experience suggests that the advice is not so needless. Under this heading there will also be occasions on which close attention must be paid to the problem of 'observer error or variation'.

6. Any form or record, which must be completed by many persons, should, as far as possible, be made simple in working and logical in the order of its questions. The amount of work required of the respondent may often be reduced by putting the question in such a form that the answer demands only a cross or a tick or ringing of one specified category. Such answers are also very easy to tabulate.

7. Much attention should be given to the number of questions. Obviously this must vary widely with circumstances, but at the same time there are probably no

circumstances in which the constructor of a form should not ask himself of every question: 'Is this question essential? Can I obtain useful answers to it? Can I analyse them usefully at the end?' Such self-discipline is likely to reduce the size of any form.

8. The informant has to be disabused of any misapprehension regarding the purpose of collection. The fear that the information may be used against him for purpose of taxation or offensive legislation makes him hesitate before giving true answers to questions on wage or income or age at marriage. A general desire to declare oneself as of higher status and superstitious beliefs are obstacles that must be faced.

9. Finally, the physical aspects of the record forms need to be considered. The schedule should be easy to handle. The space allotted to answers should be adequate. The purpose for which the investigation is attempted should be mentioned clearly in the schedule itself. The use of capitals, letters of different types, italics and rules of varying thickness and form is a device that appeals to the eye and makes the schedule agreeable as well as easily understood.

Forms should always be subjected to a preliminary trial in the field. Only in this way will minor faults be discovered. In the case of questionnaires this test is best arranged in two parts:

1. A trial by investigators who are fully experienced in questionnaire work, and who are conversant with the problem under investigation; and

2. A trial by investigators of the type that are to be employed in the survey.

The first trial will serve to determine, whether the questionnaire is in the current form most suitable for eliciting the required information from the respondents, and the second trial will provide information on whether the questions and associated instructions are understood by, and within the capacity of, the investigators.

Different methods are adopted for the collection of data for investigations. The nature of investigation, resource available, and the interests and characteristics of the people to be studied are among the factors influencing the choice of the method to be used for a particular investigation. A brief discussion of the commonly used methods is given below:

21.10.1 Personal Observation

The investigator himself studies the facts and collects the required data. By this method the investigator gets the opportunity to observe many details and to detect the salient features of complex phenomena. The investigator himself better does qualitative evaluation. For instance, the grading of the dampness of a house may have to be done only by a subjective assessment of various factors and no method other than the personal observation is likely to yield better information.

An observer, however, tends to make subjective evaluation based on his own judgment. When more than one observer is involved in an investigation, sufficient standardisation of evaluation techniques is necessary.

21.10.2 Interview

The research worker or his assistants establish contact with the respondents and interview them for the purpose of securing specific information from them. Such contacts may be made by visiting the individuals, over the telephone or meeting them in groups. The information required from the respondents is explained to them. Personal contact is preferable to phone calls in many instances because in personal contact more time can be spent, interest sustained for a longer duration and more information collected, in addition to the possibility of observing the reactions and attitude of the respondents.

A telephone interview is useful for obtaining a few facts as the duration of interview is short. Even though this method is less expensive than personal interview, the respondents who can be contacted by telephone are limited in number and consist of certain categories of persons only.

Group interviews involve bringing people together in a group to complete questionnaires and to encourage complete replies. In a group the respondents may feel less free to ask questions than if alone. They may also tend to communicate with each other to ask the opinions of others, which are then reflected in their own replies.

21.10.3 Mailed Questionnaires

Questionnaires or schedules are sent by post directly to the persons to be filled in and returned by post. This is generally less expensive method of collecting information. This method is particularly useful when the respondents are widely scattered geographically and good relationship exists between the respondents and the investigator. Respondents can spend more time thinking about replies; and the replies will be uninfluenced by interviewers.

Mailed questionnaires must be shorter than those used in personal interviews and must contain sufficient description of the information required and instructions for filling in and dispatch of the questionnaires; otherwise questions may be misunderstood and proper returns may be difficult. Non-response is usually more with this method. Those who do not respond may not be of the same types as those who respond.

21.10.4 Other Methods

Information may be collected through case studies, assembling data from existing sources, keeping systematic records by extension workers or other people, etc. The information provided by such methods may need to be carefully assessed for its suitability to meet the requirements of the investigator.

21.11 METHODS OF DEALING WITH NON-RESPONSE

Unless non-response is confined to a small proportion of the whole sample, the results cannot claim any general validity. Every effort must, therefore, be made to reduce non-response to negligible proportions.

Non-response is usually most serious in postal questionnaires. Delays in response can also sometimes be very troublesome, particularly when the results are required very quickly. A rigorous system of dealing with failure to respond and delay in response must, therefore, be instituted at the outset. The first step is to send a follow-up letter, but if this does not produce the required effect, the possibility of using more intensive methods such as telephone calls and personal visits must be considered. These will require a special regional organisation.

In interview surveys, the amount of deliberate non-response is usually small. If it is not, the questionnaire and the type of investigator used should be reviewed. Revisits by special investigators can be tried, but are not likely to be very effective. Initial non-response due to failure to contact the respondent can be very troublesome, and there is no proper way of dealing with this except by persistent call-backs. Enquiring from the neighbours when the respondent is likely to be at home, or where he can be found so that an interview can be arranged can often reduce the number of call-backs. Call-backs are also required because the respondent, though willing to give the information, is otherwise engaged at the time of the first call.

No study can succeed unless the population concerned extends willing cooperation. In achieving this first step is to explain the purposes of the study to those who are going to be asked to cooperate. To achieve cooperation it is also essential to ensure that the study causes as little inconvenience to participants as possible. The methods of investigation must cause no distress. The timings of the survey must be convenient and, if possible, transport should be provided to bring people to and from the centre where the investigation is being carried out. It is important that they suffer no financial loss in attending. In surveys inducements have been used to encourage people to attend, by money payments or even by offering prizes for lucky numbers. It is unlikely that any survey will end by producing satisfactory records for every person in the defined population or sample. It is tempting to assume that these lapses are only random sub-group, differing in no material way from the subjects successfully investigated. There is little or no evidence that this assumption is justifiable. In fact, Cochran found that a significantly increasing proportion of cases with tuberculosis in successively seen one-third of a coalmine population, in which more pressure was brought to bear on the late-comers than on those first to come to be X-rayed. The true prevalence in the entire population was 50 per cent, greater than the first two-thirds. It would reasonably be argued that in general those who fail to cooperate are either disproportionately abnormal; it is unduly optimistic to assume that neither possibility outweighs the other. In several surveys, the lapse rate has been greater in the higher social classes.

It is essential to keep the lapse rate as low as possible, and it is advisable to find some kind of ancillary information which can be obtained from everyone in the defined population or sample and which can act as an index of the representativeness of those from whom full information is obtained. Clearly, such information cannot be conclusive, but even a knowledge of age, sex and social class distribution of the lapses can help to support or throw doubt on the claim that the lapses are no more than random. Sometimes it may even be possible to obtain a reasonably comprehensive medical history in the course of an unsuccessful attempt to persuade a reluctant subject to collaborate.

21.12 DATA PROCESSING

Every good study considers at the beginning, the aspects of processing the data and of analysing them. The kind of data processing that may be applicable has to be decided beforehand, even if the trial is a small one. Chapter 25 gives some of the methods used in data processing.

The earlier chapters of this book have given some of the fundamental aspects of statistical analysis. It is advantageous for the research worker to have knowledge of statistics, so as to be able to do things himself. It also adds immensely to the interest and value of medical research. Nevertheless, for the amateur, it would be wise to have the results checked by an expert, because it is easy to make a gross error in a field in which one has not been bred. It must be admitted, however, that many research workers in the clinical field whose contributions are of great value have such an antipathy to manipulating the relatively simple formulae, which the practice of statistical analysis demands that they must rely entirely on a professional statistician to do this for them. To get the best out of their results, however, they would be wise to plan their investigation in the first place in consultation with a statistician, one of whose qualifications is that he has mastered the science of the design of experiments. In this way, material will not be wasted and so expense and labour will be saved. Appropriate statistical techniques should be applied and if the trial was planned carefully, the analysis can be made more effectively.

21.13 THE CONSTRAINTS

There are three major constraints—time, finance and ethics, which are discussed briefly.

21.13.1 Time

Most studies are time-bound and have deadlines. The problem selected must be one that can be completed within the time available. This means that the scope of the problem should be sufficiently restricted to allow enough time to be allotted for the various steps. Research activities usually require more time to accomplish than one thinks. It is, therefore, essential to draw up a time budget, indicating each step of the research process and its duration. This will be a useful guide in monitoring the research effectively.

21.13.2 Finance

The researcher should think very carefully about projected expenses before the final selection of a problem is made, as the work should be done with a limited budget. Some areas of expenditure which should be considered are:

1. Initial work such as literature review, books and journals, internet browsing, photocopying of articles.
2. Salaries and wages of personnel for the project.
3. Travel and transportation or vehicles and their maintenance.
4. Printing and stationery and related items.

5. Statistical services, computer services and consultation.

6. Postage, fax, telephone and communication, rent, electricity.

7. Subject costs: payment or incentive to subjects, if any, cost of animals if involved, and maintenance.

8. Laboratory equipment, calculators, typewriters, computers and accessories.

9. Institutional overheads.

10. Contingency.

21.13.3 Ethics

All aspects of a study including the design should satisfy ethical principles. Unethical practices such as violation of people's rights and dignity, or a denial of services that otherwise would be available, should be avoided. Indeed, if it is not possible to do an ethical study, then the study should not be done. This point cannot be compromised.

Informed consent is an important aspect of any study. The subjects should be informed, preferably in writing, in the language in which they understand the purpose, the risks and the benefits of enrolling in the study, the voluntary nature of participation, the freedom to withdraw from the study at any time without losing the benefits including the services in the future, the provisions for compensation for unpredictable adverse events, the addresses and phone numbers of the persons who could be contacted in case of any doubts regarding participation or in emergency, the legal liability of the researcher and so on, and the consent of the subjects should be obtained. Special provisions are to be made for getting informed consent from subjects such as children, the mentally ill, prisoners etc.

The study proposal needs to be sent to the organisation's Institutional Review Board (IRB). The IRB will review the proposal to make sure that ethical issues are handled correctly and that the study procedures meet all the legal requirements necessary to conduct research.

These and other constraints may affect the quality of research. These should be carefully considered at the planning stage by the researcher.

QUESTIONS

1. What are the items to be included in a research proposal?

2. Explain the following:

(a) Pilot study

(b) Literature review

(c) Theoretical framework

3. Briefly state the ethical considerations of a research study.

CHAPTER

22 Interventional Studies

22.1 INTRODUCTION

As seen from the concepts described so far, the task of making additions to scientific knowledge by experimentation involves the fulfilment of the requirements of experimental procedure and of statistical analysis. In spite of this fact, some experimenters continue to work with variable material without proper design and to obtain discordant results incapable of being fitted into a scientific system. Controversies sometimes arise because different experimenters get diverse results for the same problem. In other cases, methods of statistical analysis are employed which result in definitely misleading estimates of error. Also, methods of experimentation are used which cannot give a valid test of experimental results. The common procedure of consulting a statistician or statistical principles after an experiment or investigation has been completed, is equivalent to holding a post-mortem. Perhaps the only interpretation of the data that can be made is to name the cause from which the experiment died. But when research workers turn to sound methods of statistical analysis which involve carefully planned experimental designs, difficulties of the types enumerated above tend to disappear.

22.2 GENERAL PRINCIPLES OF EXPERIMENTAL DESIGNS

Modern ideas of experimental design differ sharply from earlier or traditional ones. It has long been an admonition in philosophical treatises of scientific experiment to hold constant all except one of the factors in a complex so that its effect may be determined. The difficulty in applying such a principle, particularly in branches of science where the data of the research worker are subject to all sorts of fluctuations, had long been recognised by critical workers. The liberation of the research worker from stereotyped experimentation is relatively recent. Procedures have been developed for experimentation under varying circumstances. The subject of the design of experiments has become large and is undergoing further development as new possibilities of the methods and of their correct application become better understood.

Some general principles are:

1. To use a simple and symmetrical pattern of experimentation and to allot patients at random to the various treatments involved. If this is done properly, not only is the subsequent numerical analysis comparatively easy, but also many sources of invalidating bias are avoided.

2. To *rely* only upon the evidence from the data themselves when explaining any inequalities which arise in certain comparisons under consideration. Once an experiment has been designed and executed, its interpretation must be based on its own evidence. Violations of this principle and arbitrary corrections based on an *a priori* basis without reference to the information provided by the data themselves are probably the source of such statements as 'anything can be proved by statistics', or of the credo 'lies, damned lies and statistics'.

A self-contained experiment implies:

1. use of control

2. provision of a means to secure a valid estimate of the experimental errors, and

3. reduction or elimination of non-sampling errors.

22.2.1 The Use of Controls

By the use of controls, experiments become comparative and not merely absolute. The need for controls is clearly illustrated by the following example:

Assume that an experimenter working with animals injected some fluid into 3 rabbits and found that all 3 got violent and suffered prolonged convulsions followed by death within a period of 24 hours. In support of his conclusion that the injected substance was the cause of the death of the animals, the experimenter might draw from his own previous experiences or from those of rabbit breeders in general. Admittedly, only rarely would three designated animals die in the way described within such a short period of time. How would the conclusion have been made stronger if the experimenter had taken the precaution to inject a number of control rabbits with a neutral substance at the same time at which he injected his experimental animals? The answer to this question provides the rationale underlying the use of controls. The controls are used to exclude, at a designated level of probability, a number of alternative interpretations of experimental results—possibilities which have individually and collectively an unknown probability of having occurred. For example, the rabbits might have been ill from tetanus, hydrophobia, cholera, or some other unsuspected epidemic disease; perhaps the needle was infected with a poisonous substance; or it might be that the experimenter's stock was genetically of a kind which reacted in this way in general to injections. Suppose, however, that the experimental rabbits had been randomly chosen from the whole herd, and the controls included. If their reaction was clearly different from that of the controls, then there was available a precise measure of probability for causes other than the experimental factor for having brought about the observed result. The weight of previous or outside evidence is even much less when the object of the experiment is quantitative, because such evidence is usually very indefinite or highly variable. Thus, the essential condition for controlling the interpretation of experimental results is the provision of comparisons between two or more unlike variants.

Different types of controls which are commonly used may be mentioned here:

(i) Historical control. (or comparisons with a group which was treated during an earlier period using another form of therapy): In dealing with such controls, care must be taken to find out if the group treated earlier had other distinguishing features or differences, e.g. in other subsidiary procedures during treatment such as in nursing care, environment, etc. Usually, this is not an ideal control.

(ii) Geographical control. (or comparisons with a group treated elsewhere with a different form of therapy): This also suffers from the drawbacks of historical controls and cannot be an ideal control.

(iii) Volunteer control. Use of volunteers as controls can be misleading if they are a selected group and differ from the non-volunteers in terms of their interest in the study, socio-economic or cultural background, etc.

(iv) Concurrent control. The control group is observed simultaneously with the treated groups. This method usually provides the best form of comparison. Here the criteria for inclusion in the trial must be the same for both the treated and the control patients, as well as later obervation and assessments.

22.2.2 A Valid Estimate of Experimental Errors

In any experiment there are factors which are susceptible to some degree of control by the experimenter but their effect cannot be entirely eliminated, owing to chance fluctuations. Many of the factors giving rise to these fluctuations which affect performance are small in size and random in incidence, so that it is impossible to present an exhaustive list of all the sources of variation in the experimental material. It is customary to designate the component of variation associated with the random variation of the experimental material as 'experimental error'.

We require replication and randomisation to provide a valid estimate of the 'experimental error'.

(i) Replication. This is the process of repeating the same treatment on more than one object of the experimental test. Replication is essential in the first place since it is a means of diminishing the experimental error. The second function of replication in an experiment is to provide the data from which the appropriate estimate of experimental error can be calculated.

(ii) Randomisation. Randomisation is the procedure of making certain that the probabilities of being subjected to like treatments are equal for every relevant pair of objects in the experiment. It is worthy of note that the object of randomisation is not to increase the precision of the experiment but only to guarantee that whatever precision the experimental arrangement is capable of providing, is neither over nor underestimated.

22.2.3 Non-sampling Errors

There are a number of factors, both quantitative and qualitative, which may contribute to make the actual error small. Some of these are: accuracy of measurement of the

criterion, improvement in the techniques of controlling non-experimental factors, care in ensuring that in the experimental material the general conditions are those occurring in the population, measurement of controls under as nearly as possible the same conditions as those for the experimental subjects including time, and the greatest possible avoidance of hidden systematic errors as well as subjective errors. Only when sufficient care has been taken to ensure that these working errors have been reduced to unimportant levels can improvement of the replication and the organisation of the structure or arrangement of the experiment be expected to achieve increased precision or sensitiveness. The process of reducing the working errors begins with the largest sources of error, and continues until sources of error that hitherto seemed inconsequential, become significant by limiting the value of the whole enterprise.

22.3 COMMON DESIGNS

Many designs of experiments and methods of analysis have been developed. Even though most of the designs have been developed to satisfy the needs of workers in other branches of scientific research, their use has been on the increase in clinical research. A few simple designs which are widely used in clinical research are described below. Apart from these there are several other designs that can be made to suit a particular research problem. It would be necessary for the medical worker as well as statistician to get together both in planning and in the analysis.

22.3.1 Single or One-way Classifications

A simple and very common pattern of experimentation is to allot patients at random to the various treatments—drugs, therapeutic procedures, etc.—and study them. Suppose, we wish to determine the efficacy of a new treatment at 3 dosages. Determine the criteria for inclusion into the study and assign the patients randomly to one of the 3 dosages to be followed up. In the successive stages of giving the treatments and of recording of results, provide safeguards against the risk of introducing any bias. If these stages of experimentation have been properly carried out, the results obtained for the different groups can be subjected to statistical analysis and valid conclusions can be drawn from responses to the different treatments. This simple design has tremendous possibilities in medicine because of its extreme simplicity of application within the conditions of ordinary clinical work.

22.3.2 Two-way Classification

If the experimenter has knowledge regarding behaviour of his experimental material, he can stratify the patients included in the study according to one or more factors that may affect the responses. For example, in a survey of family expenditures on food, classification of the results by size of family and income level is obviously relevant. Such a method is applicable to surveys also and can be used to increase the precision in an experiment. If there are K treatments to be compared, he first arranges the experimental units in groups of K, often called 'replications'. The rule is that units assigned to the same replication should be as similar in responsiveness as possible.

Each treatment is then allocated by randomisation to one unit in each replication. This produces a two-way classification, since any observation is classified by the treatment which it received and the replication to which it belonged. Sometimes, this experimental plan is also called a 'randomised block'.

22.3.3 Latin Squares

In order to take care of correlation between patients belonging to the same family or to the same locality, etc., treatments may be arranged in a particular fashion. Suppose we have 4 treatments *A*, *B*, *C*, *D*, then a plan known as *Latin squares* uses 4 replications of each treatment as follows:

A	*B*	*C*	*D*
B	*C*	*D*	*A*
C	*D*	*A*	*B*
D	*A*	*B*	*C*

The Latin square is effective in controlling two sources of variation of which the investigator has predictive knowledge. In psychology and medicine, the human subject frequently comprises a replication of the experiment, receiving all the treatments in succession, with intervening intervals in which the effects of the previous treatment will have died away. However, a systematic effect of the order in which the treatments are given can often be detected. This is controlled by making the columns of the square represent the order while rows represent subjects. In animal nutrition, the effects of both litter and condition of the animal may be removed from the estimates of treatment means by the use of a Latin square. To construct a Latin square, write down a systematic arrangement of the letters and rearrange rows and columns at random. Then assign treatments at random to the letters.

22.3.4 Factorial Design

This design enables the researcher to investigate the effects of two or more factors simultaneously. For example, one factor (type of treatment) at two levels (conventional treatment or new treatment) can be studied with another factor (dosage) also at two levels (low dosage or normal dosage). A total of 120 cases available for the study can be distributed in a factorial design as found in Table 22.1.

Table 22.1 Allocation of 120 Cases in a Factorial Design (Two Factors Each at Two Levels)

	Treatment		
Dosage	Conventional	New	Total
Low	30	30	60
Normal	30	30	60
Total	60	60	120

This design indicates that 60 cases are available to compare the conventional and new treatment at the low dose level and another sixty cases are available for the same comparison at the normal dose level. If the difference in effectiveness between the conventional and the new treatment at low dosage level is the same as at the normal dosage level, there is no interaction between the two factors, viz., type of treatment and dosage.

This analysis brings out the existence of effects due to two factors, i.e., type of treatment as well as dosage (main effects) and also interaction between the two factors.

22.3.5 The Split-plot or Nested Design

It is often desirable to get precise information on one factor and on the interaction of this factor with a second, but to forego such precision on the second factor. For example, three sources of vitamin might be compared by trying them on three mates of the same litter, replicating the experiment on 20 litters. This would be a randomised block design with high precision, providing 38 degrees of freedom for error. Superimposed on this could be some experiment with the litters as units. Four types of housing could be tried, one litter to each type, thus allowing 5 replications with 12 degrees of freedom for error. The main treatments (housings) would not be compared as accurately as the sub-treatments (sources of vitamin) for two reasons; less replication is provided, and litter differences are included in the error for evaluating the housing effects. Nevertheless, some information about housing may be got at little extra expense, and any interaction between housing and vitamin will be accurately evaluated.

22.3.6 Analysis of Covariance

The analysis of covariance has numerous uses.

1. To increase precision in randomised experiments: In such applications the covariate x is a measurement, taken on each experimental unit before the treatments are applied, that predicts to some degree the final response y on the unit.

2. To adjust for sources of bias in observational studies: An investigator is studying the relation between obesity in workers and the physical activity required in their occupations. He has measures of obesity in samples of workers from each of a number of occupations. He has also recorded the age of each worker, and notices that there are differences between the mean ages of the workers in different occupations. If obesity is linearly related to age, differences found in obesity among different occupations may be due in part to these age differences. Consequently, it is necessary to adjust for a possible source of bias in comparison among occupations.

3. To throw light on the nature of treatment effect in randomised experiments: In an experiment on the effects of soil fumigants on nematodes, which attack some farm crops, significant differences between fumigants were found both in the number of nematode cysts x_{ij} and in the yields y_{ij} of the crop. This raises the question: Can the differences in yields be ascribed to the differences in numbers of nematodes? One way of examining this question is to see whether treatment differences in yields remain, or whether they shrink to insignificance, after adjusting for the regression of yields on nematode numbers.

4. To study regressions in multiple classifications: An investigator may be interested in studying the relation between per capita income (x) and per capita health expenditure (y) in various states in India. If he has the data for a long period (10 years), he may like to examine whether the relation is the same in different states, or whether it remains the same from year to year. Sometimes, the question is whether the relation is linear or not.

Example of analysis of covariance. The haemoglobin levels of randomly selected pregnant women in rural and urban areas were recorded for a study. There was an impression of linearity of regression of haemoglobin level on the age of woman. The data including the mean haemoglobin levels before and after adjusting for age in the two samples are shown in Table 22.2.

Table 22.2 Age and Haemoglobin Level (g%) of Pregnant Women in the Rural and Urban Areas

	Rural (n = 10)		Urban (n = 10)	
	Age (years) x	Hb (g%) y	Age (years) x	Hb (g%) y
	24	9.0	25	12.0
	26	9.2	30	16.0
	28	10.0	26	13.4
	29	11.0	25	12.8
	25	9.4	28	15.2
	24	8.5	27	14.6
	24	8.8	28	15.8
	26	10.0	26	13.6
	25	9.2	27	14.8
	27	9.4	27	15.8
Total	258	94.5	269	144.0
Mean	25.8	9.5	26.9	14.4
Adjusted mean		9.8		14.1

In this example, it was first necessary to examine whether the regressions of haemoglobin level on age are the same in the rural as in the urban areas. If the differences are statistically significant, then we need to determine why, and so on. If not significant, the data from the two areas can be combined for further examination of the relation between age and haemoglobin level.

22.4 DETERMINATION OF SAMPLE SIZE

The details are given in Chapter 23.

22.5 CLINICAL TRIALS

The clinical trial is a carefully and ethically designed experiment with the aim of answering some precisely framed question. It is broadly divided into 2 varieties — therapeutic and prophylactic. In its most rigorous form, it demands equivalent groups

of subjects concurrently treated in different ways. These groups are constructed by the random allocation of subjects to one or the other treatment; such an allocation may sometimes preferably be made within smaller homogeneous sub-groups comprising the total groups. In some instances, subjects may be used as their own controls, different treatments being applied to them in random order and the effects compared.

22.6 THERAPEUTIC TRIALS

22.6.1 General Principles

The following general principles must be observed in planning and conducting a good study:

1. Careful formulation of the questions for which decisive answers are required.

2. Decision on the most appropriate form of 'controls' preferably to study concurrently. The essence of a 'controlled trial' is that a *comparative study* must be made. If no satisfactory treatment (placebo) exists, the new treatment should be compared with a pharmacologically inert preparation (dummy) to check whether it has any appreciable effect on the disease. When a treatment is already in existence which appears to be of some value, the new form of treatment should normally be compared with the established one.

3. Selection of patients for inclusion in the trial before it is known as to which treatment group they will be included in. After admission to the trial, patients should be allotted at random to one or the other treatment group.

4. When possible, especially for subjective evaluations, trials should be so arranged that neither the physician nor the patient knows which treatment is being used—the so-called *Double-Blind System*. If only the subjects are masked, it is known as *Single-Blind*; and when the person analysing the data is also kept blind, it is called *Triple-Blind*.

5. Decision on the nature and method of making systematic and pertinent observations on the patients so that relevant data are available for analysis at the conclusion of the trials.

6. Decisions on the procedure for dealing with withdrawals and other deviations from the design that arise during the trials.

7. Decisions on data processing, statistical analysis and on drawing inferences.

22.6.2 Design and Random Allocation

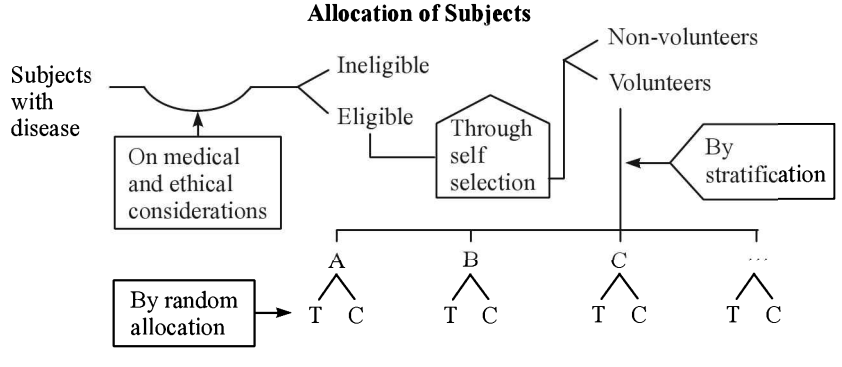
The aim is to allocate subjects to the treatment and control groups in such a way that the two groups are initially equivalent in all respects relevant to the inquiry. Individuals, it may be noted, are not necessarily equivalent; it is a group reaction that is under study. Sometimes, this allocation is made by putting patients as they present themselves alternatively into the treatment and control groups; but it has been found that we cannot always hope that a random allocation has been made that way.

By far the best way of allocating the patients is by randomisation, using when necessary the well-known table of random numbers. In certain trials, the clinician does not know to which group a particular patient belongs, and such unbiased assessments are also necessary in certain situations. If randomisation is faithfully followed, it offers three great advantages:

First, it ensures that our personal feelings or judgments applied consciously or unconsciously have not played any particular part in building up the various treatment groups; from that aspect, therefore, the groups are unbiased.

Second, it removes that very real danger, inherent in any allocation which is based upon personal judgment, that believing our judgments to be biased, we endeavour to allow for that bias and in doing so 'lean over backwards' and thus introduce a lack of balance from the other direction.

Third, having used such a random allocation one cannot be accused by critics of having set up personally biased groups for comparison. The process of selection and allocation of subjects to the different groups is shown in the following flow chart:



22.6.3 Mode of Administration of Therapy

In testing a new form of treatment, the decision on the mode of administration of therapy poses several crucial problems. Information is at first scanty. The testing may reveal the potentialities and perhaps some of the dangers or side-effects of the drug. We have to choose one of several possibilities. We can, for instance, choose one dose of a drug out of many, vary the interval of its administration, give it by different routes, for different lengths of time; and so on. Once decisions are taken about the therapy under trial, identical procedures should be followed for the control groups. Very often dummies, as we have noted, are used to preclude any differences between the psychological effects of treatment and lack of treatment, while placebos are used to produce a definite psychological effect. Sometimes, the standard form of therapy is used in the control group. As far as possible, the control group should have the form of therapy which is similar to the treated group.

22.6.4 Observations and Records

We have said that the numerical expression is the best form of expression as it permits comparative studies and as it is definite and generally gives an idea of the

degree of differences between two expressions. In the adoption of the measurement of effect of treatment, detailed planning must as usual play its part. Before the trial is set underway it must be laid down, for example, precisely when and how temperatures will be taken, when full clinical examinations will be made and what will be specifically recorded, how often and at what intervals X-rays will be taken, etc. Standard record forms must be drawn up and uniformity in completing them stressed. Unless these rules and regulations are followed by the clinicians in charge of the trial, many serious difficulties may arise in the final analysis of the results. In fact, *every* departure from the design of the experiment lowers its efficiency to some extent; too many departures may wholly nullify it. The individual may often think 'it won't matter if I do this (or don't do that) just once'; he forgets that many other individuals have the same idea. It is fundamental, too, that the same care in measurement and recording be given to both groups. The fact that some are specially treated and some are not, is wholly irrelevant. Unless the reactions of the two groups are equally noted and recorded, any comparisons of them must clearly break down.

22.6.5 Deviations from Design

Situations leading to deviations from the procedures laid down are not uncommon during the execution of the plans. Such deviations can occur at all stages, even after the patients have been included in the trial and therapy started. The following are some of the common situations requiring or resulting in removal of the patient from the study for special care and treatment:

1. Serious side-effects
2. Deterioration of condition
3. Death due to other causes
4. Lack of cooperation by the patient
5. Lost from observation (moved away, left, etc.).

These deviations must be considered in the statistical analysis. Inferences must be made taking account of these deviations.

22.7 PROPHYLACTIC TRIALS

The basic principles for a prophylactic trial are similar to those for therapeutic trials. Two groups are established alike, as far as possible, except that one is given the prophylactic and the other, the control, is not. These groups are then observed over the same period to see if the protected group has suffered a lower incidence of the particular disease than the unprotected.

As in therapeutic trials, ethical problems have to be resolved. Since the essence of a trial is the denial of a prophylactic to the control group, the first question is to decide whether a trial is justified at all. Unfortunately, so many prophylactics have been established on inadequate evidence that it is not hard to find justification for a field trial on even the most widely accepted of preventive measures.

One big difference between prophylactic trials and therapeutic trials is the scale of the trials. The proportion of the population likely to be attacked by a

specific disease is usually small. Hence, unless the two groups in the trial are large there will not be sufficient cases in the unprotected groups to provide a standard against which the incidence in the protected group be measured.

The approximate scale of a trial can be determined beforehand from estimates of the likely degree of protection afforded and of the incidence of the condition in unprotected persons. For a trial using equal numbers in protected and control groups, it can be calculated that 50 cases of the illness should occur in the controls to be reasonably confident of demonstrating significant benefit when the prophylactics on trial is expected to halve the risk. If the prophylactic is more effective less than 50 cases may be sufficient among controls and vice versa. The size of the trial can then be any appropriate combination of numbers admitted, and duration of observation likely to yield this number of cases in the unprotected group.

The aspects of controls, design and analysis are similar to those in therapeutic trials. A large-scale field trial of a prophylactic is a laborious and expensive undertaking. Such trials have been carried out particularly in infectious diseases because immunisation procedures are highly developed and as a rule, simple. Any prophylactic which requires constant use in the diet would pose special problems of allocation and administration, and would be even more difficult to assess by this method.

22.8 ETHICAL CONSIDERATIONS

Because of the random allocation of subjects to the experimental and control groups, however, certain ethical questions naturally arise. These ethical issues have become of increasing concern to investigators, institutions, and sponsoring agencies. A detailed consideration of these issues is beyond the scope of this book. The reader is referred to several excellent discussions of this issue. It is clear that before a trial can be carried out, the consent of the subjects to participate must be obtained; it is important that the subjects be informed that they may be assigned to either the experimental or control group. If the subjects, after being provided with this information, still decide to participate in the study, then they are said to have given their 'informed consent'. Such consent should be obtained in writing in accordance with regulations of governmental agencies (in the U.S.), or of the institution where the trial is to be conducted.

22.9 COMMUNITY TRIALS

Experiments that involve communities as a whole are known as *community trials*, whether they are conducted on animals or humans. It should be emphasised that in a community trial, the group as a whole is collectively studied, while in a clinical trial, it is the individual within a group (the experimental or control group) that is studied. Human community trials have been conducted when particular circumstances required that the community as a whole serve as the experimental unit with regard to testing a specific etiological hypothesis or preventive procedure. The very nature of such trials does not usually allow randomisation. Classical community trials were conducted by Goldberger and his colleagues in studying the dietary etiology and prevention of pellagra. In their studies, the human communities were closed ones—

two orphanages and two wards of a state mental institution. An interesting use of community trials was made by Watt and Lindsay in their studies of diarrhoeal diseases. One issue was to determine the effect of fly control on both the attack and death rates from diarrhoeal diseases including dysentery. Community trials are now being conducted with regard to diseases of current epidemiologic interest. A recent example is the attempt to determine whether community health education can decrease the risk of cardiovascular disease. In 1972, such a trial (termed *field experiment* by the investigators) was initiated in three California communities in an attempt to modify three risk factors (cigarette smoking, elevated plasma cholesterol levels, and high blood pressure) by community education. In two of these communities, extensive mass-media campaigns were conducted over a period of two years. In one of these, face-to-face counselling was also provided for a small sub-group of high risk individuals. The third community served as control.

In each community, people were interviewed and examined before the educational campaign began. One or two years later, their knowledge and behaviour with respect to diet and smoking were assessed. Physiological indicators of risk—blood pressure, relative weight, and plasma cholesterol were also measured. The baseline values of the risk factors were remarkably uniform in these communities. The mass media and the combination of mass media plus face-to-face instruction had significant positive effects on all factors except relative weight after two years of campaigning. The comparison of the treated and conrol communities indicated an estimated decrease in risk of developing cardiovascular disease by approximately 25%. These results were sufficiently encouraging for the investigators to extend these community trials to five communities.

22.10 REPEATED MEASUREMENT STUDIES

Attention has been restricted so far to the comparison of treatments as only a one-time outcome or response. Often, however, response to a treatment is measured several times during or after its administration, the intention being to compare treatments with respect to the trends in their effects over time as well as with respect to their mean levels of response. In spite of long-standing criticisms, the analysis of the time dimension in such repeated measurements studies continues to be performed by a straightforward but generally invalid application of the analysis of variance.

QUESTIONS

1. Explain the general principles of interventional/experimental studies.
2. Explain three commonly used experimental designs.
3. Write brief notes on:
   (a) Types of controls
   (b) Valid estimate of experimental error
   (c) Non-sampling error
   (d) Clinical trials
   (e) Deviation from designs.

CHAPTER

23 Sample Size Determination

23.1 INTRODUCTION

An important item considered while planning any investigation is the number of cases required. It is unethical to include too many cases in an experiment. Including too many cases in a study or an experiment may also increase the cost of research. On the other hand, studying too small a number of cases may result in findings that may not be significant. There is no simple answer for this frequently asked question on sample size needed. Taking into account the nature of the measurement of the observation, the design of the investigation and the variability in the characteristic in the population, the number of cases to be studied, the sample size, is calculated.

23.2 SPECIFICATIONS NEEDED TO CALCULATE SAMPLE SIZE

To estimate the sample size needed for a study or an experiment the statistician or investigator should know some important aspects about the study. They are:

1. The precision with which the parameters should be estimated or differences identified; or the minimum difference between parameters which is of sufficient clinical importance.

2. The degree of safeguard allowed against the risk of drawing wrong conclusions (confidence level). This implies controlling Type I (α) and Type II error (β) as explained in Chapter 11. Often only Type I error is considered.

3. An approximate idea of the population value, or the degree of variability.

In general, the more the variability, the precision and confidence level, the larger the number of cases to be studied in specific circumstances.

23.3 EXAMPLES OF SAMPLE SIZE ESTIMATION

23.3.1 Estimating a Proportion

Example 1: Suppose the attack rate of a disease in a locality is 5% (from past records):

(i) The investigator allows an error of 20 per cent of the attack rate on either side (i.e., the investigator wishes to get his estimated value between 4 per cent and 6 per cent).

(ii) The investigator wishes to be correct in his estimate (any value between 4 and 6) in 95 out of 100 attempts (safety factor 5 per cent; i.e., plus or minus standard error attachment is set at 2 standard error distances, at which the chances are 20 to 1 against exceeding that range of error).

Let p be the attack rate in per cent and q the non-attack rate ($p + q = 100\%$). The standard error of p is given by the formula

$$\sqrt{\frac{pq}{n}}$$

where n is the sample size. In the present example, $p = 5$, $q = 95$ and n = the sample size to be found out based on the restrictions. From the restrictions (i) and (ii), we get

$$2\sqrt{\frac{pq}{n}} = 20\% \text{ of } p$$

i.e.

$$2\sqrt{\frac{5 \times 95}{n}} = \frac{20}{100} \times 5$$

i.e.

$$2\sqrt{\frac{475}{n}} = 1$$

Squaring both sides, we get

$$4 \times \frac{475}{n} = 1, \quad \text{i.e.} \quad \frac{1900}{n} = 1$$

or

$$n = 1900$$

This means that the sample size required in the study is equal to 1900. If the investigator feels that the drop-outs can be 25%, then the sample size should be 1900 + 632 or approximately 2532.

Clue for quick work

$$\text{Sample size } n = \frac{4pq}{d^2}$$

where $p = 5$, $q = 95$ and d = difference tolerated = 20% of p (= 1) and $4 = 2^2$ (2 is the approximation of 1.96 which is the 95% confidence limit of the normal distribution.

$$n = \frac{4 \times 5 \times 95}{1^2} = 1900$$

23.3.2 Testing Difference Between Two Proportions

Example 2: *Sample size in testing difference in proportions.* Suppose an investigator wishes to detect a difference d between two proportions p_1 and p_2 at a specified significance level α. To determine if the difference d falls in the critical region, a critical ratio Z_α is calculated, which is given by

$$Z_\alpha = \text{Critical ratio} = \frac{\text{Difference in proportions}}{\text{Standard error of proportion}}$$

i.e.

$$Z_\alpha = \frac{d}{\sqrt{2pq/n}} \tag{1}$$

where each group has the sample size n, p is the pooled value of p_1 and p_2 given by

$$p = \frac{p_1 + p_2}{2}$$

Then,

$$q = 100 - p,$$

$$d = p_1 - p_2$$

To determine the sample size n, the critical ratio is reconverted and presented in the following form:

$$n = \frac{Z_\alpha^2 pq \times 2}{d^2} \tag{2}$$

Similarly, if α and β are specified, then

$$\text{Sample size } n = \frac{(Z_\alpha - Z_{1-\beta})^2\ pq \times 2}{d^2} \quad \text{or} \quad n = \frac{2(Z_\alpha + Z_\beta)^2\ pq}{d^2} \tag{3}$$

Reference to any table of the Normal distribution will show that, for the significance level $\alpha = 0.05$, the corresponding two tailed deviate $Z = 1.96$; for $\alpha = 0.01$, $Z = 2.575$; and for $\alpha = 0.001$, $Z = 3.29$.

Similarly, for the various levels of power $(1 - \beta)$, the corresponding upper two-tailed deviate in Normal distribution is given in Appendix 5.

An investigator wants to assess the efficacy of a new drug or vaccine. If the investigator feels that the vaccine will be of value only if it produces a protective effect of 50% of the control (For example, if 20 cases of the disease in question are expected among 100 controls, then only 10 cases are expected among the vaccinated), then the required sample size in the study can be determined by considering restrictions (i) and (ii) in Example 1 along with the above restriction by the following formula.

$$\frac{|\ p_1 - p_2\ |}{\sqrt{pq}\sqrt{\frac{2}{n}}} = Z_\alpha \tag{4}$$

where p_1 is the attack rate in the control group and p_2 is the attack rate in the treated group, p is the pooled value, n is the sample size, and Z_α the critical value corresponding to the Normal distribution.

$$q = 100 - p$$

If $p_1 - p_2 = d$, then,

$$n = \frac{2Z_\alpha^2 pq}{d^2} \tag{5}$$

Substituting the values $p_1 = 20$, $p_2 = 10$, we get

$$\frac{20 - 10}{\sqrt{\frac{20 \times 80}{n} + \frac{10 \times 90}{n}}} = 2$$

i.e.

$$\frac{10}{\sqrt{\frac{1600}{n} + \frac{900}{n}}} = 2$$

or

$$\frac{10}{\sqrt{\frac{1}{n} \times (2500)}} = 2$$

Squaring both sides, we get

$$\frac{100}{1/n \times (2500)} = 4$$

i.e.

$$\frac{n}{25} = 4 \quad \text{or} \quad n = 100$$

This means that the size of the sample required in the study is equal to 100 in each of the two groups.

Clue for quick work (without considering the power of the test)

$$\text{Sample size } n = \frac{2Z_\alpha^2 pq}{d^2} \text{ for each group or each arm}$$

where $p = 15 = (p_1 + p_2)/2$, $q = 85$, d = difference tolerated = 10, i.e. $(p_1 - p_2)$ and $= Z_\alpha$ is 1.96 which is the critical value at 5% level of the normal distribution rounded off as 2.

$$n = \frac{2 \times 2^2 \times 15 \times 85}{10^2} = 100$$

Clue for quick work (considering the power of the test)

$$\text{Sample size } n = \frac{2(Z_\alpha + Z_\beta)^2 pq}{d^2} \text{ for each group or arm}$$

If the investigator additionally states that a 80% power $(1-\beta)$ of the test is needed, then $Z_\alpha = 1.96$, $Z_\beta = 0.842$ (from Appendix 5), $p = 15$ and $q = 85$.

$$\text{Sample size } n = \frac{2(1.96 + 0.842)^2 \times 15 \times 85}{10^2} = 200 \text{ for each group or arm}$$

23.3.3 Testing Difference between Two Proportions—Additional Example

The investigator hypothesizes that the percentage of children developing emotional problems among those beginning school with both parents at home is $p_1 = 5$ and that the percentage of children beginning school with at least one parent absent is approximately $p_2 = 15$. The significance level α is set at 0.05 and power $1 - \beta$ at 0.90. Using the formula 3, the sample size needed in each group is

$$n = \frac{\left(1.96 - (-1.282)\right)^2 (10 \times 90) \times 2}{10^2} = 190$$

If the investigator drops the power $(1 - \beta)$ and considers only significance level (α), then using formula 2, the sample size needed in each groups is

$$n = \frac{(1.96)^2 \times (10 \times 90) \times 2}{10^2} = 69$$

Clue for quick work (considering the power of the test)

$$\text{Sample size } n = \frac{2(Z_\alpha + Z_\beta)^2 pq}{d^2} \text{ for each group or arm}$$

The investigator wants a 90% power $(1 - \beta)$ of the test, so $Z_\alpha = 1.96$, $Z_\beta = 1.282$ (from Appendix 5), $p = 10$ and $q = 90$.

$$\text{Sample size } n = \frac{2(1.96 + 1.282)^2 \times 10 \times 90}{10^2} = 190 \text{ for each group or arm}$$

23.3.4 Estimating a Mean

Sometimes, the investigator is interested in estimating a population mean. For determining the sample size, the investigator needs to know the variation in the variable under consideration and the expected difference between the sample mean and the population that can be tolerated. If σ is the standard deviation, that represents the variation and d is the difference tolerated, then with 95% confidence interval the sample size needed is given by

$$\text{Sample size } n = \frac{Z_\alpha^2 \sigma^2}{d^2}$$

Example 3: An investigator wanted to estimate the mean birth weight in a population, the standard deviation from a sample was known to be 450 gm. The expected difference between the population mean and the sample mean was considered to be 30 gm. Then the sample size n needed is

$$\text{Sample size } n = \frac{Z_\alpha^2 \sigma^2}{d^2} = \frac{1.96^2 \times 450^2}{30^2} = 864$$

The sample size needed is 864.

23.3.5 Testing Difference Between Two Means

Example 4: *Sample size in testing difference in means.* Suppose an investigator wishes to detect a difference d between two means at a specified significance level α. To determine if the difference d falls in the critical region, a critical ratio, Z_α, is calculated, which is given by

$$\text{Critical ratio, } Z_\alpha = \frac{\text{Difference in means}}{\text{Standard error of difference in means}}$$

i.e.

$$Z_\alpha = \frac{d}{\sigma\sqrt{(2/n)}} \tag{6}$$

where each group has the sample size n, and σ is the standard deviation in the population.

When σ is not available, the sample standard deviation, s can be used instead of σ, where

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

By referring to a table of a normal (or similar probability distribution), the value of Z is determined.

To determine the sample size n, this critical ratio is reconverted and presented in the following forms:

$$n = 2 \times \left(\frac{Z_\alpha (\sigma)}{d}\right)^2 \tag{7}$$

If, in addition to α we consider the specified power $1 - \beta$, the sample size is determined by the formula:

$$\text{Sample size } n = \frac{2(Z_\alpha - Z_{1-\beta})^2 \cdot \sigma^2}{d^2} \quad \text{or} \quad n = \frac{2(Z_\alpha + Z_\beta)^2 \sigma^2}{d^2} \tag{8}$$

In the study of serum uric acid levels, the mean level in the diseased group is 5.4 mg/100 ml with standard deviation 1.1 mg/100 ml. Using a new drug, the investigator expects the clinical difference between the untreated and the treated group to be at least 0.2 mg/100 ml. The investigator wishes to conduct the experiment with the significance level $\alpha = 0.05$ and power $(1 - \beta) = 0.90$. Then the appropriate sample size needed for each group, using formula 8 is

$$n = 2 \times \left(\frac{(1.96 - (-1.282)) \times (1.1)}{0.2}\right)^2 = 636$$

Supposing the investigator specifies the significance level only, then the sample size needed in each group using formula 7 is

$$n = 2 \times \left(\frac{(1.96)\,(1.1)}{0.2}\right)^2 = 232$$

Clue for quick work (considering the power of the test)

$$\text{Sample size } n = \frac{2(Z_\alpha + Z_\beta)^2 \sigma^2}{d^2} \text{ for each group or arm}$$

The investigator wants a 90% power ($1 - \beta$) of the test, so $Z_\alpha = 1.96$, $Z_\beta = 1.282$ (from Appendix 5), $\sigma^2 = 1.1^2$, d = difference between the means expected = 0.2.

$$\text{Sample size } n = \frac{2(1.96 + 1.282)^2 \times 1.1^2}{0.2^2} = 636 \text{ for each group or arm}$$

23.4 USEFUL HINTS

It is impossible to estimate sample size without specifying the items mentioned as required in Section 23.2. These can be obtained through relevant published literature or through a pilot study. Inappropriate specification of these may result in inaccurate value of the size of the sample.

The procedure for sample size estimation was given only for a few common situations encountered by an investigator. For calculation of sample size to estimate correlation coefficient, odds ratio, relative risk, testing between several means, multivariate analysis and so on, the reader may refer to books mentioned in the bibliography or other advanced books.

QUESTIONS

1. An investigator wants to find out the proportion of diabetic persons developing ulcer on their feet. Help him to determine the sample size needed for his study. Literature search showed that 20.5% of diabetic persons developed such ulcers. Fix Type I error at 5% and consider 95% confidence interval.
2. In conventional treatment of warts, 40% reach complete resolution over a period of 8 weeks. A researcher wants to try a new method which he believes will attain 60% complete resolution at 8 weeks of treatment. Determine the sample size needed for the study. Clearly state the assumptions and conditions imposed.
3. An educationalist was interested in knowing the extent of knowledge about leprosy acquired by the students graduating from college. Previous studies pointed out that a batch of 60 students had a knowledge score on leprosy as mean 27.7 and standard deviation 9.2. Estimate the sample size needed for this study. State the conditions and assumptions involved while estimating it.
4. Among village children the mean serum protein level was known as 7.2 gram per cent with a standard deviation of 2.5%. A company introduced a new high protein food which is supposed to increase the level to 9%. If the high protein food is to be tested with a control group, how many subjects are needed in each group?

CHAPTER

24 Observational Studies

24.1 INTRODUCTION

Medical studies can be broadly classified as *Experimental* or *Non-experimental* (observational). As described in Chapter 22, the investigator actively manipulates the explanatory or independent variables, assigning subjects to experimental and control groups randomly, and then observes the outcomes.

In observational studies, individuals 'self-select' into comparison groups, based on differing levels of exposure to the risk factor as they take place in nature, with no active intervention on the part of the investigator. Observational studies are often concerned with identifying factors which could be regarded as causal in the development of a disease or condition (risk factors), or with determining predictors of outcome or prognosis (without experimental manipulation) in patients who already have a specific condition. Depending on the time frame, observational studies are either cross-sectional or longitudinal, the latter further classified as prospective (cohort or follow-up) study or retrospective (case-control) study.

Experimental studies are more efficient in proving causal relationships, but practical and ethical considerations limit their use; observational studies, on the other hand are more feasible to evaluate disease-risk factor associations, despite their potential for greater bias.

The role played by the observational method of enquiry in the acquisition of scientific knowledge can never be overemphasised. In social sciences, for example, the investigator cannot experiment, but has to deal with circumstances operating entirely beyond his control. Often, observational studies prove to be a valuable indicator of a problem providing the starting point for further research. Many medical breakthroughs were achieved in the past without knowledge of cause or even of diagnosis of a disease. Identification of the conditions necessary to the appearance of the phenomenon provided an important aid in the control and prevention of many diseases. For example, Sir John Snow's finding that cholera was spread by polluted water, and Goldberger's demonstration that pellagra was a deficiency disease came to light several years before the cause was actually discovered; yet they were important discoveries.

In an experimental approach, we can design a study with matching groups thus controlling several disturbing factors. In observational studies, we have to observe the events as they occur. Many correlations which are observed in such studies may not bring us anywhere near the cause-and-effect phenomena. It will be necessary to examine and try to elucidate the multiplicity of factors influencing the

results. Tests of significance play a less important role in observational studies. The more valuable aspect is the examination of the data for possible selection (or bias) and the consideration of all associated factors. The method of data collection and the design of a medical study are intimately related to the purpose for which the data are collected, and if the study design is inappropriate to, or inadequate for, the purpose of the study, no amount of statistical manipulation or analysis can compensate for or redeem it.

24.2 CROSS-SECTIONAL STUDIES

Essentially, a cross-sectional study does not take into account the temporal relationship between the factors studied, and usually involves the examination of a cross-section of a particular population at one point of time. It gives a 'snapshot' view. Cross-sectional studies may be based on a random sample from a defined population, or on a presented sample of patients with a particular condition.

If a study is hospital-based, one must be careful not to over-generalise the results, and be wary of the selection biases which lead to hospital admission.

Cross-sectional studies may be employed to study associations between different diseases, but they cannot often determine which disease might have occurred first. In terms of explaining disease aetiology (cause of disease), they provide only limited information. The prevalence of a disease can be broadly defined as the proportion of persons in a population who have the disease in question at a particular point of time (point prevalence), or over a short time period (period prevalence). Cross-sectional studies are ideally suited to study prevalence. Such studies also have a use in estimating health needs or health attitudes and behaviour in a community, and the results of such studies may help in planning services or appropriate health education programmes.

24.3 PROSPECTIVE STUDIES

A prospective study (also known as a cohort study, incidence study, or follow-up study) is of two types:

24.3.1 Aetiological

Here, one starts with an initially healthy (but susceptible) cohort of individuals, exposed to different levels of a suspected risk factor (or "cause"), who are followed in time to determine the subsequent incidence of the event (or "effect") in question in each of the groups under study and compared. Sometimes, the cohort of exposed and nonexposed subjects is assembled from past records and followed up to determine present incidence of the disease. This is known as a historical cohort or nonconcurrent cohort study. As in the prospective cohort study, comparison groups are defined on the basis of exposure to the suspected risk factor.

24.3.2 Prognostic

Here, patients with a specific disease are followed up to determine which factors relate to further morbidity (illness) and/or mortality. Studies of prognosis are usually

based on a presented sample of patients with a disease, and thus patients do not enter into observation at the same point in calendar time, but often over an extended period. The follow-up may require periodic examination of the individuals studied, or may just be based on the notification of the date of occurrence of any specific event (complication or death).

A prospective study resembles a planned experiment in several ways. It is, however, different from a planned experiment in that the occurrence of the suspected cause is determined by nature, and not by the intentional interference of the investigator.

For a successful analysis of the data collected in a prospective study, the following two requirements are most important:

1. Sufficient information about the people in the various groups is necessary on factors which are known to have effect on the responses under study, in order to make allowance for their effect while analysing the results. For example, age and sex influence responses in several studies.

2. The follow-up must be complete; at least very nearly so. If the follow-up is incomplete to a large extent, it will affect the validity of the conclusions to be drawn from the results.

When the data are in the form of counts, the strength of the correlation between the risk factor and the disease may be expressed as the relative risk (RR) defined as follows:

$$RR = \frac{\text{Incidence of disease among Exposed}}{\text{Incidence of disease among Nonexposed}}$$

If these results are presented in a 2 × 2 table as shown below, then, $RR = ad/bc$.

	Outcome	
Exposure	Present	Absent
Present	*a*	*b*
Absent	*c*	*d*

24.4 RETROSPECTIVE STUDIES

From most diseases, the occurrence of new cases is a relatively rare event. Thus, in a cohort study, a large number of study subjects may have to be examined for exposure status and followed up for a long time to obtain a sufficient number of cases. Such a study is often neither practical nor feasible. The problem can be avoided by using exposure information from a sample of the study population rather than including the entire study population. Retrospective studies are based on these principles.

In a prospective study, two or more groups which are distinguished with reference to the supposed 'cause' are studied for the extent of the 'effect' in each group. However, in a retrospective study (also known as case-control, or case-referent study), the groups for study are chosen with reference to the 'effect'.

Individuals not exhibiting such effect (control group) are chosen for comparative study. Conclusions are based on the extent of the supposed 'cause' in the groups under study. Thus, while a prospective study goes forward from a risk factor to the development of the end-point of interest, the retrospective study starts with the end-point and moves backwards in time to the identification of risk factors.

Let us consider, for illustration, the study of the effect of caffeine on bladder cancer. In the prospective study, the coffee drinkers based on their level of consumption, and non-coffee drinkers will constitute the groups for study. The incidence of bladder cancer in these groups will be recorded and conclusions drawn from the results. On the other hand, in a retrospective study, bladder cancer cases, suitably chosen, will constitute the other group. Inference will be made based on the comparison of the proportion of coffee drinking in both the groups.

In the retrospective study, an attempt is made to compare the frequency of the suspected cause in the affected group with the non-affected group in the population from which the patients come. Instead of true relative risk, only an estimated RR known as Odds Ratio (OR) is computed as follows:

$$\text{Odds Ratio} = \frac{\text{Odds that exposed individual will have disease}}{\text{Odds that nonexposed individual will have disease}}$$

$$= \frac{a/c}{b/d} = \frac{ad}{bc}$$

This is sometimes known as risk-odds ratio. On the other hand, the exposure-odds ratio is defined as

$$\frac{\text{Odds in favour of exposure among cases}}{\text{Odds in favour of exposure among non-cases}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

is the same as before.

The determination of the population from which the patients come is a crucial step in the design of a retrospective study. Also, the choice of a suitable control group as a representative sample of the above population is also very important. Any mistake committed at this stage will introduce bias into the investigation, resulting in artificial relationships. The logical validity of the results can be questioned.

The following are some methods of choosing the control group:

(i) Random sampling. This may not be often feasible. Also, this may not be appropriate when the disease under study affects certain specific categories of persons only. For example, males above thirty years may be affected by a particular disease.

(ii) Technique of matching. For every patient, a control is chosen who is of the same age, sex, social conditions etc. The object of matching is to eliminate, between the study and control groups, differences in factors likely to have relationship with the responses under study. In this process, there is the danger that the factor which the enquiry was designed to seek, may also be eliminated.

(iii) Patients with other diseases. Patients having other diseases are chosen as controls. It is better to use patients with many different conditions. Greater confidence can be attached to the findings when it is possible to construct two or more control groups, each group composed of patients with different conditions, and when the results obtained in the different control groups are the same.

Interpretation of results from retrospective studies may be extremely difficult. The association which is observed may reflect a real relationship, or it may also be an artefact produced by the method of enquiry—data collection and choice of the control group. In fact, it is impossible to exclude all bias from any retrospective enquiry. What has to be sought is the most reasonable interpretation of results, considering the various factors operating, and assessment of the plausibility of the various explanations.

The results of a retrospective study may point to the existence of a real association between a disease and some aetiological factor, but they can never prove cause and effect. In the case of an association between A and B, (i) A may cause B; (ii) B may cause A; or (iii) both A and B may be caused by a common factor. The time relations may enable us to exclude either (i) or (ii); but the third possibility can never be easily excluded by retrospective methods alone. Prospective studies and/or experiments will have to be undertaken to provide further evidence when this may not be possible. The investigation will have to be repeated under different conditions and with different techniques.

24.5 SOURCES OF BIAS IN OBSERVATIONAL STUDIES

These could be generally discussed under confounding, selection and information bias.

1. *Confounders* are variables, which compete with the hypothesised risk factor as explanations for the observed response. Or, by virtue of their association with both the suspected risk factor and the outcome in question, they create an artificial relationship between the two. Alternatively, confounding may result when two or more variables "travel together" e.g., smoking and alcohol consumption.

2. *Selection bias* occurs when an investigator inadvertently assigns subjects to comparison groups such that they differ with respect to extraneous factors. Migration bias, another form of selection bias, may occur when individuals drop out of the study or move from one comparison group to another.

3. *Information bias*, results from the systematic differences in the measurement of the response variables among the comparison groups. This happens when knowledge of a particular study subject is exposed to the suspected risk factor and influences both the outcome and the rigour of the diagnostic process. Or, in case-control studies, knowledge of a subject's disease status affects the diligence with which exposure information is collected. Diseased individuals (i.e., cases) and their families may search their memories more industriously for "causes" (i.e., prior exposure to the suspected risk factor) than controls would. As a result, the cases may not only remember exposure more accurately than controls, but also systematically overestimate or underestimate their exposure, resulting in recall bias.

Further, another bias due to sampling can lead to systematic differences between the study population and the target population that pose a threat to external validity.

Case-control studies based on prevalent rather than incident cases are susceptible to sampling bias because they exclude patients who die, or recover rapidly, as well as those with mild symptoms whose illness goes undetected. These "survivors" are unlikely to be representative of the total patient population.

Case-control studies based on hospitals rather than the population, are prone to sampling bias, if hospitalised patients are generally more ill or "selective". Similarly, if the probability of attending hospitals varies, another type of sampling bias known as Berkson's bias can occur leading to false risk estimates.

Sampling bias is also common due to nonresponse and when dropouts differ systematically from individuals who remain in the comparison groups.

24.6 CONTROLLING FOR BIAS

The bias may be controlled by using the following techniques:

(i) Restriction. The effect of a potentially confounding variable may be minimised by restricting study participants to those falling within a narrow range of values for the variable. Obviously, the relationship between the restricted factor and the disease cannot be studied.

(ii) Matching. Pairs of subjects may be matched with respect to values for a potentially confounding variable (matching variable). Identification of matching subjects becomes progressively demanding as the number of matching variables increases. Overmatching may result when subjects are matched on a variable that is related to exposure, artificially equalising significant differences between the comparison groups.

(iii) Stratification. By grouping study subjects into subsets separately, the effect of a confounding variable can be eliminated.

(iv) Standardisation of rates. Specific rates for a potentially confounding variable can be statistically applied to the overall study population or to a standard population (e.g., the 1991 Census population of India) to obtain adjusted (standardised) rates.

(v) Multivariate methods. Regression procedures, such as logistic regression, Cox's proportional hazards regression, and analysis of covariance, can be used to adjust values of the dependent variable for the influence of one or more confounding independent variables (covariates).

24.7 CAUSALITY

Causal associations are always sought after to guide prevention, diagnosis, and treatment. Proof of cause, however, is neither simple, straightforward, nor irrefutable.

While experimental studies provide the strongest proof, the following criteria postulated by Bradford Hill, known as Hill's criteria can be used to establish causality:

1. Study design
2. Strength of association
3. Consistency
4. Correct temporal relationship
5. Dose-response relationship
6. Plausibility
7. Specificity
8. Analogy.

24.8 RELATIVE ADVANTAGES AND DISADVANTAGES OF THE PROSPECTIVE AND RETROSPECTIVE METHODS

The cohort design might seem to be a more appealing choice than the case-control design. It has a simple, logical structure that leads to measurements of disease incidence for the exposed and unexposed groups, or each category of exposure if several are used. This design permits absolute as well as relative comparisons of disease incidence among the exposed and the unexposed. The case control design provides rather less information in that usually, only the relative risk can be estimated. Absolute measures of incidence can also be made in case-control studies, but only if the sampling fraction for the controls or the base-line incidence rate for the total study population is known.

In the prospective studies, there is less chance of introduction of bias into the results. In particular, the data relating to the 'cause' cannot be biased by knowledge of the results, because they are recorded before the effect is known. The extent of risk attributable to the factor under study can be measured and expressed in terms of incidence rate. However, prospective studies are often extremely difficult to organise. A long period of observation may be necessary before definite results are obtained. They are, therefore, generally unsuitable for use in a pilot study or when a large number of factors of rather doubtful significance have to be investigated. It is, perhaps, of greatest value when a precise hypothesis has already been formulated and it is desirable to obtain clear evidence to support or refuse it.

Retrospective studies have the advantage that patients who are readily available with a common disease are examined. Results can be obtained quickly and with relatively little expenditure and effort. When a disease is rare, a prospective study may be impossible because the required minimum number of cases will not be available in a population of manageable size. However, as mentioned before, retrospective studies can only help to formulate hypotheses of association but can never prove cause-and-effect.

Perhaps the most frequently mentioned disadvantage of the case-control design is the difficulty in selecting a satisfactory control group, with the consequent problem of introduction of systematic error in the study (by the selection of controls). This

problem is an issue only for non-random selection of controls. When the control group is defined as a random sample of the study population, control selection is a simple technical procedure and introduces no systematic error beyond what would be present in a cohort study using the entire study population.

In case-control studies involving contact with study subjects or their relatives, questions about previous exposure are usually answered by the study subjects only after the cases have fallen ill. The cases, therefore, tend to spend more time on recalling past exposures and the causes of their disease, while the controls have no motivation to do so. This difference between cases and controls in the accuracy and completeness of exposure information can introduce a "recall bias" into the study. A similar bias may be seen in cohort studies in which, for some reason, exposure information is obtained only after the cases have been identified.

Both retrospective and prospective studies have their places in epidemiologic and clinical research. With appropriate techniques, both types of studies will help in acquiring knowledge of great value.

24.9 CONFIDENCE INTERVALS (CI) OF ODDS RATIO (OR)

As in the case of any descriptive statistics, the Odds Ratio (OR) is also subject to sampling error. The SE of OR is approximately given by the formula

$$\text{OR}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

using the standard notation in a 2 × 2 table.

More exactly, Woolf's method can be used where the natural logarithm (ln) of OR is considered to find SE of OR.

$$\text{The SE of ln OR} = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

The 95% CI and ln OR can be obtained by ln OR ± 1.96 SE of ln OR.

The upper and lower limits of the 95% confidence interval estimate of the OR are then derived as the antilogs of the upper and lower limit of the 95% confidence interval for ln OR.

A confidence interval estimate of OR can be used to test the significance of an association between a disease and a suspected risk factor. If 1 is included in the calculated confidence interval, then H_0 is not rejected. If the calculated confidence interval does not include 1, H_0 is rejected.

24.10 MEASURES OF POTENTIAL IMPACT

Once a cause-and-effect relationship between the disease or other health-related outcome, and exposure to the risk factor have been established, a few other statistics can be computed.

(i) Attributable risk (AR). This is also known as *risk difference*, excess risk, or rate difference. AR defines the excess risk of disease that can be ascribed to exposure

to the risk factor, over and above that experienced by people who are not exposed. It thus provides an estimate of the number of cases of the disease that might be prevented if exposures to the risk factor were eliminated and is useful for determining the magnitude of the public health problem posed by such exposures. AR is calculated using the formula:

AR = Incidence of disease among those exposed – Incidence of disease among those not exposed.

(ii) Population attributable risk (PAR). This is a measure of the excess risk of disease in a population that can be solely attributed to the risk factor. It is obtained by multiplying the AR by the prevalence of exposure to the risk factor in the population.

(iii) Attributable fraction in exposed (AF_E). This is also known as "attributable risk percent in the exposed", "aetiologic fraction in exposed", or "simple aetiologic fraction". It is the difference in the incidence of disease between people who are exposed to the risk factor and those who are not exposed, divided by the rate among those exposed. In simple terms,

$$AF_E = \frac{RR - 1}{RR}$$

(iv) Attributable fraction in total population (AF_T). This is also known as "population attributable risk percent", "*total aetiologic fraction*", or "population attributable fraction" and is the proportion of the total risk of disease in a population that can be attributed to exposure to a suspected risk factor. It is calculated as

$$AF_T = \frac{P(E+)(RR - 1)}{1 + P(E+)(RR - 1)}$$

where $P(E+)$ is the proportion of people in the population who have been exposed to the risk factor.

(v) Number needed to treat (NNT)*, **. A recently introduced measure, 'number needed to treat' is a meaningful way of expressing the benefit of an active treatment over a control. It can be used either for summarising the results of a therapeutic trial or for making medical decision about a patient. In a trial comparing a new treatment with a standard one, the NNT gives the estimated number of patients who need to be treated with the new treatment rather than the standard treatment for one additional patient to benefit. The NNT is the inverse of the absolute risk reduction—that is, the difference between the proportion of events in the control group (P_c) and the proportion of events in the treatment group (P_t).

---

* Laupacis, A., Sackett, D.L. and Roberts, R.S. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med*, **318**: 1728–33, 1988.

** Cook, R.J. and Sackett, D.L. The number needed to treat: A clinically useful measure of treatment effect. *BMJ*, **310**: 452–4, 1995.

$$\text{Number needed to treat} = \frac{1}{(P_c - P_t)}$$

It can also be calculated by the formula,

$$\frac{1}{P_c \times \text{relative risk reduction}}$$

For instance, if the relative risk reduction is 50% and the spontaneous risk (P_c) of events is 20%, then

$$\text{Number needed to treat} = \frac{1}{0.2 \times 0.5} = 10$$

This means the treatment would lead to 'saving' one patient for every 10 patients treated. A large treatment effect leads to small NNT and if there is no treatment effect, the NNT will be very large (infinity).

Example. The effect of insulin-glucose infusion followed by multidose insulin treatment in patients with diabetes mellitus and myocardial infarction was studied in a randomised controlled trial*. The results of one-year follow up are given in the following table:

	Outcome of first year follow up		
Groups	Dead	Alive	Total
Insulin (Treated)	58	248	306
Control	82	232	314
Total	140	480	620

$$\text{Risk of death among the treated} = P_t = \frac{58}{306} = 0.190$$

$$\text{Risk of death among the controls} = P_c = \frac{82}{314} = 0.261$$

$$\text{Relative Risk (RR)} = \frac{P_t}{P_c} = \frac{0.190}{0.261} = 0.728$$

$$\text{Relative Risk Reduction} = 1 - \left(\frac{P_t}{P_c}\right) \quad \text{or} \quad \frac{(P_c - P_t)}{P_c} = 0.272$$

$$\text{Absolute Risk Reduction} = P_c - P_t = 0.071$$

The number needed to treat is the reciprocal of the absolute risk reduction.

* Malmberg, K. Prospective randomised study of intensive insulin treatment on long-term survival after acute myocardial infarction in patients with diabetes mellitus. *BMJ*, **314**: 1512–15, 1997.

$$\text{NNT} = \frac{1}{P_c - P_t} = \frac{1}{0.261 - 0.190} = \frac{1}{0.071} = 14.08$$

This means 14 patients would have to be treated to prevent one death.

The Confidence Interval for the NNT can be obtained using Normal Distribution. For the difference between two rates, $(R_1 - R_2)$, the 95% confidence interval is

$$(R_1 - R_2) \pm 1.96\ SE\ (R_1 - R_2)$$

where $SE\ (R_1 - R_2) = \sqrt{\left[R_1 \frac{(1-R_1)}{n_1} + R_2 \frac{(1-R_2)}{n_2}\right]}$

and R_1 and R_2 are the rates of two groups and n_1 and n_2 are the sample sizes of the two groups respectively.

Applying these, the 95% confidence limits for the difference between the proportions of events in the control group and the insulin (treatment) group, that is of $P_c - P_t$, are estimated as 0.006 and 0.137. Taking the reciprocal of these limits and reversing them, obtain the confidence limits of NNT. Hence, the 95% confidence interval of the NNT is 7.3–166.7. The smallest number needed to treat is obtained from the lower limit of the 95% confidence interval of the NNT, which indicates the highest benefit associated with the treatment. Similarly, the upper limit is interpreted.

24.11 SOME COMMENTS

A broad overview of observational studies including their strengths and limitations is described in this chapter. The main objective of observational studies is to identify factors and establish causal relationships. Statistical methodology both in terms of designing and analyses of observational studies has made great strides in the recent past. Several excellent books are available for further study and application. In particular, the methods for handling multiple variables or taking into account person-years of observation and use of statistical models such as logistic or loglinear or Cox proportional hazards are just some of the innovations available. Investigators should refer to these methods and consult a proper statistician for additional guidance and help.

QUESTIONS

1. Distinguish between prospective studies and retrospective studies.
2. What is meant by bias? Mention the sources of bias. Describe how it can be controlled.
3. Define the Odds Ratio. What does it measure? When will you apply it?

CHAPTER

25 Data Processing

25.1 INTRODUCTION

Research data are collected in relation to the objectives formulated. If the sample studied is small and the number of items on which information collected is also limited, then data processing constitutes a simple matter especially if the records and questionnaires have been designed properly. On the other hand, if the sample size is large and/or data on a multiplicity of items requiring study of many correlations are collected, then it is necessary to plan for efficient data processing. Such processing of data should preferably begin as early as possible after data collection has started in order to check for inconsistency, errors or incompleteness. These exercises, in addition to helping in quality control of data also help in visualising final analyses. Data processing thus constitutes an important intermediate stage and if planned well, can result in speedy processing and statistical analyses.

In large-scale studies, direct handling of records for data management becomes quite cumbersome. It would also be necessary to keep track of the information collected in a manner which would facilitate a quick review of the progress and quality of the data obtained. Basically, two systems can be used in practice:

1. Registers with appropriate columns designed for entering relevant information from each record in a brief and concise manner or as coded material.

2. Unit record systems where individual records are maintained for each unit which can be arranged in any manner.

Information recorded can also be briefly described or coded. The registers facilitate a general review and analysis by page totals or simple one-way tables. If multiple correlations are required, the register becomes difficult to handle. In comparison, the unit record system is flexible and permits rearrangements or review of any specified group. One advantage of the register may be that by a serial number, the accumulation can be determined more easily.

In any project, the relative advantages of these two systems should be considered and a suitable method followed. With the advent of computers, data management using register method (spread sheets) or individual records have become faster and more efficient, using several kinds of user-friendly software.

25.2 COMPUTER HARDWARE

Computers are electronic devices that enable storage of large quantities of data and the analysis of such data at very high speeds. Non-specialised computers are often

categorised into three groups—microcomputers, minicomputers and mainframes, according to size, power and cost, although the distinction between these groups is somewhat imprecise (for example, a large minicomputer may be much more powerful than a small mainframe) and it is quite likely that these categories will soon become obsolete in light of the rapid technological developments. Best known to the general public are the so-called microcomputers, which range from the very small 'personal' or 'home' computer to the compact desktop type computer suitable for business, academic and similar applications.

A computer consists of a number of physical units (hardware) such as the CPU (central processing unit), the monitor or VDU (visual display unit), the I/O (input-output) devices, e.g., a typewriter type of keyboard for entry of data into the computer, magnetic storage devices such as floppy diskettes, hard disks, CD, cassette tapes, and printing and plotting units. The floppy diskettes or 'floppies' come in various sizes, densities (high density and low density) and capacities (360 KB–1.44 MB). A CD and the newer storage devices can have a capacity of 500 MB or more.

The capacity of a computer is usually expressed in terms of 'bytes' of information it can store and handle at a time (memory), which ranges from 1 KB (1024 bytes) to 4000 KB (4 MB) or more. Depending on the capacity, large quantities of data can be stored in computer tape reels and on disk packs. The data available in these media can be read into a computer at great speeds and analysed quickly. Once the accuracy of the fed data is ensured, the computer can be relied upon to provide error-free analyses of these data, provided the computer program which analyses these data is correctly written.

Basic to all these machine types is the microprocessor, comprising thousands of transistors on a silicon chip. It is the technical development in the production of microprocessor components and other advances in computer technology that have brought about a tremendous increase in the power of computers and reduction in their cost.

In addition to the direct costs for the basic computer system, one needs to add the costs for a UPS (Uninterruptible Power Supply), air-conditioning, modems, storage facilities for records/floppies/diskettes, printers, suitable furniture and similar items.

25.3 COMPUTER SOFTWARE

A special kind of program called an Operating System controls the flow of information between the different components. The most popular ones are MS DOS (Microsoft Disk Operating System) and Unix. One of its most important jobs is to enable the hardware to understand the instructions of the software. In addition, the operating system manages files on the disks, listing the directory of files. The medium of communication with computers has changed over the years. Programming a computer requires knowledge of the computer languages such as COBOL (Common Business Oriented Language), FORTRAN (formula translation), BASIC (Beginners All-purpose Symbolic Instruction Code), Pascal, C, Java, Visual Basic, etc. Using these languages (known as software), a program (set of instructions) is prepared for the computer to execute any given job of analysis.

Today, however, many software packages such as MS Word, MS Excel, etc. require practically no prior knowledge of the programing languages.

Several statistical package programs such as SAS, BMDP, SPSS, STATA, EPI-Info, Egret, etc. have been developed which can be used to summarise statistical data and to carry out all the common statistical analyses on these data. Special programs can be written in a suitable computer language when pre-written package programs are not available. A brief description of Excel, EPI-Info, SPSS and STATA is as follows:

25.3.1 Spreadsheets

There are several varieties of worksheets with cells formed by rows and columns, like in a register, known commonly as spreadsheets. The most popular one is known as Excel or Microsoft Excel. In this, you can enter data, use formulae for arithmetical and statistical operations, and even do error checking. Excel is useful for carrying out univariate statistical analyses, keeping accounts, monitoring data acquisitions, to preview your data entry and simple tabulations, as well as print them. You can even prepare simple statistical diagrams like histograms, scatter diagrams, etc. using Excel.

Several simple books are available for guidance.

25.3.2 Epi-6, Epi-2000, Epi-2002

These are easy-to-use statistical packages for data entry and analyses, developed by the Center for Disease Control (CDC), Atlanta, Georgia, USA. These epidemiological data management software can be freely downloaded through internet, and has tutorials as well as worked out examples. The main advantage of this package is that a template can be created for data entry in any format, as for example, like a hospital chart or interview questionnaire or survey pro forma. Subsequently, simple statistical analyses can be easily carried out. This is menu-based, user-friendly, which can be easily learnt and adopted.

25.3.3 SPSS

SPSS (Statistical Package for Social Sciences) is probably the most popular, and carries virtually all the possible statistical analyses from the simplest to the most advanced multivariate statistical procedures, including nonparametric statistics, survival analyses and so on. The higher version of SPSS, say 12.0, is available with new options such as the exporting files .xls and so on.

25.3.4 STATA

STATA is another statistical package for managing particularly large datasets, analyse the statistical properties of new or existing estimators and tests. These may include topics such as simulations, power analyses; comparisons of tests or estimators. The STATA help menu gives a brief description about STATA commands and its features. The higher version STATA 8.0 is a window-based version and user friendly.

SPSS takes up considerable space on the hard disk and many of the procedures may not even be used. Also the outputs are very comprehensive, not all of them required. On the other hand, STATA is more specific and occupies less memory.

Familiarity with any one package will suffice to carry out the most intricate statistical analyses.

25.4 CHOOSING A COMPUTER SYSTEM

The computer market is flooded with a plethora of models varying in their power, flexibility, price and availability of service facilities. Both hardware and software have had revolutionary changes to match the needs of requirements of consumers. Both individual (stand-alone) and networks (LAN, WAN) are now popular and very user friendly. Further desktop and laptop models are easily available with ever increasing power, size and reduced price-tags.

In the recent past, several companies have introduced microprocessor-based computers on the Indian market which are suitable for meeting the data processing needs of small and medium-sized organisations. As is invariably the case with new technologies, there is a 'learning-by-doing' process which requires adaptation and the acquisition of new skills and practices, and most of all an understanding by the user of what a computer can and cannot do, in the particular application for which it is designed and installed. Too often, purchasers have ambitious but vaguely articulated expectations of what a computer can do, and are subsequently disappointed, when the computer cannot perform the tasks expected of it because the wrong system has been purchased.

25.5 ROLE OF COMPUTERS IN DATA MANAGEMENT

Computers are useful not just for statistical analyses, but also to monitor the accuracy and completeness of the data as they are collected. Thus, computers help in data entry, data editing, data management including follow up actions, etc. Computers also allow for greater flexibility in recording the data while they are collected as well as greater ease during the analysis of these data.

In research studies, the preparation and inputting data is generally the most labour-intensive and time-consuming aspect of the work. Typically, the data will be initially recorded on a questionnaire or record form, and it is necessary to convert this into a form suitable for acceptance by the computer. To do this the researcher, in conjunction with the statistician and the programmer, will prepare a list of coding instructions, which specify precisely how the data are to be converted to computer-readable form. Unless the questionnaire or record sheet has been expressly designed for subsequent computerisation, an intermediate state, by which the original data are transferred onto a coding form or sheet, may be required to facilitate conversion to tape or disk.

25.6 FURTHER USES OF COMPUTERS

Needless to say, computers have demonstrated their veracity in word processing, preparing reports and manuscripts for publication.

With the availability of Internet facilities, computer can access several literature databases, and help in review of literature and bibliographic researches. It can thus be used for storing relevant published articles, to be retrieved whenever needed.

Another interesting application of computers is in simulation techniques, of which there are various examples in medicine and health care. Briefly, this involves the formulation of a mathematical model to a system (which may be based on actual or hypothetical data), and the use of the model to 'simulate' the consequences of particular events or decisions, which are specified by the model user. These are designed to reflect 'real life' possibilities and to indicate, with varying degrees of probability, what might happen in particular situations. The use of simulation models in medicine and health care includes the following:

1. Analysis of the effects of epidemics under varying assumptions about population movement and rates of infection.
2. Determination of 'queuing' or waiting times at clinics, under varying assumptions about rates of referral, treatment times and patient throughput.
3. Training of health care administrators or managers in resource allocation decisions through simulation of the operations of health care systems (e.g., hospitals).

Simulation techniques are also being developed for training in diagnosis. The diagnostic checks and procedures used in simulation may be also used with real patients, the computer in this case undertaking the evaluation on the basis of the patient information with which it is supplied and printing out 'its' diagnosis or, more likely, a list of possible diagnoses.

In a more comprehensive way, medical record linkage provides a continuous record of individual patients from birth to death, including illnesses, hospitalisation, operation, allergies and so on. Prior to the use of computers, the linking together of different events in an individual's medical case history was extremely difficult. Individual events (treatment by family doctor, stay in hospital, etc.) were, and indeed usually still are, recorded separately at different locations with no systematic procedures to bring together these elements of a case history into a single record. With computer-based recording systems, such linking becomes more feasible, medical records can be put to better use and information stored efficiently; moreover, this information can be retrieved more rapidly than before. Cumulative files for individuals can be compiled and assembled into family groups, socio-economic categories, and so on, for purposes of analysis. Longitudinal studies, which involves following up a given group of cohort over a period of time, are a well-established application of the concept of medical record linkage, and the large and comprehensive computerised databases established through record linkage open up a wide range of potential research studies.

25.7 A NOTE OF CAUTION

It is wise to remember that a computer is just a tool and a resource. If the methods of handling the data described briefly in this chapter are to be applied efficiently, adequate planning and suitable organisation is necessary not only for recording of information in the proper format but also for collection of the required data. No facility can replace this aspect of planning. Further, it would be a disaster to replace the statistician by a computer, no matter how powerful, since statistical analyses are built on sound principles of design, implementation and handling of exigencies in data collection, all of which require the expertise of a qualified statistician.

25.8 CLUES FOR USING SPSS AND EXCEL

In the menu driven SPSS package, click "Analyse", choose the command "Descriptive Statistics", then choose the subcommand "Frequencies" to get the frequency distribution. One must be cautious in using this command, when the raw data is a continuous variable. Histogram is obtained using the subcommand "Charts".

To get both the measures of central tendency and measures of variation, click "Analyse" in SPSS package, choose "Statistics" under the "Frequencies" Command. When a variable is recoded into categories, such commands will give the measures of the codes instead of the variable. Measures of central tendency and dispersion may also be obtained using alternative commands.

Excel is one of the programs useful for data entry, data editing and data analysis. It contains a number of worksheets (spreadsheets). The main features of the worksheet are:

1. The Title Bar
2. The Menu Bar
3. The Tool Bar
4. The Column Heading
5. The Row Heading
6. The Formula Bar.

The Title Bar, Column Heading and Row Heading are used for creating data files.

The Tool Bar contains "Add-ins" or "Analysis tool pack" from which descriptive statistics such as mean, median, mode, standard deviation, skewness, and kurtosis, and also charts like histogram, bar chart can be obtained.

QUESTIONS

1. Describe the terms 'software' and 'hardware'.

2. Explain the features of some commonly used computer software packages.

3. Write a note on the role of computers in research data management.

CHAPTER

26 Presentation of Data

26.1 INTRODUCTION

In Chapter 5, the process of presentation of data as frequency tables and diagrams was described. Reporting of results is an important part of any investigation as it constitutes the means of communication. Careful consideration of the principles of presentation will greatly help in successful reporting. We shall consider in this chapter further details of presentation of numerical data of certain types by means of tables and diagrams. The time series and its importance in medicine are also discussed.

26.2 BASIC PRINCIPLES

A properly constructed and adequately labelled table can be read and understood independently without consulting the accompanying text. On the other hand, a table which is not itself understandable may confuse the meaning of a text which is otherwise clear. For people in the field of medicine and public health, knowing how to organise data into a table is as important as knowing how to write. By the same reasoning, the ability to read a table is as important as the ability to read printed words. Tables can be brought under the following two classes based on the purpose of presentation:

26.2.1 General Purpose Table

A general purpose table is prepared to present information which is intended for use by an outside party. Data are usually broken down in considerable detail and completely classified according to every characteristic for which they are presented. It is desirable that such a table is accompanied by extensive notes defining the terms used and describing the method of collection of the data, changes in classification and definitions or other factors which will lead to noncomparability of the data. Examples of such tables are found in census reports.

26.2.2 Special Purpose Table

A special purpose table is prepared to form part of some report in order to provide detailed evidence for some point made in that report. It may be used in conjunction with a graph giving the exact values plotted on the graph or it may constitute the sole presentation of some figures.

26.3 ELEMENTS OF A TABLE

While presenting data as a table the following parts of the table require to be considered:

(i) Title. A clear-cut and concise title consisting of statements of the following items is an essential part of every table:

1. What the data represent
2. The source of the data
3. The period to which they relate
4. The classification used.

A table with an inadequate or unsatisfactory title may cause many readers to discard an article without reading it, or to misinterpret the data presented.

(ii) Unit of measurement. The unit of measurement of each characteristic presented in the table must be shown in the table itself. When there is only one characteristic, its unit of measurement may be shown just below the title, forming almost part of it. If more than one characteristic is involved, each unit of measurement may be stated against the corresponding characteristic shown in the captions (column headings) or in the stubs (labels for the rows).

(iii) Captions and stubs. The descriptions of the characteristics shown in the captions and/or in the stubs must be concise and clear cut.

(iv) Interpretative aids. Inclusion of certain interpretative aids such as totals, per cent difference, per cent of totals, previous figures, etc. may be found useful in special purpose tables.

26.4 CLASSIFICATIONS

In the process of tabulation, the following types of classification are encountered:

26.4.1 Classification by Space

In classification by space, data are classified by location of occurrence. Arrangement of the set of categories in alphabetical order of the terms defining these categories, or in the order of their geographical location, may be found to be suitable in many cases (Table 26.1).

26.4.2 Classification by Attribute

When the data represent observations made on a qualitative characteristic (as distinct from a quantitative characteristic) the classification in such a case is made according to these qualities. Alphabetical arrangement of categories may be suitable for a general-purpose table; but in the case of a special-purpose table, arrangement may be made in the order of importance of these categories, the importance being measured by the number of observations in each category (Table 26.2).

Tabel 26.1 Number of Literates in Rural and Urban Areas in India (statewise list)

States	No. of literates (in thousands)	
	Rural	Urban
Andhra Pradesh	14,821	10,119
Assam	7921	1709
Bihar	20,358	6485
Maharashtra	22,164	20,774
Gujarat	12,096	9179
Kerala	16,443	6228
Madhya Pradesh	14,464	9027
Tamil Nadu	17,424	12,958
Karnataka	12,267	8812
Orissa	10,303	2608
Punjab	6253	3699
Rajasthan	8189	5428
Uttar Pradesh	33,079	13,968
West Bengal	20,337	12,382
Total	216,119	123,376

Total 26.2 Patients who Attended Leprosy Clinics during 1993

Type of leprosy	No. of patients
Tuberculoid	604
Lepromatous	272
Indeterminate	72
Borderline	48
Total	996

26.4.3 Classification by Time

In this case, data are classified by time of occurrence of the observations. Arrangement of categories is almost always in chronological order (Table 26.3).

Table 26.3 Number of Primary Health Centres in India at 5-year Intervals during 1970–1990

Year	No. of PHCs
1970	5015
1975	5293
1980	5484
1985	7284
1990	18,981

Source: 1. Health Information of India, 1989, Ministry of Health and Family Welfare, Govt. of India, New Delhi.
2. Bulletin on Rural Health Statistics in India, Sep. 1993. Ministry of Health and Family Welfare, Govt. of India, New Delhi.

26.4.4 Classification by Size of Observation

When the data represent observations of some characteristic on a numerical scale, classification is made on the basis of the size of individual observations. The range of the observations is suitably divided into smaller divisions called class intervals and the number of observations falling in each of these intervals is counted and recorded. The numerical scale adopted may be either discrete as in the case of W.B.C. counted on a slide, or continuous as in the case of haemoglobin. It is called continuous, because any value in the observed range on the scale is theoretically possible. Class intervals should be determined keeping the following guiding principles in mind:

1. The number of class intervals should be sufficient to condense the data bringing out their significant features. From ten to thirty class intervals are usually found to be satisfactory for a fairly long series.

2. Uniform sized class intervals (e.g., Table 26.4) are preferable unless some important details are lost by this arrangement; because in such cases interpretation is easier and complications are avoided in graphing and arithmetical calculation.

Table 26.4 Protein Intake of 400 Families from the Rural Population (Grams per Consumption Unit Per Day)

Protein intake (g) per consumption unit per day	No. of families
15–25	30
25–35	40
35–45	100
45–55	110
55–65	80
65–75	30
75–85	10
Total	400

3. Sometimes open-end class intervals (e.g., Table 26.5) are used. (As far as possible, they must be avoided.)

Table 26.5 Distribution of Single Live Born with respect to Birth Weight

Birth weight (grams)	No. of children
< 1500	24
1500–1999	60
2000–2499	422
2500 and above	3202
Total	3708

Source: Outcomes of Pregnancies, Monograph No. 17: Longitudinal studies in child growth and development, Department of Biostatistics, Christian Medical College, Vellore, 1992.

4. Class limits must be precisely and clearly stated.

5. For many arithmetic calculations the mid-point of each class interval will be used as a representative of each item in that interval. The following precautions are, therefore, necessary:

(a) the intervals should be so chosen that the mid-point of each interval is approximately the average of the items in that interval;

(b) any tendency for the items to concentrate at certain points must be recognised; and

(c) it will be convenient if mid-points and limits are whole numbers.

Table 26.6 Family Planning Acceptors ('thousands) in India: 1977–91

Year	Sterilisations (000)	IUD insertions (000)	Total acceptors (000)
1977–78	949	326	4528
1978–79	1484	552	5505
1979–80	1778	635	5482
1980–81	2053	628	6490
1981–82	2792	751	8102
1982–83	3983	1097	11,028
1983–84	4532	2134	15,056
1984–85	4085	2562	16,442
1985–86	4902	3274	18,920
1986–87	5043	3935	20,632
1987–88	4940	4356	22,702
1988–89	4678	4851	24,368
1989–90	4188	4942	26,082
1990–91	4123	5368	27,396

Source: Ministry of Health and Family Welfare, Govt. of India, Family Welfare Programme in India: Year Book 1990–91, New Delhi.

26.5 PRESENTATION BY GRAPHS AND DIAGRAMS

In addition to tables, graphs and diagrams are widely used in the presentation of statistical data either alone or in conjunction with the tables. They furnish a visual method of examining quantitative data. They make a more lasting impression than detailed numbers and convey an idea forcefully. They help us have a real grasp of the overall picture rather than the details. In order that these graphs and diagrams present ideas truthfully and emphasise correct ideas, they must be drawn according to certain basic rules which are dependent partly on convention, partly on mathematical considerations and partly on personal preferences. Many types of graphs and diagrams are commonly used. We shall consider some important types in relation to the type of data to be presented. As in the case of tables, a title indicating what is being presented and labels for each axis to describe the unit of measurement and the categories are essential parts of a graph.

26.6 TIME COMPARISON

The following diagrams are used for time comparison.

26.6.1 Bar Diagram

A bar diagram is commonly used to provide a visual comparison of figures in a time series. Figure 26.1 is an illustration of a bar diagram for the data in Tabe 26.3. The figure for each period is represented by a bar drawn for that period. Each bar

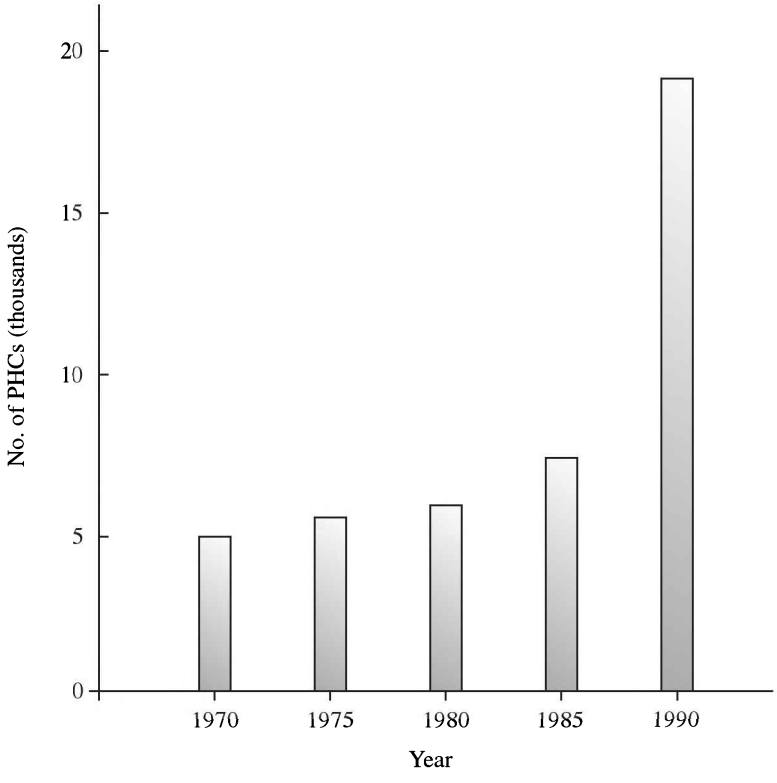


Figure 26.1 Number of primary health centres (PHCs) in India: 1970–1990.

must be of the same width. The height of the bar over a period represents the figure for the corresponding period. In order to avoid erroneous comparisons, the scale of measurements of the bars must start at zero. The use of divided bars is common when a series and its component parts are to be shown.

26.6.2 Line Graph

A line graph is to be preferred when emphasis is on the trend of the time series over the period rather than on the comparison of relative sizes of the different figures in the series. In Figure 26.2, the use of a line graph is illustrated for comparison of two time series.

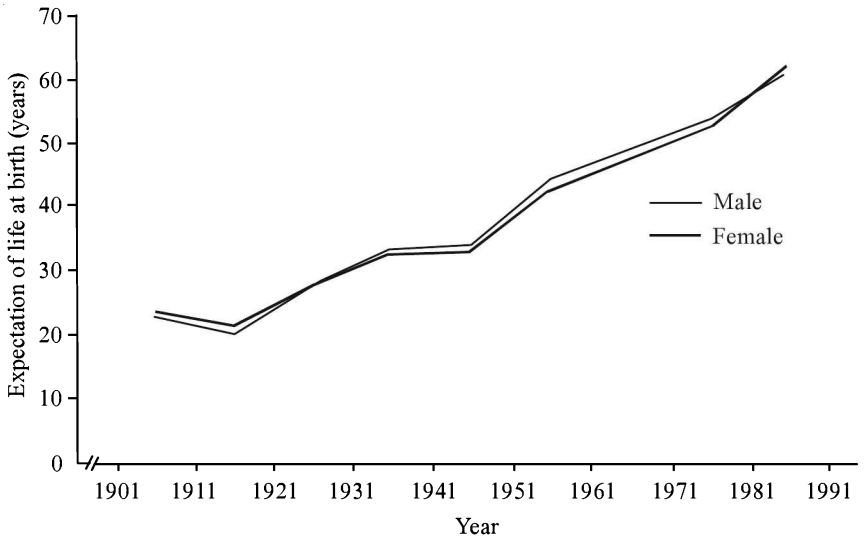


Figure 26.2 Expectation of life at birth in India for males and females: 1901–1991.

When relative changes in the series are to be studied, the correct idea is conveyed by the use of logarithmic scale to represent the figures in the series (Table 26.6; Figure 26.3). Figure 26.4 represents the same data on an arithmetic scale.

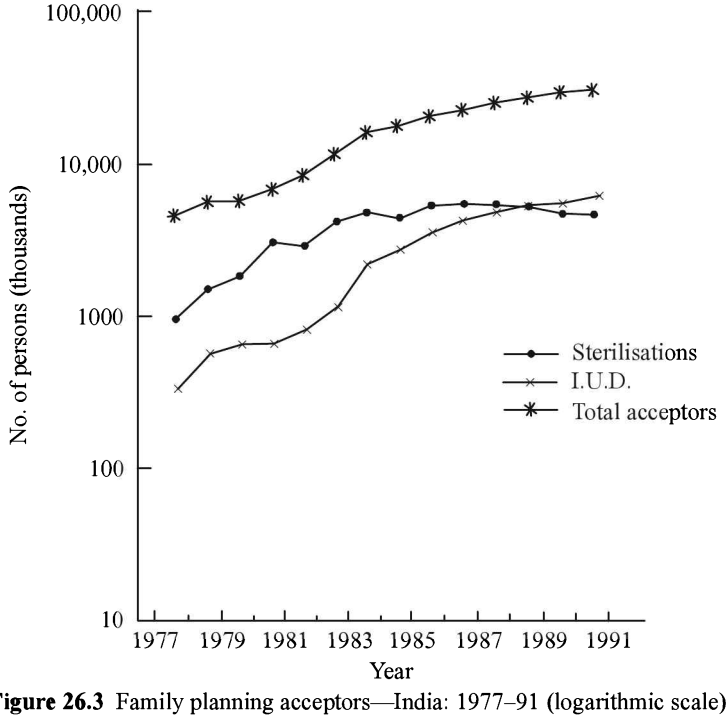


Figure 26.3 Family planning acceptors—India: 1977–91 (logarithmic scale).

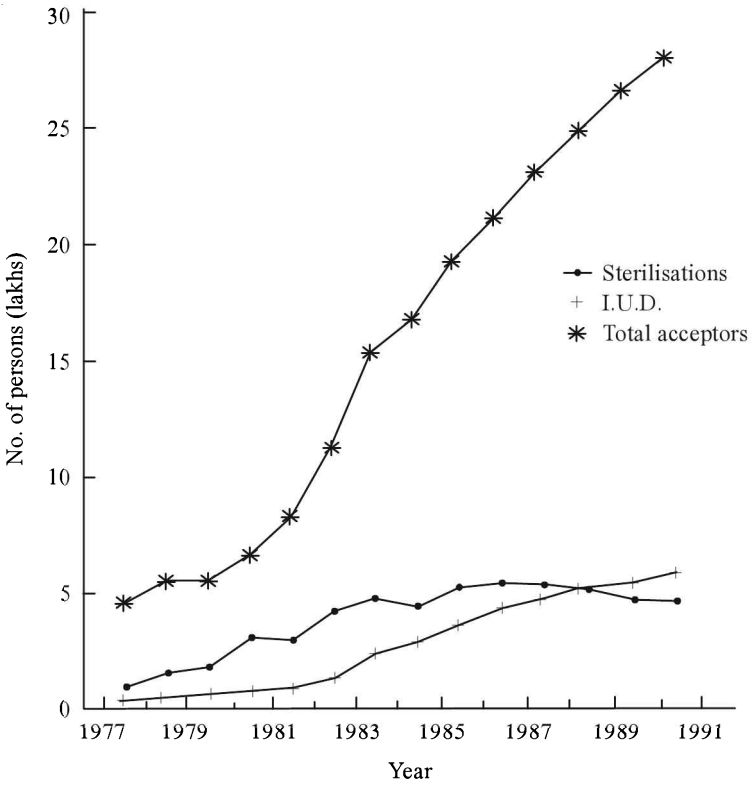


Figure 26.4 Family planning acceptors—India: 1977–91 (arithmetic scale).

When it is thought desirable to break the scale of measurement either to use less space or to magnify fluctuations, one must consider the danger of conveying thereby a false idea, and so the break should be clearly indicated on the graph (Figure 26.5). By a change in the relative size of the two scales used, it is possible to make the graph appear very different (see Figure 26.6) which will convey a different idea of the series graphed (Figures 26.5 and 26.6 use the same data).

It is, therefore, necessary to make sure that the correct idea is conveyed by the graph that is drawn.

26.7 SPACE AND ATTRIBUTE COMPARISONS

The commonly used diagrams for space and attribute comparison are described here.

26.7.1 Bar Diagram

To compare figures corresponding to different places or different categories of a qualitative characteristic a bar diagram may be used. Horizontal bars are used frequently in contrast to time series where almost always the bars are vertical. The

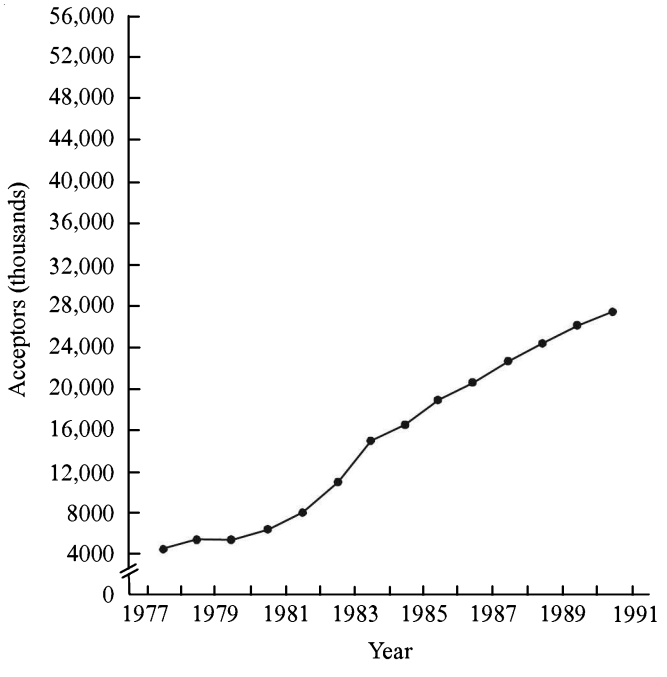


Figure 26.5 Family planning acceptors—India: 1977–91.

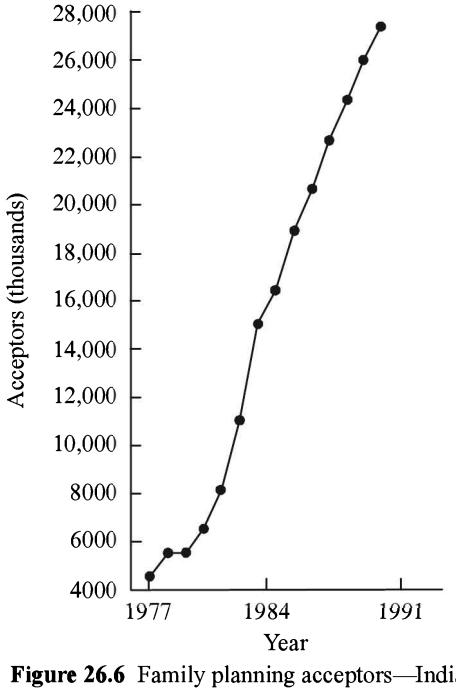


Figure 26.6 Family planning acceptors—India: 1977–91.

length of the bar represents the figure for the corresponding category. Arrangement of categories in ascending or descending order of the corresponding figures for these categories may assist comparison of the figures more easily. Figure 26.7 illustrates the use of bar diagram in space comparison.

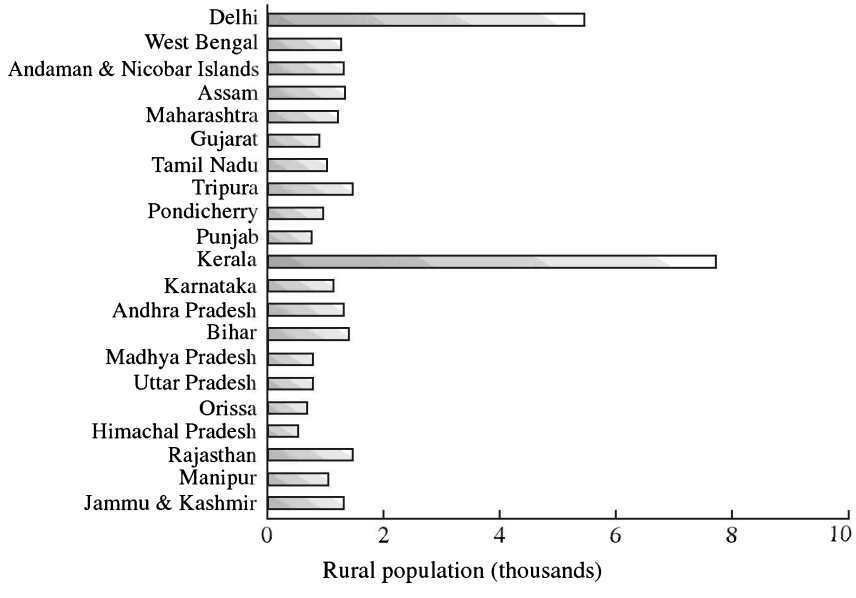


Figure 26.7 Rural population served by a trained *dai* in India.

26.7.2 Component Bar Diagram and Pie Diagram

Either a component bar diagram or a pie diagram is chosen when component parts of the figure relating to each category are to be shown. Figures 26.8(a) and 26.8(b) illustrate the use of these two diagrams. On the circles, the size of the figure for each category is shown by the total area of the circle, and the sizes of the component parts are shown by the areas (or the angles subtended at the centres) of sectors of the circles. Since the eye can measure and understand a linear distance much more easily than angles or areas, a bar diagram is preferable.

26.8 ANALYSIS OF TIME SERIES

One of the first elements in an epidemiologic enquiry is to study the time distribution of cases. This method is now applied to almost all diseases. The prompt recognition of epidemic movements demands a dynamic conception of statistics of epidemic diseases and a knowledge of the analysis of time series. For the application of this type of analysis, information about the event is needed over a reasonably long period of time. It is customary to maintain charts to give rapid visual information about the progress of various diseases. Figure 21.3, for instance, presents birth rate and death rate in India for the period 1881–1991.

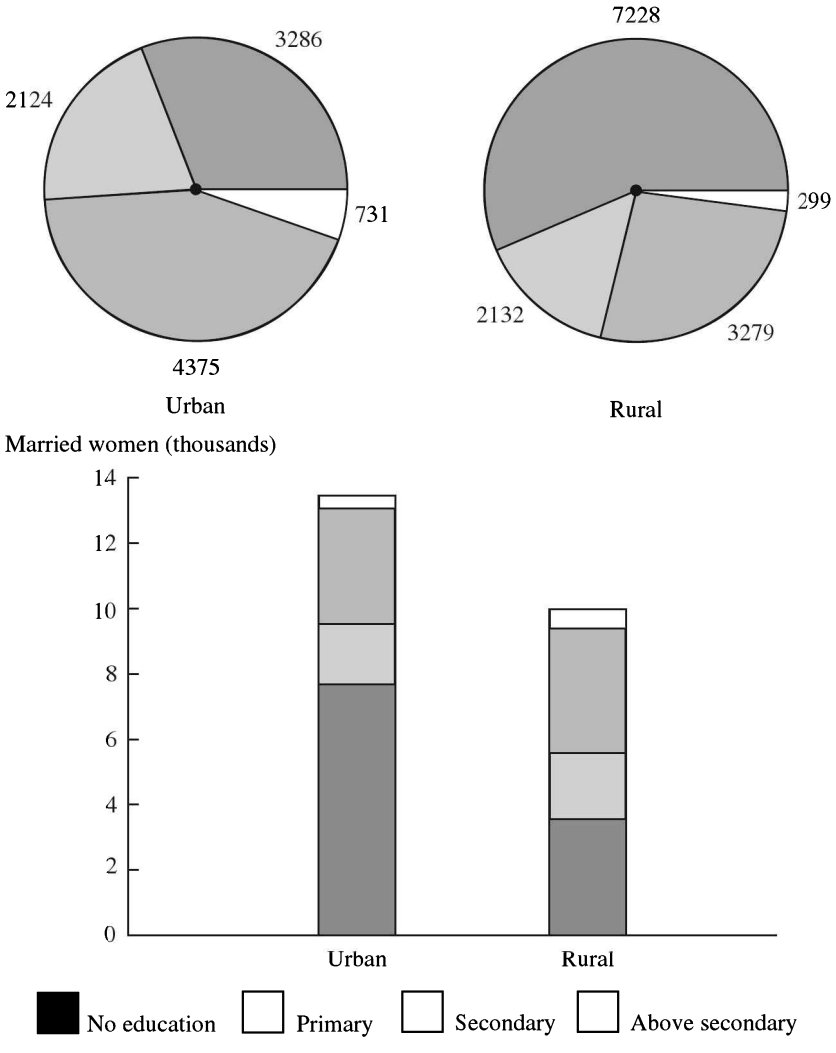


Figure 26.8 Educational status of married women.

A time series has the following four components:

(i) Long-term movement (or secular trend). The general nature of the movement observed in the data over a long period of time is referred to as secular trend. The annual crude death rate (Figure 21.3) shows a steady decline during the period under study. When such increase or decrease appears to be a straight line, the trend is determined by fitting a linear regression line. The technique of moving average is often used to find the trend.

(ii) Seasonal variation. Many diseases exhibit increased incidence during some months of the year. For example, smallpox cases are seen more during the summer

months. In the data on such diseases, the seasonal influence will be noticed along with that of the secular trend. In Figure 26.9 is shown the seasonal pattern of cholera cases in Calcutta during the years 1950 to 1956.

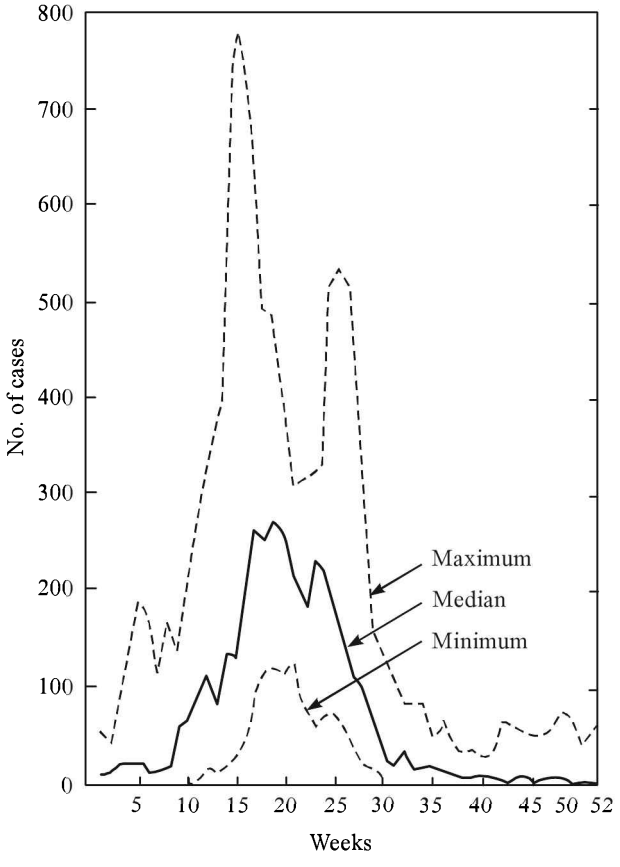


Figure 26.9 Seasonal pattern of Cholera cases—Calcutta, 1950–1956.

(iii) Cyclic trend. Cyclic trend refers to a regular cyclic pattern observed in the data over a longer period than the seasonal pattern considered above. Smallpox in India shows a cyclic periodicity of 5–7 years and measles of 2–3 years.

(iv) Irregular fluctuations. In addition to one or more of the three characteristics mentioned above, every longitudinal data on illness has another component known as irregular fluctuation. It represents the random variation in biological phenomena in general.

A study of these basic characteristics of a time series of illness provides a logical basis for projections into the future which will be useful in planning for timely action for the prevention or control of diseases. A study of the time series will also help in understanding the phenomenon and its relation to certain basic factors affecting it.

26.9 SPECIAL DIAGRAMS FOR CLINICAL RESEARCH

Some special diagrams which are commonly used for presenting the findings or the pattern of distribution of clinical data are given here. Modified versions of these are also used frequently.

26.9.1 Dot Diagram

This is mainly used for presenting the results of experimental studies. The values of outcome (variable) are represented as dots, one for each subject, belonging to the experimental groups and control group. Sometimes, the mean for each group is also represented along with them. Figure 26.10 is an example of this.

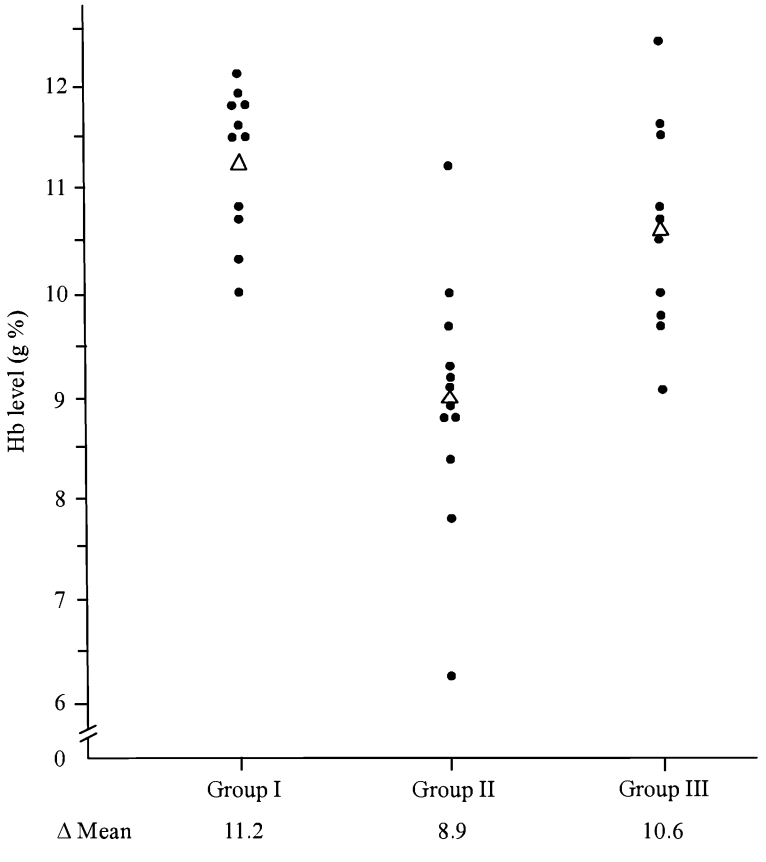


Figure 26.10 Haemoglobin (g%) of three groups of children fed with different diets.

26.9.2 Bar Diagram with Confidence Interval (CI)

In any presentation of the means of a phenomenon, giving the confidence interval helps interpretation. When the means are represented as bars, the confidence interval is represented as a vertical line at the top of the bar. Sometimes, this is also extended inside the bar to represent full interval. Figure 26.11 illustrates this.

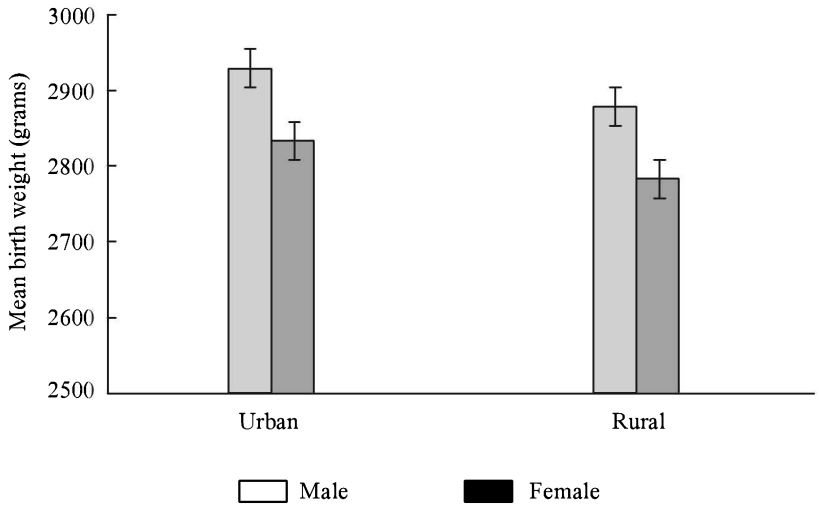


Figure 26.11 Mean birth weight of male and female babies in urban and rural sectors.

26.9.3 Special Line Diagram

When several groups of experimental subjects are studied over time or over different doses or over different ages, not only are the differences between means at specific points compared but the trend over the characteristic is also studied. This is represented as a line diagram for each group in which the means are joined to study the trend. At each mean, one standard deviation is marked above and below it. An example is given in Figure 26.12.

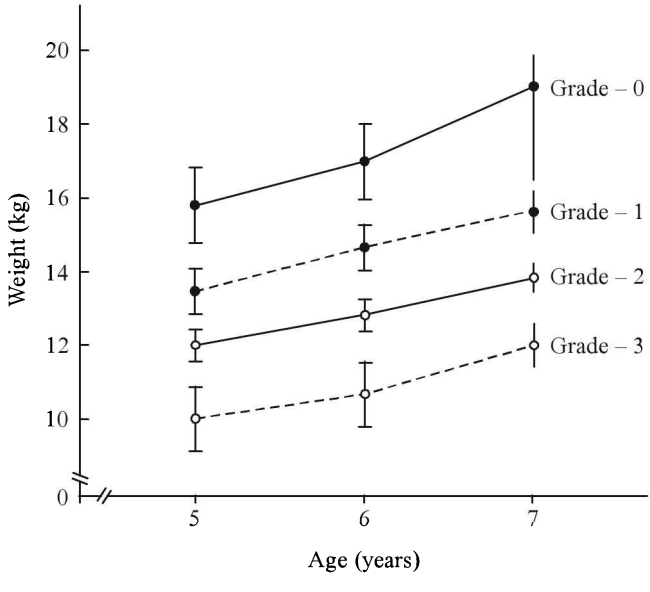


Figure 26.12 Weight of girls by different nutritional grades (nutritional grade was determined at age 5).

26.9.4 Box Plot

A box plot diagram is used to describe the data when large numbers are studied. This has a box with a horizontal dividing line representing the median or Q_2, the top of the box represented by Q_3 and the bottom of the box by Q_1. The vertical line at the bottom of the box extends upto the 10th percentile and at the top of the box goes upto the 90th percentile. An example is given in Figure 26.13.

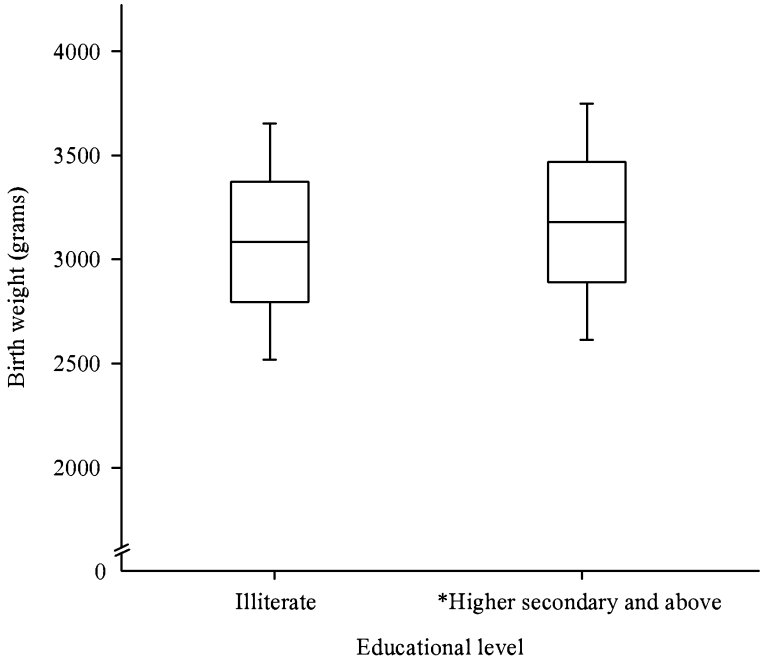


Figure 26.13 Birth weight of babies of illiterate and educated (HS* +) mothers.

26.9.5 Stem and Leaf Plot

All observations of the study are presented as a stem and leaf plot wherein the accuracy of the individual observation is maintained along with depiction of the pattern of distribution.

With the availability of many computer graphics programs it is possible to draw these easily. Many more methods of presenting clinical data are also available through relevant computer software. The following tabular data shows the stem and leaf plot for the data given in Table 7.1.

Stem and Leaf Plot

10	4885	4
11	846,276	6
12	2,930,472,692	10
13	32,508	5
14	2	1
		26

QUESTIONS

1. Describe the types and elements of a 'table'.
2. State the advantages of using charts and diagrams.
3. How will you represent trends as a chart/diagram?
4. Write a note on special diagrams used in clinical research.

Appendices

Appendix 1

Table A.1.1 500 Haemoglobin Values

11.6	10.3	9.7	12.4	11.4	10.7	11.1	9.5	10.6	8.0
13.0	10.3	9.9	9.0	11.4	9.6	11.0	11.7	11.0	8.9
8.9	10.8	10.0	11.5	7.5	9.8	12.3	11.4	11.9	12.3
11.9	10.3	8.5	9.8	11.8	10.7	8.9	9.5	12.9	12.8
11.8	10.0	7.9	10.3	11.4	11.8	11.8	8.8	9.2	10.7
11.2	10.4	11.1	13.3	8.2	11.6	9.1	8.4	12.0	12.1
12.3	12.9	12.7	12.1	11.4	11.2	6.3	10.0	9.7	11.2
10.5	9.0	12.2	11.5	6.6	11.5	12.1	10.8	12.1	10.3
12.5	8.9	10.7	12.4	10.8	12.5	10.8	11.8	11.0	11.8
12.5	7.6	9.4	9.2	12.2	12.8	9.3	11.7	12.0	11.5
9.7	9.3	10.7	10.0	12.9	11.4	11.9	10.7	12.1	8.8
11.5	12.9	10.7	9.7	12.3	11.5	10.7	11.4	10.8	12.3
12.1	7.8	11.4	12.5	9.8	10.3	10.0	11.7	9.3	8.4
12.1	13.3	10.5	11.5	11.8	12.7	11.5	11.1	12.0	13.3
12.3	9.0	8.2	9.4	10.5	10.7	11.4	11.6	10.2	11.2
11.8	9.7	7.6	9.1	7.9	9.9	10.8	12.1	10.7	8.9
11.0	8.0	11.3	10.8	11.2	11.2	12.8	11.8	6.6	12.8
12.2	7.5	9.2	12.0	12.4	12.7	10.0	8.0	8.4	9.1
9.8	11.0	12.5	10.6	11.7	8.0	10.8	10.6	8.2	10.0
7.7	10.1	9.5	11.3	11.7	11.6	9.0	11.1	11.0	11.7
10.2	11.7	10.6	10.3	12.2	11.0	10.6	10.5	12.2	8.0
8.2	10.9	11.0	9.8	8.2	10.2	12.2	8.8	11.5	9.7
10.5	11.1	11.8	11.1	9.1	7.7	10.8	10.5	9.3	10.2
10.0	12.4	9.0	11.1	11.6	9.7	9.0	11.0	11.1	11.6
10.5	11.0	7.7	10.5	12.6	9.5	9.8	11.5	11.6	10.8
11.6	11.5	10.8	9.8	11.0	12.0	10.6	16.3	9.8	14.6
10.2	12.2	11.5	11.9	7.8	12.8	12.7	10.8	10.1	10.7
10.5	11.3	10.5	14.5	11.0	11.8	16.0	12.1	13.4	11.2
10.6	12.1	8.0	11.6	10.4	12.8	13.2	12.8	9.6	10.4
11.6	13.0	7.0	12.7	11.3	11.5	11.7	10.5	12.6	14.2
10.4	12.7	8.9	9.6	14.8	11.4	8.4	13.2	10.1	13.2
10.1	11.4	11.0	11.9	13.2	8.4	10.4	11.7	12.1	11.2
10.6	12.7	11.8	12.4	11.6	10.5	13.9	12.1	11.4	9.6
12.8	12.7	12.2	11.5	10.2	10.6	10.7	9.6	11.5	11.3
11.6	11.3	9.8	8.0	12.8	10.5	10.5	7.8	13.2	9.7

Table A.1.1 500 Haemoglobin Values (Cont.)

10.1	16.0	14.5	14.2	15.9	13.0	13.5	14.8	15.4	16.3
12.8	13.2	9.9	9.6	9.4	11.6	9.5	10.0	9.9	11.6
10.4	11.9	12.0	10.2	10.9	10.5	11.8	10.7	10.6	10.8
8.7	12.6	11.6	9.0	11.1	10.2	11.2	7.5	12.5	12.0
11.4	10.8	11.8	8.0	11.2	8.6	10.5	11.9	10.7	9.8
11.4	12.2	11.3	11.1	11.2	7.8	11.4	12.7	13.3	12.0
10.2	7.3	10.1	12.1	9.4	11.9	11.4	10.4	10.0	12.2
12.2	10.7	11.8	11.1	10.0	8.6	10.8	10.0	9.5	8.5
9.3	8.9	12.1	11.9	11.9	10.3	10.2	6.6	9.9	10.6
11.6	8.4	12.0	9.2	9.9	9.6	9.5	10.2	10.2	7.4
9.3	11.0	11.5	11.6	9.8	9.7	11.1	9.1	10.5	10.0
12.4	10.7	9.7	11.4	15.3	9.5	15.9	12.1	11.9	13.9
10.3	13.5	11.7	12.3	13.7	9.9	9.5	12.6	11.7	13.7
12.8	12.7	12.1	10.4	11.9	7.0	10.4	8.9	10.3	10.7
15.3	14.6	12.3	7.4	11.9	11.5	10.2	11.3	11.4	11.6

Appendix 2

Table A.2.1 Random Numbers

Row No.	Col. Nos. 1–10	Col. Nos. 11–20	Col. Nos. 21–30	Col. Nos. 31–40
1	13 50 63 04 23	25 47 57 91 13	52 62 24 19 94	91 67 48 57 10
2	78 66 28 55 80	47 46 41 90 08	55 98 78 10 70	49 92 05 12 07
3	51 57 32 22 27	12 72 72 27 77	44 67 32 23 13	67 95 07 76 30
4	66 86 65 64 60	56 59 75 36 75	46 44 33 63 71	54 50 06 44 75
5	19 83 52 47 53	65 00 51 93 51	30 80 05 19 29	56 23 27 19 03
6	51 78 57 26 17	34 87 96 23 95	89 99 93 39 79	11 28 94 15 52
7	79 68 96 26 60	70 39 83 66 56	62 03 55 86 57	77 55 33 62 02
8	73 52 93 70 50	48 21 47 74 63	17 27 27 51 26	35 96 29 00 45
9	63 99 25 69 02	09 04 03 35 78	19 79 95 07 21	02 84 48 51 97
10	86 28 30 02 35	71 30 32 06 47	93 74 21 86 33	49 90 21 69 74
11	97 96 47 59 97	56 33 24 87 36	17 18 16 90 46	76 27 28 52 13
12	93 41 69 96 07	97 50 81 79 59	42 37 13 81 83	92 42 85 04 31
13	40 24 74 36 42	40 33 04 46 24	35 63 02 31 61	34 59 48 36 96
14	06 06 16 25 98	17 78 80 36 85	26 41 77 63 37	71 63 94 94 33
15	97 81 26 03 89	39 46 67 21 17	98 10 39 33 15	61 63 00 25 92
16	65 99 59 97 84	90 14 79 61 55	56 16 88 87 60	32 15 99 67 43
17	16 91 21 32 41	60 22 66 72 17	31 85 33 69 07	68 49 20 43 29
18	62 03 89 26 32	35 27 99 18 25	78 12 93 09 70	50 93 19 35 56
19	92 27 73 40 38	37 11 05 75 16	98 81 99 37 29	92 20 32 39 67
20	45 51 94 69 04	00 84 14 36 37	95 66 39 01 09	21 68 40 95 79

Table A.2.1 Random Numbers (Cont.)

Row No.	Col. Nos. 1–10	Col. Nos. 11–20	Col. Nos. 21–30	Col. Nos. 31–40
21	00 81 06 28 48	12 08 05 75 26	03 35 63 05 77	13 81 20 67 58
22	05 06 42 24 07	60 60 29 99 93	72 93 78 04 36	25 76 01 54 03
23	12 68 46 55 89	60 09 71 87 89	70 81 10 95 91	83 79 68 20 66
24	07 79 26 69 61	67 85 72 37 41	85 79 76 48 23	61 58 87 08 05
25	52 16 16 23 56	62 95 80 97 63	32 25 34 03 36	48 84 60 36 56
26	16 01 92 58 21	48 79 74 73 72	08 64 80 91 38	07 28 66 61 59
27	89 84 05 34 47	88 09 31 54 88	97 96 86 01 69	46 13 95 65 96
28	51 99 98 44 39	12 45 10 60 36	80 66 39 94 97	42 36 31 16 59
29	08 12 60 39 23	61 73 84 89 18	26 02 04 37 95	96 18 69 06 30
30	69 99 47 26 52	48 06 30 00 18	03 30 28 55 59	66 10 71 44 05
31	88 91 28 79 50	71 42 14 95 55	98 59 96 01 36	88 77 90 45 59
32	59 45 27 08 51	85 64 23 85 41	64 72 08 59 55	67 98 36 65 56
33	84 27 17 30 37	48 69 49 02 58	98 02 50 58 11	95 39 06 35 63
34	65 45 53 41 07	14 83 46 74 11	76 66 63 60 08	90 54 33 65 84
35	48 28 01 83 84	09 11 21 91 73	97 28 44 74 06	22 30 95 69 72
36	93 18 31 83 45	54 52 62 29 91	53 58 54 66 05	47 19 63 92 75
37	18 26 32 96 83	50 58 45 27 57	14 96 39 64 85	73 87 96 76 23
38	45 62 63 40 88	34 69 34 10 94	32 22 52 04 74	69 63 21 83 41
39	26 22 59 28 27	38 88 22 14 79	24 32 12 38 42	33 56 90 92 57
40	33 26 74 03 30	74 22 19 13 48	30 28 01 92 49	58 61 52 27 03
41	80 94 77 47 60	06 25 36 88 56	81 71 02 91 26	03 46 75 07 79
42	37 45 67 66 72	21 95 60 70 36	45 20 45 84 57	14 81 22 50 29
43	08 35 36 41 16	38 34 64 17 67	76 64 62 47 83	62 12 57 52 65
44	16 01 11 43 72	72 39 88 83 56	94 77 58 70 64	25 17 25 97 92
45	77 97 41 21 96	03 57 23 46 30	55 49 05 93 57	61 72 00 72 27
46	76 20 83 10 95	00 71 16 62 59	76 19 37 49 91	21 21 85 93 35
47	20 96 52 70 07	93 39 50 27 22	09 25 57 92 10	40 58 06 96 36
48	68 03 70 16 10	55 63 96 77 54	35 91 26 13 53	25 74 85 24 06
49	35 66 93 10 26	04 86 07 47 65	22 37 12 22 39	47 12 28 27 08
50	64 28 00 86 62	45 32 47 57 07	78 47 62 17 08	57 82 29 56 09
51	86 33 26 17 91	76 96 02 48 07	72 69 61 31 87	80 34 17 72 78
52	66 32 80 56 10	91 91 58 73 03	40 84 90 96 40	47 67 75 08 76
53	26 12 79 36 14	53 48 12 17 42	71 28 36 36 65	61 75 22 03 59
54	94 36 16 81 08	51 34 88 88 15	53 01 54 03 54	56 05 01 45 11
55	04 18 24 87 55	83 90 32 65 07	85 54 03 46 62	51 35 77 41 46
56	68 53 37 57 01	71 59 43 11 45	34 34 01 88 56	65 77 79 71 79
57	83 47 70 44 46	40 68 32 24 45	48 72 78 70 23	35 28 74 93 93
58	51 82 62 63 12	24 98 63 67 51	00 84 88 27 94	79 83 42 00 53
59	92 15 39 92 48	74 80 82 59 59	28 61 45 74 58	13 59 03 71 61
60	55 88 34 56 96	02 52 60 65 68	86 18 03 40 33	81 75 79 63 59

Table A.2.1 Random Numbers (Cont.)

Row No.	Col. Nos. 1–10	Col. Nos. 11–20	Col. Nos. 21–30	Col. Nos. 31–40
61	39 96 04 15 70	15 92 10 09 74	03 19 26 99 80	36 10 90 38 05
62	73 46 94 00 32	92 81 65 32 06	70 35 52 27 73	40 85 15 42 25
63	86 21 41 85 67	27 27 70 12 27	36 96 64 96 48	89 26 97 33 16
64	93 99 55 75 15	62 58 21 98 24	49 09 03 48 87	35 36 04 99 59
65	43 34 03 47 48	93 20 25 55 90	81 26 85 71 25	32 93 55 75 63
66	80 91 05 36 65	22 54 09 14 63	01 38 03 84 67	11 23 84 00 72
67	96 27 33 11 20	10 25 25 31 42	97 00 21 96 40	76 37 10 33 72
68	00 86 35 01 49	16 25 11 12 74	17 75 83 24 96	46 06 11 10 48
69	37 53 01 74 79	34 34 83 92 10	91 63 47 14 78	88 35 77 65 96
70	27 40 32 39 30	54 99 91 37 78	31 95 10 40 20	22 31 93 91 96
71	39 19 68 71 56	85 81 47 73 10	20 80 41 96 33	75 57 00 79 67
72	45 77 78 97 27	57 59 92 73 98	76 87 84 15 15	95 96 36 46 05
73	02 15 72 54 53	78 38 61 34 48	94 94 52 21 13	25 73 17 10 22
74	58 07 79 48 17	74 68 36 17 86	23 92 28 20 85	33 06 29 37 71
75	19 10 96 53 12	14 39 21 52 98	83 34 13 52 71	13 22 91 93 12
76	39 90 13 10 93	38 26 01 55 71	14 24 78 50 45	31 01 33 55 19
77	59 67 89 41 79	87 33 35 75 79	97 35 30 42 84	09 70 53 53 64
78	58 72 11 43 91	83 69 11 22 47	15 59 48 88 71	98 92 49 13 95
79	72 40 18 27 32	81 07 05 44 79	55 98 99 85 81	70 33 67 69 28
80	22 68 42 27 58	44 07 00 69 07	96 68 66 70 00	97 06 86 63 11

Appendix 3

Table A.3.1(a) Random Permutations of 10 Numbers

51246	43985	04628	54670	93108
38097	27160	97531	32918	67542
10783	41237	72168	87194	80472
95426	89605	35904	23506	69351
78059	76901	24618	85102	30642
61432	35482	57093	93647	15987
81973	35719	83924	21593	98432
65042	42086	67501	40678	57601
72896	74312	57346	16248	68102
43105	85960	82019	97305	53794
01325	52839	25301	24150	71306
74896	70614	78649	63987	94258
32516	57819	23461	15894	20935
07498	04623	80957	06723	78641

Table A.3.1(b) Random Permutations of 20 Numbers

5	2	3	14	13	18	9	2	12	16	0	1	18	4	15
18	9	0	12	16	3	10	7	8	13	17	8	13	6	16
10	6	15	17	7	15	14	17	0	6	14	9	12	19	3
8	19	1	11	4	11	19	4	1	5	11	5	2	10	7
7	16	11	19	17	2	11	0	1	7	10	4	12	5	19
5	14	18	9	3	19	13	18	8	6	8	7	14	2	18
0	10	4	8	12	9	15	3	17	16	0	3	9	15	13
15	13	1	6	2	4	12	5	10	14	6	17	1	16	11
17	19	0	14	11	4	10	7	19	2	17	10	3	0	14
3	2	18	8	16	0	16	3	8	6	2	9	18	19	1
4	10	6	12	1	14	11	17	18	15	8	15	6	11	12
7	15	5	9	13	12	9	13	5	1	4	5	13	16	7
2	18	3	8	6	11	1	6	8	9	6	1	10	11	15
4	7	14	0	10	18	12	15	0	19	8	7	2	14	4
19	1	9	11	13	13	2	10	14	5	13	17	0	9	19
15	12	16	5	17	4	3	7	16	17	16	3	12	18	5

Appendix 4

Table A.4.1 Cumulative Normal Distribution

Z	Area	Z	Area	Z	Area	Area	Z
− 3.25	0.0006	− 1.00	0.1587	1.05	0.8531	0.00001	− 4.265
− 3.20	0.0007	− 0.95	0.1711	1.10	0.8643	0.0001	− 3.719
− 3.15	0.0008	− 0.90	0.1841	1.15	0.8749	0.001	− 3.090
− 3.10	0.0010	− 0.85	0.1977	1.20	0.8849	0.005	− 2.576
− 3.05	0.0011	− 0.80	0.2119	1.25	0.8944	0.01	− 2.326
− 3.00	0.0013	− 0.75	0.2266	1.30	0.9032	0.02	− 2.054
− 2.95	0.0016	− 0.70	0.2420	1.35	0.9115	0.25	− 1.960
− 2.90	0.0019	− 0.65	0.2578	1.40	0.9192	0.03	− 1.881
− 2.85	0.0022	− 0.60	0.2743	1.45	0.9265	0.04	− 1.751
− 2.80	0.0026	− 0.55	0.2912	1.50	0.9332	0.05	− 1.645
− 2.75	0.0030	− 0.50	0.3085	1.55	0.9394	0.06	− 1.555
− 2.70	0.0035	− 0.45	0.3264	1.60	0.9452	0.07	− 1.476
− 2.65	0.0040	− 0.40	0.3446	1.65	0.9505	0.09	− 1.341
− 2.60	0.0047	− 0.35	0.3632	1.70	0.9554	0.15	− 1.036
− 2.55	0.0054	− 0.30	0.3821	1.75	0.9599	0.25	− 0.674
− 2.50	0.0062	− 0.25	0.4013	1.80	0.9641	0.35	− 0.385
− 2.45	0.0071	− 0.20	0.4207	1.85	0.9678	0.45	− 0.126
− 2.40	0.0082	− 0.15	0.4404	1.90	0.9713	0.50	0
− 2.35	0.0094	− 0.10	0.4602	1.95	0.9744	0.60	0.253
− 2.30	0.0107	− 0.05	0.4801	2.00	0.9772	0.70	0.524
− 2.25	0.0122			2.05	0.9798	0.80	0.842
− 2.20	0.0139			2.10	0.9821	0.90	1.282
− 2.15	0.0158	0.00	0.5000	2.15	0.9842	0.92	1.405
− 2.10	0.0179			2.20	0.9861	0.94	1.555
− 2.05	0.0202			2.25	0.9878	0.95	1.645
− 2.00	0.0228	0.05	0.5199	2.30	0.9893	0.96	1.751

Table A.4.1 Cumulative Normal Distribution (Cont.)

Z	Area	Z	Area	Z	Area	Area	Z
– 1.95	0.0256	0.10	0.5398	2.35	0.9906	0.97	1.881
– 1.90	0.0287	0.15	0.5596	2.40	0.9918	0.975	1.960
– 1.85	0.0322	0.20	0.5793	2.45	0.9929	0.98	2.054
– 1.80	0.0359	0.25	0.5987	2.50	0.9938	0.99	2.326
– 1.75	0.0401	0.30	0.6179	2.55	0.9946	0.995	2.576
– 1.70	0.0446	0.35	0.6368	2.60	0.9953	0.999	3.090
– 1.65	0.0495	0.40	0.6554	2.65	0.9960	0.9999	3.719
– 1.60	0.0548	0.45	0.6736	2.70	0.9965	0.99999	4.265
– 1.55	0.0606	0.50	0.6915	2.75	0.9970		
– 1.50	0.0668	0.55	0.7088	2.80	0.9974		
– 1.45	0.0735	0.60	0.7257	2.85	0.9978		
– 1.40	0.0808	0.65	0.7422	2.90	0.9981		
– 1.35	0.0885	0.70	0.7580	2.95	0.9984		
– 1.30	0.0968	0.75	0.7734	3.00	0.9987		
– 1.25	0.1056	0.80	0.7881	3.05	0.9989		
– 1.20	0.1151	0.85	0.8023	3.10	0.9990		
– 1.15	0.1251	0.90	0.8159	3.15	0.9992		
– 1.10	0.1357	0.95	0.8289	3.20	0.9993		
– 1.05	0.1469	1.00	0.8413	3.25	0.9994		

Z = deviations from Mean in multiples of σ.

Appendix 5

Table A.5.1 Selected Values of Normal Distribution

(a) Z-values for selected levels of significance (α)

Level of significance (α)	Z-Values	
	Two-tailed test	One-tailed test
0.10	1.645	1.282
0.05	1.96	1.645
0.02	2.326	2.054
0.01	2.576	2.326
0.001	3.291	3.090

(b) Z-values for power of the test ($1-\beta$)

Power ($1-\beta$)	Z_β	$Z_{(1-\beta)}$
0.90	1.282	– 1.282
0.85	1.037	– 1.037
0.80	0.842	– 0.842
0.75	0.675	– 0.675

Appendix 6

Table A.6.1 Table of *t*-Distribution (Two-tailed probability)

Degrees of Freedom (df)	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850

Table A.6.1 Table of *t*-Distribution (Two-tailed probability) (Cont.)

Degrees of Freedom (df)	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.785	3.725
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.126	0.254	0.387	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
100	0.126	0.254	0.386	0.526	0.678	0.846	1.043	1.293	1.660	1.984	2.364	2.641	3.390
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.575	3.290

Appendix 7

Table A.7.1(a) Values of F at the 5% Significance Level

n_2 (df of the denominator)	n_1 (df of the numerator)								
	1	2	3	4	5	6	7	8	9
1	161	200	216	225	230	234	237	239	241
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04

Table A.7.1(b) Values of *F* at the 1% Significance Level

n_2 (df of the denominator)	n_1 (df of the numerator)								
	1	2	3	4	5	6	7	8	9
1	4052	5000	5403	5625	5764	5859	5928	5982	6022
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72

Appendix 8

Table A.8.1 Table of Values of χ^2 Corresponding to Different Levels of Significance

	P					
df	0.50	0.10	0.05	0.02	0.01	0.001
1	0.46	2.71	3.84	5.41	6.63	10.83
2	1.39	4.61	5.99	7.82	9.21	13.82
3	2.37	6.25	7.82	9.84	11.34	16.27
4	3.36	7.78	9.49	11.67	13.28	18.47
5	4.35	9.24	11.07	13.39	15.09	20.52
6	5.35	10.65	12.59	15.03	16.81	22.46
7	6.35	12.02	14.07	16.62	18.48	24.32
8	7.34	13.36	15.51	18.17	20.09	26.13
9	8.34	14.68	16.92	19.68	21.67	27.88
10	9.34	15.99	18.31	21.16	23.21	29.59
11	10.34	17.28	19.68	22.62	24.72	31.26
12	11.34	18.55	21.03	24.05	26.22	32.91

P = probability of getting a larger value of χ^2 by chance alone.

Appendix 9

Table A.9.1 Critical Values of the Correlation Coefficient '*r*'

	Level of significance for two-tailed test			
df	0.10	0.05	0.02	0.01
1	0.988	0.997	0.9995	0.9999
2	0.900	0.950	0.980	0.990
3	0.805	0.878	0.934	0.959
4	0.729	0.811	0.882	0.917
5	0.669	0.754	0.833	0.874
6	0.622	0.707	0.789	0.834
7	0.582	0.666	0.750	0.798
8	0.549	0.632	0.716	0.765
9	0.521	0.602	0.685	0.735
10	0.497	0.576	0.658	0.708
11	0.476	0.553	0.634	0.684
12	0.458	0.532	0.612	0.661
13	0.441	0.514	0.592	0.641
14	0.426	0.497	0.574	0.623
15	0.412	0.482	0.558	0.606
16	0.400	0.468	0.542	0.590
17	0.389	0.456	0.528	0.575
18	0.378	0.444	0.516	0.561
19	0.369	0.433	0.503	0.549
20	0.360	0.423	0.492	0.537

Table A.9.1 Critical Values of the Correlation Coefficient '*r*' (Cont.)

	Level of significance for two-tailed test			
df	0.10	0.05	0.02	0.01
21	0.352	0.413	0.482	0.526
22	0.344	0.404	0.472	0.515
23	0.337	0.396	0.462	0.505
24	0.330	0.388	0.453	0.496
25	0.323	0.381	0.445	0.487
26	0.317	0.374	0.437	0.479
27	0.311	0.367	0.430	0.471
28	0.306	0.361	0.423	0.463
29	0.301	0.355	0.416	0.456
30	0.296	0.349	0.409	0.449
35	0.275	0.325	0.381	0.418
40	0.257	0.304	0.358	0.393
45	0.243	0.288	0.338	0.372
50	0.231	0.273	0.322	0.354
60	0.211	0.250	0.295	0.325
70	0.195	0.232	0.274	0.303
80	0.183	0.217	0.256	0.283
90	0.173	0.205	0.242	0.267
100	0.164	0.195	0.230	0.254

Appendix 10

Table A.10.1 Probabilities (One-tailed) Associated with Values as Small as Observed Values of X in the Binomial Test ($P = 0.5$)

N \ x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	031	188	500	812	969	*										
6	016	109	344	656	891	984	*									
7	008	062	227	500	773	938	992	*								
8	004	035	145	363	637	855	965	996	*							
9	002	020	090	254	500	746	910	980	998	*						
10	001	011	055	172	377	623	828	945	989	999	*					
11		006	033	113	274	500	726	887	967	994	*	*				
12		003	019	073	194	387	613	806	927	981	997	*	*			
13		002	011	046	133	291	500	709	867	954	989	998	*	*		
14		001	006	029	090	212	395	605	788	910	971	994	999	*		
15			004	018	059	151	304	500	696	849	941	982	996	*	*	*
16			002	011	038	105	227	402	598	773	895	962	989	998	*	*
17			001	006	025	072	166	315	500	685	834	928	975	994	999	*
18			001	004	015	048	119	240	407	593	760	881	952	985	996	999
19				002	010	032	084	180	324	500	676	820	916	968	990	998
20				001	006	021	058	132	252	412	588	748	868	942	979	994
21				001	004	013	039	095	192	332	500	668	808	905	961	987
22					002	008	026	067	143	262	416	584	738	857	933	974
23					001	005	017	047	105	202	339	500	661	798	895	953
24					001	003	011	032	076	154	271	419	581	729	846	924
25						002	007	022	054	115	212	345	500	655	788	885

* 1.0 or approximately 1.0.

Appendix 11

Table A.11.1 Critical Values of T in the Wilcoxon Signed Rank Test

n (No. of pairs)	Level of significance (two-tailed)		
	0.05	0.02	0.01
6	0		
7	2	0	
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38

Appendix 12

Table A.12.1 Critical Values of the Rank-Correlation Coefficient (Spearman's rho)—r_s

n	Level of significance (two-tailed)		
	0.05	0.02	0.01
6	0.89	0.94	1.00
7	0.79	0.89	0.93
8	0.74	0.83	0.88
9	0.68	0.78	0.83
10	0.65	0.73	0.79
11	0.61	0.71	0.77
12	0.59	0.68	0.75
13	0.56	0.65	0.71
14	0.54	0.63	0.69
15	0.52	0.60	0.66
16	0.51	0.58	0.64
18	0.48	0.55	0.61
20	0.45	0.53	0.58

Bibliography

Ahlbom, A., and Norell, S. (1990): *Introduction to Modern Epidemiology*. Chestnut Hill (Mass.): Epidemiology Resources.

Altman, D.G. (1992): *Practical Statistics for Medical Research*. London: Chapman and Hall.

Anderson, T.W. (1984): *Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons.

Armitage, P. (1960): *Sequential Medical Trials*. London: Blackwell Scientific Publications.

Armitage, P. (1989): *Statistical Methods in Medical Research*. London: Blackwell Scientific Publications.

Ashish, B., Devendra, B.G., and Gaurisankar, R. (1977): *Population Statistics in India*. New Delhi: Vikas Publishing House.

Atkinson, A.C. (1985): *Plots, Transformations and Regression*. Oxford: Clarendon Press.

Bahn, A.K. (1972): *Basic Medical Statistics*. New York: Grune & Stratton.

Bailey, N.T.J. (1964): *The Elements of Stochastic Process with Applications to the Natural Sciences*. New York: John Wiley & Sons.

Barclay, G.W. (1959): *Techniques of Population Analysis*. New York: John Wiley & Sons.

Bishop, O.N. (1966): *Statistics for Biology*. Boston: Houghton Mifflin.

Bishop, Yvonne, M.M., Fienberg, Stephen, E., and Holland, Paul, W. (1988): *Discrete Multivariate Analysis*. Cambridge (Mass.): The MIT Press.

Bland, M. (1989): *An Introduction to Medical Statistics*. London: ELBS.

Bliss, C.I. (1967): *Statistics in Biology*. New York: McGraw-Hill.

Bourke, G.J., and McGilvary, J. (1985): *Interpretation and Uses of Medical Statistics*. London: Blackwell Scientific Publications.

Bradley, J.V. (1968): *Distribution-Free Statistical Tests*. Englewood Cliffs, New Jersey: Prentice-Hall.

Brockwoll, Peter, J., and Davis, Richard, A. (1987): *Time Series Theory and Methods*. New York: Springer-Verlag.

Castle, W.M. (1972): *Statistics in Small Doses*. London: Churchill Livingstone.

Chatfield, C., and Collins, A.J. (1986): *Introduction to Multivariate Analysis*. London: Chapman and Hall.

Chatfield, C. (1989): *The Analysis of Time Series*. London: Chapman and Hall.

Checkoway, Harvey, Pearce, Neil, E., and Crawford-Brown, Douglas, J. (1989): *Research Methods in Occupational Epidemiology*. New York: Oxford University Press.

Chiang, C.L., (1968): *Introduction to Stochastic Processes in Biostatistics*. New York: John Wiley & Sons.

Cochran, W.G., and Cox, G.M. (1962): *Experimental Designs*. Bombay: Asia Publishing House.

Cochran, W.G. (1977): *Sampling Techniques*. New York: John Wiley & Sons.

Colquhoun, D. (1971): *Lectures on Biostatistics*. Oxford: Clarendon Press.

Colton, T. (1974): *Statistics in Medicine*. Boston: Little Brown & Co.

Conover, W.J. (1980): *Practical Non-Parametric Statistics*. New York: John Wiley & Sons.

Cox, D.R. (1976): *Planning of Experiments*. New York: John Wiley & Sons.

Crowder, M.J., and Hand, D.J. (1990): *Analysis of Repeated Measures*. London: Chapman and Hall.

Croxton, F.E. (1953): *Elementary Statistics with Applications in Medical and Biological Sciences*. New York: Dover.

Croxton, F.E., and Cowden, D.J. (1964): *Applied General Statistics*. New Delhi: Prentice-Hall of India.

Daniel, W.W. (1987): *Biostatistics: A foundation for analysis in the health sciences*. New York: John Wiley & Sons.

Das, M.N., and Giri, N.C. (1979): *Design and Analysis of Experiments*. New Delhi: Wiley Eastern.

Diggle, Peter, J. (1990): *Time Series: A biostatistical introduction*. New York: Oxford University Press.

Dixon, W.J., and Massey, F.J. (1969): *Introduction to Statistical Analysis*. New York: McGraw-Hill

Draper, N.R., and Smith, H. (1981): *Applied Regression Analysis*. New York: John Wiley & Sons.

Dunn, O.J. (1984): *Basic Statistics*. New York: John Wiley & Sons.

Dunn, O.J., and Clark, V. (1987): *Applied Statistics*. New York: John Wiley & Sons.

Dwyer, James, H. (1991): *Statistical Models for Longitudinal Studies of Health*. New York: Oxford University Press.

Elandt-Johnson, Regina, C. (1971): *Probability Models and Statistical Methods in Genetics*. New York: John Wiley & Sons.

Elston, R.C., and Johnson, W.D. (1987): *Essentials of Biostatistics*. Philadelphia: F.A. Davis Co.

Everitt, B.S. (1992): *The Analysis of Contigency Tables*. London: Chapman and Hall.

Federer, W. (1967): *Experimental Design*. New Delhi: Oxford & IBH.

Feller, W. (1969): *An Introduction to Probability Theory and Its Applications*. Vols. I and II. New Delhi: Wiley Eastern.

Finney, D.J. (1978): *Statistical Methods in Biological Assay*. London: Charles Griffin & Co.

Fleiss, Joseph, L. (1981): *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons.

Fleiss, Joseph, L. (1986): *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons.

Flury Bernard, and Riedwyl, Hans. (1988): *Multivariate Statistics: A practical approach*. London: Chapman and Hall.

Gallant, Ronald, A. (1987): *Nonlinear Statistical Models*. New York: John Wiley & Sons.

Gilbert, Neil (1989): *Biometrical Interpretation*. New York: Oxford University Press.

Gnanadesikan, R. (1977): *Methods for Statistical Data Analysis of Multivariate Observations*. New York: John Wiley & Sons.

Goldstein, H. (1987): *Multilevel Models in Educational and Social Research*. London: Charles Griffin & Co.

Goldstone, L.A. (1985): *Understanding Medical Statistics*. London: ELBS.

Govindarajulu, Z. (1988): *Statistical Techniques in Bioassay*. New Delhi: Karger.

Hamilton, M. (1961): *Lectures on the Methodology in Clinical Research*. Edinburgh and London: E. & S. Livingstone.

Hand, D.J., and Taylor, C.C. (1987): *Multivariate Analysis of Variance and Repeated Measures*. London: Chapman and Hall.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953): *Sample Survey Methods and Theory*. New York: John Wiley & Sons.

Hastie, T.J., and Tibshirani, R.J. (1990): *Generalized Additive Models*. London: Chapman and Hall.

Healy, M.J.R. (1990): *GLIM: An Introduction*. New York: Oxford University Press.

Hill, A.B. (1962): *Statistical Methods in Clinical and Preventive Medicine*. Edinburgh: Livingstone.

Hill, A.B. (1981): *Principles of Medical Statistics*. London: Edward Arnold.

Holland, W.W. (1970): *Data Handling in Epidemiology*. New York: Oxford University Press.

Hooke, Robert, (1983): *How to Tell the Liars from the Statisticians*. New York: Marcel Dekker.

Hosmer, David, W., and Lemeshow, S. (1989): *Applied Logistic Regression*. New York: John Wiley & Sons.

Ingelfinger, T.A. (1983): *Biostatistics in Clinical Medicine*. New York: Macmillan.

Johnson, Richard, A., and Wichern, Dean W. (1988): *Applied Multivariate Statistical Analysis*. London: Prentice-Hall International.

Jolliffe, I.T. (1986): *Principal Component Analysis*. New York: Springer-Verlag.

Jones, Byron, and Kenward, Michael, G. (1990): *Design and Analysis of Crossover Trials*. London: Chapman and Hall.

Kahn, Harold, A. (1983): *An Introduction to Epidemiological Methods*. New York: Oxford University Press.

Kempthorne, C. (1966): *Introduction to Genetic Statistics*. New York: John Wiley & Sons.

Kendall, M., and Gibbons, J.D. (1990): *Rank Correlation Methods*. London: Edward Arnold.

Kendall, M., and Ord, J.K. (1990): *Time Series*. London: Edward Arnold.

Keyfitz, Nathan (1985): *Applied Mathematical Demography*. New York: Springer-Verlag.

Kiefer, Jack, C. (1987): *Introduction to Statistical Inference*. New York: Springer-Verlag.

Kirkwood, B.R. (1988): *Essentials of Medical Statistics*. London: Blackwell Scientific Publications.

Kish, Leslie, (1987): *Statistical Design for Research*. New York: John Wiley & Sons.

Kleinbaum, David, G., Kupper, Lawrence, L., and Morgenstern, Hall (1982): *Epidemiologic Research Principles and Quantitative Methods*. Life Time Learning Publications. New York: Van Nostrand Reinhold.

Krzanowski, W.J. (1990): *Principles of Multivariate Analysis: A user's perspective*. New York: Oxford University Press.

Lancaster, H.O. (1974): *Introduction to Medical Statistics*. New York: John Wiley & Sons.

Lilienfeld, A.M., and Lilienfeld, D.E. (1980): *Foundations of Epidemiology*. New York: Oxford University Press.

Little, Michael, A., and Hall, Jere, D. (1989): *Human Population Biology*. New York: Oxford University Press.

Lutkepohl, Helmat (1991): *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.

Lwanga, S.K., and Lemeshow, S. (1991): *Sample Size Determination on Health Studies*. Geneva: WHO.

Mainland, D. (1963): *Elementary Medical Statistics*. London: W.B. Saunders.

Manly, Bryan, F.J. (1986): *Multivariate Statistical Methods*. London: Chapman and Hall.

Mathews, D.E. (1988): *Using and Understanding Medical Statistics*. New Delhi: Karger.

Maxwell, A.E. (1977): *Multivariate Analysis in Behavioural Research*. London: Chapman and Hall.

McCullagh, P., and Nelder, T.A. (1990): *Generalized Linear Models*. London: Chapman and Hall.

Miettenen, Olis (1985): *Theoretical Epidemiology*. New York: John Wiley & Sons.

Montgomery, D.C. (1984): *Design and Analysis of Experiments*. New York: John Wiley & Sons.

Mood, A.M., and Graybill, F.A. (1974): *Introduction to Theory of Statistics*. New York: McGraw-Hill.

Moolgavkar, S.H., and Prentice, R.L. (1986): *Modern Statistical Methods in Chronic Disease Epidemiology*. New York: John Wiley & Sons.

Moroney, M.J. (1951): *Facts from Figures*. Baltimore: Penguin Books.

Morrison, Donald, F. (1984): *Multivariate Statistical Methods*. Singapore: McGraw-Hill.

Moser, C.A., and Kalton, Graham (1989): *Survey Methods in Social Investigations*. Hants (UK): Gower Publishing Group.

Murray, J.D. (1989): *Mathematical Biology*. Berlin: Springer-Verlag.

Murthy, M.N. (1977): *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

Oakes, Michael (1990): *Statistical Inference*. Chestnut Hill (Mass.): Epidemiology Resources.

Rao, C.R. (1952): *Advanced Statistical Methods in Biometric Research*. New York: John Wiley & Sons.

Rao, C.R. (1974): *Linear Statistical Inference and Its Applications*. New Delhi: Wiley Eastern.

Rees, D.G. (1989): *Essential of Statistics*. London: Chapman and Hall.

Regier, Mary, H. (1985): *Biomedical Statistics with Computing*. New York: Research Studies Press (A Division of John Wiley & Sons).

Reid, Norma, G. (1987): *Research Methods and Statistics in Health Care*. London: Edward Arnold.

Rohatgi, V.K. (1985): *An Introduction to Probability Theory and Mathematical Statistics*. New Delhi: Wiley Eastern.

Rothman, Kenneth, J. Ed. (1988): *Causal Inference*. Chestnut Hill (Mass): Epidemiology Resources.

Sathe, P.V., and Sathe, A.P. (1991): *Epidemiology and Management for Health Care for All*. Bombay: Popular Prakashan.

Saxena, P.C., and Talwar, P.P. (1987): *Recent Advances in the Techniques for Demographic Analysis*. Bombay: Himalaya Publishing House.

Schlesselman, James, J. (1982): *Case Control Studies*. New York: Oxford University Press.

Seal, H.L. (1968): *Multivariate Statistical Analysis for Biologists*. London: Methuen.

Selvin, S. (1991): *Statistical Analysis of Epidemiological Data*. New York: Oxford University Press.

Sen, Ashish, K., and Srivastava, V. (1990): *Regression Analysis*. New York: Springer-Verlag.

Siegel, S. (1988): *Non-Parametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Singh, Daroga, and Choudhary, F.S. (1986): *Theory and Analysis of Sample Survey Designs*. New Delhi: Wiley Eastern.

Snedecor, G.W., and Cochran, W.G. (1967): *Statistical Methods*. Calcutta: Oxford & IBH.

Sokal, R.R., and Rohlf, F.J. (1969): *Biometry*. San Francisco: W.H. Freeman.

Sprent, Peter (1989): *Applied Nonparametric Statistical Methods*. London: Chapman and Hall.

Stibitz, G.R. (1966): *Mathematics in Medicine and the Life Sciences*. Chicago: Year Book Publishers.

Stilson, D.W. (1971): *Probability and Statistics in Psychological Research and Theory*. Bombay: Vakils, Feffer & Simons.

Swaroop, S. (1960): *Introduction to Health Statistics*. Edinburgh: E.&S. Livingstone.

Taylor, T.R. (1967): *The Principles of Medical Computing*. London: Blackwell Scientific Publications.

Tippett, L.H.C. (1968): *Statistics*. London: Oxford University Press.

Tuckwell, Henry, C. (1988): *Elementary Applications of Probability Theory*. London: Chapman and Hall.

Underwood, B.J., Duncan, C.P., Taylor, J.A., and Cotton, J.M. (1954): *Elementary Statistics*. New York: Appleton-Century-Crofts.

Valanis, Barbara (1986): *Epidemiology in Nursing and Health Care*. Norwalk: Appleton-Century-Crofts.

Weisberg, S. (1985): *Applied Linear Regression*. New York: John Wiley & Sons.

Whitehead, J. (1983): *The Design and Analysis of Sequential Clinical Trials*. New York: John Wiley & Sons.

Witts, L.J. (1964): *Medical Surveys and Clinical Trials*. London: Oxford University Press.

Woolson, Robert, F. (1987): *Statistical Methods for the Analysis of Biomedical Data*. New York: John Wiley & Sons.

World Health Organization (WHO) (1992): *International Statistical Classification of Diseases and Related Health Problems (ICD-10)*, Tenth Revision, Vols. 1–3, Geneva.

Yates, F. (1981): *Sampling Methods for Censuses and Surveys*. New York: Charles Griffin & Co.

Young, P.V. (1968): *Scientific Social Surveys and Research*. New Delhi: Prentice-Hall of India.

TABLES

Fisher, R.A., and Yates, F. (1957): *Statistical Tables for Biological, Agricultural and Medical Research*. London: Oliver & Boyd.

Hald, A. (1952): *Statistical Tables and Formulas*. New York: John Wiley & Sons.

Kokoska, Stephen, (1989): *Statistical Tables and Formulae*. New York: Springer-Verlag.

Pearson, E.S., and Hartley, H.O. (1962): *Biometrika Tables for Statisticians*. Vol. I. London: Cambridge University Press.

Rao, C.R., Mitra, S.K., and Matthai, A. (1966): *Formulae and Tables for Statistical Work*. Calcutta: Statistical Publishing Society.

JOURNALS

American Statistician. American Statistical Association, Washington (DC).

Applied Statistics (*Series C*). Royal Statistical Society, London.

Biometrics. The Biometric Society, Tucson, Arizona.

Biometrika. Biometrica Trust, University College, London.

Demography India. Indian Association for the Study of Population, Hindustan Publishing Corporation, Delhi.

Journal of American Statistical Association. American Statistical Association, Alexandria.

Population and Development Review. Center for Policy Studies of the Population Council, New York.

Royal Statistical Society, Series B (Methodological). The Royal Statistical Society, London.

Sample Registration Bulletin. Registrar General, Ministry of Home Affairs. New Delhi.

Statistics in Medicine. John Wiley & Sons, New York.

Vital and Health Statistics. Public Health Service, Department of Health, Education and Welfare, Washington (DC).

World Health Statistics. World Health Organisation, Geneva.

WEBSITE

Population Reference Bureau: World Population Data Sheet

Census of India 2011

Registrar General of Birth and Death: India

Answers to Selected Questions

Chapter 5

4.

Hb levels (g%)	No. of children
9.0– 9.9	6
10.0–10.9	5
11.0–11.9	8
12.0–12.9	4
13.0–13.9	2
Total	25

Note: A frequency distribution can be formed for the given data in many ways. This is one correct answer.

Chapter 6

Q4. Mean = 204 Median = 210 Mode = 160, 210, 230

Note: There are three modes for the data.

Chapter 7

4. Range: 2.5 to 3.5, Mean = 3.0, Median = 2.95, Standard deviation = 0.368, Coefficient of variation = 14.72%

5. Protein supplement group Mean = 12.0, Standard deviation = 0.914
Normal diet group Mean = 11.0, Standard deviation = 1.019

Chapter 8

4(a). P (normal birth wt.) = 0.7

4(b). P (Hypertension or anaemia) = 0.65
P (Hypertension and anaemia) = 0.105

Chapter 12

7. Birth weight: Rural infants: Mean = 2.89, SD = 0.166, SE of mean = 0.053
Urban infants: Mean = 3.17, SD = 0.245, SE of mean = 0.078
Test statistic $t = 2.989$, $df = 18$, $p = 0.008$.
Inference: Birth weight of rural infants are different from the urban infants.
Conclusion: The birth weight of rural infants is significantly ($p = 0.008$) lower than the urban infants.

Chapter 13

4. Pearson's correlation coefficient, $r = 0.906$, $p = 0.000$
   The correlation is significant.

Chapter 14

4. Chi-square = 3.857, $df = 1$, $p = 0.0495$
   Inference: Malnutrition and area of residence are significantly associated ($p = 0.0495$).
   Concluion: Rural children are more malnourished than urban children.

Chapter 16

4. Spearman's rank correlation coefficient = 0.855, $p = 0.002$
   Inference: The rank correlation is significant.
   Conclusion: Both the nutritionists agree in their judgments.

Chapter 17

2. Neonatal mortality rate = 9.26 per 1000 live births
   Post neonatal mortality rate = 28.04
   Infant mortality rate = 92.6 per 1000 live births

Chapter 18

4. Incidence = 0.67 per 1000 population
   Prevalence = 126.67 per 1000 population

Chapter 23

1. Sample size needed, $n = 372$
   (Assumptions: The investigator allows an error of 20% of the proportion.)
2. Sample size needed is 94 **for each group**.
   (Assumptions: Type I error 5% and power of the test 80%)
3. Sample size needed, $n = 42$
   (Assumptions: The investigator allows an error of 10% of the mean. Use 95% confidence interval.)
4. Sample size needed is 15 **for each group.**
   (Assumptions: Type I error 5% and power of the test 80%)

Index

ANOVA
  assumptions of, 113
  illustration of, 109
Arithmetic mean, 26
Association, test for, 101
Average, 26

Bayes' theorem, 47, 49
Bias
  controlling for, 213
  in observational study, 212
Binomial distribution, 55
Biostatistics, the term, 1
Box plot, 239

Causes of death, survey of, 126
Census, 166
Centile, 33
Chi-square test, 100
Classification
  of diseases, 147
  of healthy and sick, 143
  single, 192
  two-way, 192
Clinical trials, 195
Coefficient of variation, 42
Collection of sickness data, problems in, 142
Comparison (large samples)
  of means of two samples, 78
  of sample mean with population mean, 77
  of sample proportion with population proportion, 79
  of two sample proportions, 80
Comparison (small samples)
  of means of two correlated samples, 85
  of means of two independent samples, 83
  of sample mean with population mean, 82
Computer, 219
  Epi-6, 221
  Epi-2000, 221
  Epi-2002, 221
  hardware, 219
  software, 220
  spreadsheets, 221
  SPSS, 221
  STATA, 221
  system, 222
  uses of, 222
Confidence interval, 72, 86, 215
Controls, 190
Correlation, 90
  linear, 91
  multiple, 98
  partial, 98
  Pearson's, 92
  Spearman's rank coefficient of, 120
  spurious, 97
Correlation coefficient, standard error of, 94
Covariance, 93
  analysis of, 194
Critical ratio, 76
Crude birth rate, 135
Crude death rate, 129
Crude rate, 127
Curve
  cumulative frequency polygon, 24
  frequency, 24
  ROC, 51

Data, collection of, 183
Data management, 219, 222
Data processing, use of computer in, 222
Decennial rate, 129
Demography, 166
Dependency ratio, 169
Deviation
  mean, 37
  standard, 38
Diagrams, 21, 229
Diseases, international classification of, 147
Distribution
  binomial, 55
  characteristics of, 25
  chi-square, 101

'*F*', 61, 108
normal, 58
Poisson, 57
'*t*', 60, 83
Distribution-free statistics, 114

Errors
non-sampling, 64, 191
type I and type II, 74
Estimate
interval, 73
point, 72
Estimation, 72, 76, 86
of experimental errors, 191
of sample size, 201
Ethics, 188, 199
Experimental design, 189
general principles of, 189
use of 'controls' in, 190

Factorial design, 193
False negative rate, 50
False positive rate, 50
Fertility rate, 136
age-specific, 136
Fisher's exact probability test, 116
Foetal death ratio, 134
Frequency
curve, 23
distribution, 20
polygon, 22
probability, 45
F-test, 108

General fertility rate, 136
Goodness-of-Fit, test for, 104
Gross reproduction rate, 137

Half-yearly rate, 129
Health statistics, 140
sources of, 141
Histogram, 22
Hospital statistics, 146
Hypothesis, testing of, 3, 76

Inter-quartile range, 36

Kolmogorov–Smirnov test, 115
Kruskal–Wallis test, 118

Latin squares, 193
Level of measurement, 18
Life tables, 159
application to health problems, 162
construction of, 159
structure of, 160
Location, measures of, 26

Mann-Whitney U-test, 115
McNemar test, 116
Mean
arithmetic, 26
geometric, 31
harmonic, 32
Mean deviation, 37
Measurement of sickness, 144
Median, 28
Methods of dealing with non-response, 185
Mode, 29
Monthly rate, 129
Morbidity indices, 144
Mortality
determinants of (variations in), 171
figure, comparative, 157
Mortality rate, 129
infant, 132
maternal, 135
neonatal, 133
perinatal, 134
post-neonatal, 133
specific death rate, 130

Net reproduction rate, 137
Non-parametric statistical tests, 114
Normal distribution, 58
Null hypothesis, 76, 77
Number needed to treat, 216

Observation
qualitative, 14
quantitative, 15
Observational studies, 208
Odds ratio, 21

Percentile, 33
Poisson distribution, 57
Population estimation, 172
arithmetic progression method of, 173
geometric progression method of, 174
Population growth, 166
Power-efficiency, 120

Predictive value
negative, 50
positive, 50
Presentation of data
basic principles of, 225
classification, 226
graphs and diagrams, 229
space and attribute comparisons, 232
table, 225
time comparison, 230
Probability
applications of, 51
Bayes' theorem, 47
classical, 45
conditional, 46
distribution, 55
frequency, 45
laws (Addition, multiplication), 46
measurement of, 45
scale, 44
Problems in collection of sickness data, 142
Prophylactic trials, 198
Prospective studies, 209
advantages and disadvantages of, 214
Pyramid, population, 168

Quarterly rate, 129
Quartile, 33
Questionnaires and record forms, 184
Quinquennial rate, 128

Random allocation, 196
Random numbers, 66
permutation of, 244
table of, 24
Randomisation, 191
Range, 35
Registration system
civil, 124
model, 126
sample, 125
for vital events, 124
Regression, 91
coefficient, 95
equation, 95
linear, 95
multiple, 97
non-linear, 97
Replication, 191
Research
methods, 176
protocol, 177
question, 176

Residual, 110
Risk
attributable, 215
population attributable, 216
relative, 210

Sample, 63
Sample size, determination of, 201
Sampling, 62
distribution, 71
techniques, 65
variation and bias, 64
Scatter diagram, 90
Sensitivity, 49
Sign test, 117
Significance tests, 76
Specificity, 49
Specific rate, 128
Standard deviation, 35, 38
Standard error of
correlation coefficient, 92
difference between means, 78
difference between proportions, 80
mean, 77
proportion, 79
regression coefficient, 96
Standardised death rates, 153
Statistical inference, 71
Statistical methods, the need in clinical medicine, 7
Statistics
areas of application in, 8, 11
in preventive medicine, 11
Stem and leaf plot, 239
Still-birth rate, 134
Study
cross-sectional, 209
interventional, 189
observational, 208
prospective, 209
retrospective, 210

Tabulation, purpose of, 20
Tests of significance, 76
examples for large samples, 77
examples for small samples, 81
procedure, 76

Theoretical distribution, 55
Theoretical framework, 177
Therapeutic trial, 196
Total fertility rate, 137
Type I error, 74
Type II error, 74

Variables, qualitative and quantitative, 14, 15
Variance, 38
  analysis of, 108
  ratio, 108
Variation, coefficient of, 42
Vital statistics, 122
  basic formulae for calculation of, 127
  definition of, 122
  mechanism for collection of, 123
  uses of, 122

Weekly rate, 129
Wilcoxon's signed-rank test, 117

Yates correction, 105