Biostatistics

Basic and Advanced

Manju Pandey

MV Learning

Biostatistics

Basic and Advanced

Manju Pandey

MVLearning
London • New Delhi

Every possible effort has been made to ensure that the information contained in this book is accurate at the time of going to press, and the publisher and author cannot accept responsibility for any errors or omissions, however caused. No responsibility for loss or damage occasioned to any person acting, or refraining from action, as a result of the material in this publication can be accepted by the editor, the publisher or the author.

Every effort has been made to trace the owners of copyright material used in this book. The author and the publisher will be grateful for any omission brought to their notice for acknowledgement in the future editions of the book.

MV Learning

3, Henrietta Street
London WC2E 8LU
UK

4737/23, Ansari Road,
Daryaganj, New Delhi 110 002
India

A Viva Books imprint

ISBN: 978-81-309-2753-4

Printed and bound in India.

SUB-0030, SSUB-0072

Biostatistics

Basic and Advanced

Dedicated to My Loving Father

Dadaji, author of many still read books, I could ultimately fulfill your repeatedly expressed desire of my writing books instead of research papers.

Preface

It is well-known that the foundation and development of the statistical theory came up with a search for interpretation of results of research experiments and, hence, decision making in biological, agricultural and medical fields. This endeavour gave birth to a new stream of mathematical sciences named as "Statistics". The statistical concepts are independent of various subject areas, since the procedures developed for agricultural or biological experiments are equally applicable in industrial, technical, psychological, medical, educational, social and commercial fields.

This led to recognition of need for formal courses on statistical methods in biological, medical and agricultural fields and they were started in UK and USA around the end of 19th and beginning of 20th century. In India, the foundation of development in agriculture, food and nutrition, population control and human health was laid soon after. The importance of biostatistics is now deeply felt all over India. A few decades ago this course was introduced in various life science, earth science, education and technology degree syllabi at Masters level also, in most of the Universities and Institutes. Some of these institutions are offering short term and degree courses in Biostatistics.

Explaining the simplest concepts of Statistics, which stands on the foundation of mathematics, to students belonging to non-mathematics fields, is an arduous task. The present book aims to provide basic knowledge and applications of Biostatistics to Masters level students pursuing courses in different disciplines and also to help young researchers in solving their problems. This book effectively summarizes the author's many years of teaching and research experience. While herself possessing a pure mathematics background, about three-fourth of the author's professional career has been devoted to teaching biostatistics to Masters level students of Zoology with non-mathematics background. The author has also drawn ideas from her interaction with researchers of a variety of disciplines such as education, social science, life sciences, agricultural sciences, medical sciences, earth sciences and engineering.

The book covers almost all parts of the recent UGC Model Syllabus in biological sciences relating to Quantitative Zoology. Each chapter contains relevant figures and numerical examples to help clarify the concepts. The book explains how to choose an appropriate statistical analysis for various biological experimental data. The computational steps and interpretation of the findings are also given in detail. Some of the statistical techniques, often used by researchers of all the life sciences disciplines as well as those belonging to social sciences, psychological sciences, earth sciences and education are also included, briefly, if not in detail.

This book may be used as a textbook by the undergraduate and postgraduate students of Biostatistics in biological, agricultural and health sciences. It will help the researchers in these and other aforesaid disciplines, after acquiring sound knowledge of basic concepts of statistical techniques, to choose advanced topics for a proper analysis of their research data and draw valid conclusions.

This book is divided into two parts: Part I on "Basic" consists of chapters 1 to 11 and Part II on "Advanced" consists of chapters 12 to 23. The book contains solved examples to illustrate the computational steps of various statistical procedures. Tables and Figures supplement the text for clear illustration of the theory behind the procedures and their applications.

The material in the book assumes the readers to have the knowledge of High School algebra and arithmetic in order to understand the logic behind the statistical procedures. Applications of these on experimental data require simple calculators for computations. However, if a researcher is handling huge project data, easy access to computer systems enables large or repeated computations. Therefore, two chapters (Chapters 11 and 23) are included, which provide a fundamental idea of computer systems, methods of using these for common simple problems and also give some idea of using statistical softwares.

Abbreviations used in the book are listed in the beginning. For performing various tests of significance and discussing some advanced topics, some Statistical and Mathematical Tables are required. These are covered under Appendix A. Many students and researchers of life sciences may be unfamiliar with various mathematical symbols and expressions, use of which could not be avoided in the book. Therefore, Appendix B is included to define and explain these. It becomes very easy to understand concepts of probability theory using Venn diagram and rules of Set Theory. Appendix C on Matrix Algebra contains the results (without proof) which enable the logic behind complex multivariate techniques and a description of these procedures in an easy and concisely presentable form. Appendix D provides meaning of the rules and principles of Set Theory used in the chapter on Probability.

Now a brief description of the various chapters is provided for the reader. Chapter 1 is introductory. It defines Statistics and Biostatistics, types of data that arise for analysis of experiments and a brief history of growth of statistical theory and practice in India and abroad. Chapter 2 gives the method of summarising and presenting data in tables and graphs. Chapter 3 defines and determines the descriptive statistics viz. measures of location, dispersion, skewness and kurtosis. Chapter 4 discusses the concepts of probability and Chapter 5 defines the random variables and associated functions giving their nature and characteristics. The concepts of probability distributions, expected values, raw and central moments of continuous and discrete random variables are also given in this chapter.

Chapter 6 discusses the distributions of various continuous and discrete random variables which are extensively used in data analysis in almost all the fields. Chapter 7 deals with the fundamental theory behind problems of estimation of parameters and testing of hypotheses. The criteria of good estimators, concepts of two types of error and choice of rejection region are also explained. Chapter 8 gives the one and two sample procedures of testing means and variances. Equality of more than two means and variances are also considered in this chapter.

Chapter 9 deals with measuring and testing relationships in bivariate and multivariate data in terms of correlation and regression. Chapter 10 deals with analysis of categorical data by establishing independence and measuring association in 2 × 2 and $r \times c$ contingency tables. Chapter 11 considers data handling electronically, indicating components and types of computers, machine and high level computer languages. Basic ideas of working on DOS, Windows, MS Office and computer network are also given. Chapter 12 illustrates the methods of planning medical and biological studies in lab and field, while Chapter 13 considers complete enumeration or census (of human and animal populations) and commonly used sampling methods in brief.

Chapter 14 gives methods of comparison of two means when the commonly assumed equality of two variances does not hold. Subsequently analysis of variance in various designs, such as randomized block, Latin square, nested design, BIBD, PBIBD and factorial experiments have been described. Lastly, multiple comparison methods, if hypothesis of equality of means is rejected, have been given.

Chapters 15, 16 and 17 consider one sample, two sample and k-sample non-parametric tests when the distribution, from which the samples have come, may not be known. Chapter 18 defines components of time series and provides procedure for analysis of these under conditions of being stationary or autocorrelated. Chapter 19 considers the problem and analysis of bioassays.

Chapters 20 and 21 deal with analysis of multivariate data. The Hotelling's T^2, Mahalonobis D^2, Discriminant Analysis, Principal Component Analysis and Cluster Analysis have been explained with examples of real data and use of statistical packages.

Chapter 22 deals with concepts of bioinformatics and some procedures of sequence and microarray data analysis. Chapter 23 on Computer Techniques gives some ideas of programming in FORTRAN, C and C^{++} for use of enthusiastic readers willing to develop their own program for simple but repeatedly required computations. For other problems a brief description of three standard statistical software packages (SPSS, BMDP and SAS) with illustration of their use is included.

In the course of preparation of the manuscript some of the standard books on Statistics and Biostatistics were found to be very helpful, especially the works by Steel and Torrie; Milton and Tsokos; Daniel; Zar; Sokal and Rohlf; Kramer; Finney; Mood and Graybill; Meyer; Yule and Kendall; Croxton, Cowden and Klein; Cochran and Cox; Anderson; Press; Dillon and Goldstein; Hohn; Sukhatme and Sukhatme; Goon, Gupta and Dasgupta; Gupta and Kapoor, to name a few.

The author gratefully acknowledges some of these authors and their publishers for granting permission to use the data and figures in their books for illustrating the application of statistical tools. The statistical and mathematical tables required for completing various statistical testing procedures have been included from some of the said books and from the Formula and Tables for Statistical Work by Rao, Mitra, Mathai and Ramamurthy.

The infrastructural and administrative support extended by the respective Head of the Departments of Zoology and Statistics, Banaras Hindu University in this endeavour, is highly appreciated.

The author very respectfully acknowledges her highly esteemed Ph.D. supervisor Prof. J. Singh, former Vice Chancellor, Vikram University, Ujjain, India for his painstaking efforts in reading many of the book's chapters and for giving his expert suggestions to improve the original draft.

The author is also grateful to Late Prof. M.S. Kanungo and Prof. B.N. Singh for devoting their valuable time to read some of the initial chapters and for their valuable suggestions to make the text more comprehensible.

Critical comments by Dr. G.K. Shukla, former Professor, Department of Mathematics, Indian Institute of Technology Kanpur, and former President, International Biometric Society (Indian Region), on Chapters 12 and 14 were very useful in giving final form to the same. The author deeply expresses her regards to him.

Mr. P.K. Chakravorty, Maintenace Engineer, Computer Center; Prof. K.K. Shukla; Prof. A.K. Agrawal (both from Department of Computer Engineering, Institute of Technology, BHU); Dr. Rajnikant Mishra and Dr. A.K. Maurya, my colleagues in the department, helped by extending their critical advice on technical points, respectively, pertaining to Chapters 11, 22, 23 and a part of Chapter 13. The author expresses her sincere thanks to all for their expert suggestions.

Last but not the least, the elegant typing and preparation of the figures by Dr. Vinod Kumar Upadhyay is sincerely and thankfully acknowledged. Without his patient and meticulous work, this book may not have taken this form. High appreciation is accorded to Mr. P.K. Sinha for his willing cooperation in speedy mathematical typing of Part II of the book.

The author hopes that the good presentation of the book and its well-arranged contents will make for useful reading not only to researchers in biological sciences, but also to researchers in various other application areas like education, social sciences, life sciences, agricultural sciences, medical sciences, earth sciences and engineering etc.

The basic bio-statistical concepts in Part I, as well as the new and advanced topics contained in Part II of this book will remain useful for all times to come. It is, therefore, hoped that the book will prove useful as a textbook for Graduate and Postgraduate students as well as a reference book for researchers.

MANJU PANDEY

Acknowledgements

The author thankfully acknowledges the Department of Science and Technology (DST), Ministry of Science and Technology, Government of India, New Delhi, India for extending financial support towards preparation of this book under Utilisation of the Scientific Expertise of Retired Scientists (USERS) Scheme vide Grant No. HR/UR/15/2006 dated 26 October 2006. The financial support extended by DST has made it possible to complete the book well in time.

Abbreviations

a	Adenine
ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
AIC	Akaike Information Criterion
ALU	Arithmetic/Logical Unit
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
ANSI	American National Standard Institute
AR	Attributable Risk
AR(q)	Autoregressive Model of Order q
ARIMA	Autoregressive Integrated Moving Average
ARMA	Auto Regressive Moving Average
ASCII	American Standard Code for Information Interchange
BASIC	Beginners' All purpose Symbolic Instruction Code
BIBD	Balanced Incomplete Block Design
BJ	Box-Jenkins
BLAST	Basic Local Alignment Search Tool
BLOSUM	BLOcks SUbstitution Matrix
BLUE	Best Linear Unbiased Estimator
BMDP	BioMeDical Program
c	Cytosine
CATH	Class Architecture, Topology, Homologous Super-family
CBR	Crude Birth Rates
CD	Compact Disks
cdf	Cumulative Distribution Function, $F(x) = P[X \leq x]$
CF	Characteristic Function
C.F.	Correction Factor
CI	Confidence Interval
C.I.	Class Interval
COBOL	Common Business Oriented Language
CPU	Central Processing Unit
c.q.d.	Coefficient of Quartile Deviation
CRD	Completely Randomized Design
CRT	Cathode Ray Tube

C.V.	Coefficient of Variation
DDBJ	DNA Data Bank of Japan
d.f.	Degrees of Freedom
DF	Dickey Fuller
DNA	Deoxyribo Nucleic Acid
DoE	Design of Experiment
DS	Difference Stationary
EBCDIC	Extended Binary Coded Decimal Interchange Code
ED	Effective Dose
EDSAC	Electronic Delay Storage Automatic Calculator
EDVAC	Electronic Discrete Variable Automatic Computer
EF	Etiologic Function
EMBL	European Molecular Biology Laboratory
ENIAC	Electronic Numerical Integrator and Calculator
ExPASy	Expert Protein Analysis System
FA	Factor Analysis
FORTRAN	FORmula TRANslation
fpc	Finite Population Correction
ftp	File Transfer Protocol
g	Guanine
GB	Giga Byte
Gbp	Giga Base Pairs
GHz	Giga Hertz
GUI	Graphics User Interface
Hb	Hemoglobin
Huges	Human Genome Equivalents
IBD	Incomplete Block Designs
IBM	International Business Machines
ID	Incidence Density
I/O	Input Output
ISP	Internet Service Provider
IT	Information Technology
IUGR	IntraUterine Growth Retarded
JIPSD	Japanese International Protein Sequence Database
KB	Kilo Byte
LAN	Local Area Network
LB	Ljung-Box
LCD	Liquid Crystal Display
LD	Lethal Dose
L.H.S.	Left Hand Side
LISP	List Processing
LSD	Least Significant Difference
LS$_q$D	Latin Square Design
LSI	Large Scale Integrated
MA	Moving Average

MANOVA	Multivariate Analysis of Variance
MB	Mega Byte
MCSE	Microsoft Certified Systems Engineer
m.d.	Mean Deviation
m.e.	Mutually Exclusive
MGF	Moment Generating Function
MIPS	Munich Information Centre for Protein Sequences
mips	Mega Instructions Per Second
m.l.	Maximum Likelihood
m.l.e.	Maximum Likelihood Estimator
MNIC	Multipurpose National Identity Card
MRT	Multiple Range Test
MS	Mean Squares
MSA	Multiple Sequence Alignment
MSE	Mean Square Error
MSGr	Mean Sum of Square due to Groups
MSS	Mean Sum of Square
MSTr	Mean Sum of Square due to Treatments
MTBF	Mean Time Between Failure
MVN	Multivariate Normal
NBRF	National Biomedical Research Foundation
NCBI	National Centre for Biotechnology Information
NIH	National Institute of Health
NMR	Nuclear Magnetic Resonance
NTD	Non Tea Drinkers
NW	Needleman-Wunsch
OOPS	Object Oriented Programming Languages and Systems
OR	Odds Ratio
OR_{o}	Outcome Odds Ratio
OR_{E}	Exposure Odds Ratio
OS	Operating System
PACF	Partial Autocorrelation Function
PAM	Percent Accepted Mutation
PBIBD	Partially Balanced Incomplete Block Design
PCA	Principal Component Analysis
PDB	Protein Data Bank
pdf	Probability Density Function, $f(x)$ so that $f(x)dx = P[x \le X \le x + dx]$
PE	Prevalence Exposure
PGF	Probability Generating Function
PIAR	Percentage of Identically Aligned Residues
PIR	Protein Information Resource
pmf	Probability Mass Function, $p(x) = P[X = x]$
Q.d.	Quartile Deviations
QS	Quadrant Sum
RBD	Randomized Block Design

RC	Renal Cancer
RCT	Randomized Clinical Trials
RHS	Right Hand Side
RNA	Ribo Nucleic Acid
RR	Relative Risk
SAS	Statistical Analysis System
SBC	Schwartz Bayesian Criterion
SCOP	Structural Classification of Proteins
s.d.	Standard Deviation
SE	Standard Error
SI	Seasonal Index
SIB	Swiss Institute of Bioinformatics
SLP	Serum Lipid Peroxidase
SM	Scoring Matrix
SNOBOL	StriNg Oriented SymBOlic Language
SPSS	Statistical Package for the Social Sciences
SRS	Sequence Retrieval System
SRSWOR	Simple Random Sampling Without Replacement
SRSWR	Simple Random Sampling With Replacement
SS	Sum of Squares
SSB	Sum of Square due to Blocks
SSE	Sum of Square due to Error
SSGr	Sum of Square due to Groups
SSR	Sum of Square due to Regression
SST	Sum of square due to Treatments
SSTr	Sum of Square due to Treatments
SW	Smith-Waterman
t	Thymine
TD	Tea Drinkers
TMC	Tabulating Machine Company
TS	Trend Stationary
TSS	Total Sum of Squares
URL	Uniform Resource Locator
UT	Union Territory
VAR	Vector Auto Regression
VLSI	Very Large Scale Integrated
WAN	Wide Area Network
WBC	White Blood Cell
WS	Weighted Sum
www	World Wide Web