ers.com, for detailed information on medical statistics books, visit our website www.jaypeebrothers.com, for detailed information on medical statistics

Basics of Biostatistics

A Manual for Medical Practitioners



Jatinder Bali
Anil Kant

Foreword
Barun K Nayak

JAYPEE

Contents

Zero Point: Approach Road to Learning 1
Statistics in a Nutshell 8

Section 1: Bioethics, Evidence-based Medicine

1. **Bioethics in Clinical Research** 13
Major Divisions of Ethics 13; Bioethics 14; Main Branches of Bioethics 14
Etymology, History and the Famous Nonethical Experiments 14
Dachau and Auschwitz Concentration Camp Hypothermia Experiments 15
Animal Ethics 34; Environmental Ethics 35; Anthropocentrism 36
2. **Evidence-based Medicine and Levels of Evidence** 38
Evidence-based Medicine 38; Fundamental Principles 38
Components of Evidence-based Practices 39
3. **Introduction to Statistics** 49
Etymology 49; History 49; Notable Contributors 50
India in Modern Context 51; Applications 51; Summing Up 52; Biostatistics 52

Section 2: Research Methodology and 'Back to School'

4. **Research Methodology** 57
Objectives of Research 57; Types of Research 58
Approaches to Research 59; Research Methods 60; Research Methodology 60
5. **Inside Research Methodology: The Nuts and Bolts** 63
What Constitutes a Research? 63; Research Plan 72
6. **Back to School** 74
Greek Alphabets 74; Common Mathematical Operators/Symbols 75
Modulus 76; Basics of Set Theory and Venn Diagrams 76; Indices 79
Progressions: Arithmetic and Geometric Progressions 79
Logarithms and Antilogarithms 80; The Exponential Function 'e' 81
Ratio 81; Proportion 82; Factorial; 82; Permutations and Combinations 83

Section 3: Describing the Data

7. **Frequency Distribution** 87
Classification of Data 87; Variables: Discrete and Continuous 89
Frequency Distributions: Discrete and Continuous 89
Relative Frequency Distribution 93
Bivariate or Two Way Frequency Distribution 93; Marginal Distribution 94
8. **Measures of Central Tendency** 95
Central Value Concept 95; Measures of Central Tendency 95
Types of Measures of Central Tendency 96; Median (M_d) 100
Quartiles, Deciles and Percentiles 103; Mode (M_0) 105
Harmonic Mean 109; Which Measure to Use? 111
9. **Measures of Dispersion (or Variation)** 114
Understanding Dispersion 114; Types of Measures of Dispersion 115
Quartile Deviation 116; Percentile Deviation 116; Mean Deviation 116
Standard Deviation 117; Lorenz Curve 119
Coefficient of Dispersion and Its Measures (Also referred as a Relative Measure of Variation) 120; Which Measure of Dispersion to Use? 121

10. Skewness, Moments and Kurtosis 122
Skewness 122; Kurtosis 125; Moments 126; Significance 127

Section 4: Probability, Probability Distribution and Sampling

11. Theory of Probability 131
Terminology in Probability Theory 132; Approaches to Probability 133
Conditional Probability 137; Bayes Theorem 139

12. Theoretical Discrete Probability Distributions 142
Underlying Concepts and Terminology 142
Types of Theoretical Discrete Distributions 147

13. Theoretical Continuous Probability Distributions 163
Normal Distribution 163; Rectangular Distribution 174
Lognormal Distribution 176; Exponential Distribution 176

14. Sampling Theory 178
What is a Sample and why it is Required? 179; Types of Sampling Methods 180
Advantages and Disadvantages of Sampling 182
What should be the Sample Size? 183; Sampling Errors 185
Standard Error 186; Underlying Basis of Sampling Theory 186

Section 5: Correlation, Regression, Partial and Multiple Correlation and Regression, Theory of Attributes

15. Correlation 191
What is Correlation? 191; Correlation Measures 192

16. Regression Analysis 205
Components of Regression Analysis 205; Types of Regression 206
Linear Regression Analysis and Line of Best Fit 207; Logistic Regression 212

17. Partial and Multiple Correlation and Regression 216
Partial Correlation; 216; Multiple Correlation 217
Partial and Multiple Regression 218; Multicollinearity 221

18. Theory of Attributes 225
Terminology of Attribute Theory 225; Association of Attributes 228

Section 6: Statistical Testing

19. Hypothesis Testing 235
Basic Concepts 236;
Practical Considerations in Framing and Testing Hypothesis 246

20. Tests of Significance (Z and Student's 't'-Test) 248
Large Sample Tests 249; Small Sample Tests (Student's t-tests) 256
What is P-value? 267

21. F-Test and Analysis of Variance and Covariance 269
Constants and Properties of F-Distribution 269; Applications 270

22. Chi-square Test 286
Degrees of Freedom 286; Salient Points of Chi-square 287
Chi-square Distribution 288; Applications of χ^2 Distribution 289
Chi-square Test of Homogeneity of Correlation Co-efficient 298
Remarks on Chi-Square 300

23. Nonparametric Tests 301
Advantages 301; Disadvantages 302
Run Test: Test for Randomness in a Series of Observations 302
The Signed-rank Test for Testing Specified Mean or Median of a Population 303
Wilcoxon Signed Rank Test for Paired data 304

Kolmogorov-Smirnov Test for One Sample 307
Kolmogorov-Smirnov Test (For Comparing Two Populations) 308
Mann-Whitney 'U' Test for Equality of Two Means 310
Wilcoxon-Wilcox Test for Comparison of Multiple Treatment on a Series 311
Kruskal-Wallis Rank Sum Test (H-Test) 313
Friedman Test (Two-way ANOVA) 314
Test of Significance of Spearman's and Kendall Rank Correlation 317
McNemar's Test for Paired Samples 317; Cochran Q Test 318

24. Which Statistical Test to Use? 321
Summary of Statistical Tests 322

25. Special Considerations about Inferential Issues in Biostatistics 327
Cause-Effect Relationship: Is it Really True? 328; Bias 329
Confounding 329; Spurious Relationship 331; Effect Modification 331
Bradford Hill Causality Criteria 332; Exposure-Disease Association 333
Odds Ratio 334; Attributable Risk 335

26. Measurement and Error Analysis 337
Measurement 337; Levels of Measurement 338
Evaluating Measurement 340; Reliability v/s Validity 343
Assessment of Validity: Specificity and Sensitivity 344
Errors in Measurement 350

Section 7: Other Statistical Concepts Frequently Used in Research

27. Vital (Demographic) Statistics and Life Tables 357
Basics of Vital Statistics 357; Life Tables 367
Survival Analysis and Kaplan-Meier Curves 374; Log-Rank Test 378

28. Time Series Analysis 381
Time Series Components 382; Measuring the Components of Time Series 384

29. Interpolation, Extrapolation and Forecasting Methods 398
Terminology 399; Methods of Interpolation 400
Extrapolation and Forecasting 406

Section 8: Statistical Concepts for Medical Administration

30. Index Numbers 411
Key Decisions in Creating Index Number 412; Weighted Index Numbers 417
Quantity and Value Index Numbers 423; Chain Index Numbers 423
Base Shifting and Splicing 426; Index Number Tests 427

31. Statistical Quality Control 429
Control Charts 430; Types of Control Charts 431
Charts based on Actual Measurements 432
'C-Control Chart' for Number of Defects Per Unit 435
P-Control Chart (Fraction Defectives or Proportion of Defects) 436
Acceptance Sampling 437

32. Decision Theory 440
Notation/Terminology 441; Decisions Methods 443
Principles of Decision Making 444; Decision Tree Analysis or Tree Diagram 449

Section 9: Report Writing and Applying for Grants

33. Research Report Writing: How to do? 453
Organization of the Research Report 453
Manuscript Preparation for Journals 460

34. Applying for Grants **467**

Preparing for Application Process 468
Ground Reality: The Background Check 469
Main Application: Contents and Pitfalls 470
Background and Significance or Introduction 471
Preliminary Studies/Progress Report 471; Research Design and Methods 471
Results 472; Monitoring and Evaluation 472
Protection of Human Subjects and Institutional Review Boards 473
Bibliography 474; Contracts 474; Abstracts 474; Title 474
Biographical Sketches 474; Budget 475
Formatting and Instructions to Applicants 475
Cover Letter 476; Decision Process Explained 476; Pay Lines 477
Outcomes 477; Common Problems Encountered 478;
Revision of Application 479; Voila! Manna from Heaven: The Funding 480
Records and Accounting 480

Appendix **481**

Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence 482
Random Number Table 484; Binomial Coefficients 485
Binomial Probabilities 486
$e^{-\lambda}$ *Values for Calculating Poisson Probabilities for* $(0 < \lambda < 1)$ 490
Area Under the Standard Normal Curve 491
Area Under the Standard Normal Curve 493
Area Under the Standard Normal Curve 495
Critical Values of 't' 497; Critical Values of Chi Square (χ^2) *498*
Critical Values of F (0.05 or 5%) 500; Critical Values of F (0.01 or 1%) 503
Factors for Construction of Control Charts (Statistical Quality Control) 506
Studentized Range: Critical Values 508
Studentized Range: Critical Values 510
Pearson's Correlation Coefficient (r): Critical Values 512
Spearman Rank Correlation Coefficient (ρ)*: Critical Values 513*
Run Test: Acceptance Region for Values of 'R' 514
Sign Test: Critical Values of 'T' 515
Kolmogorov-Smirnov Test: Critical Values of 'D' 517
Wilcoxon-Wilcox Two-Sided Test: Critical Values 518
Mann-Whitney Test (U Test): Critical Values of 'U' 521
Critical Values of τ*: Kendall Rank Correlation Significance Test 522*
Wilcoxon Signed Rank Test 'W': Lower and Upper Critical Values 523
Common Logarithms ($\log_{10}x$) *524; Antilogarithms 530*

Index **537**

CHAPTER

3

Introduction to Statistics

Statistics is "the branch of science that deals with the collection, organization, analysis, and interpretation of numerical data." It is also defined as 'the study of the collection, analysis, interpretation, presentation, and organization of data.' It studies the methods of collection, summarizing and drawing of conclusion from the data. Statistics is widely applicable in different academic and applied functional fields such as physical and social sciences, humanities, business, government and industry. The sum total of attributes studied are of no use if some information regarding their characteristics cannot be derived or some actionable intelligence cannot be gained from them. The aim of collection of data is to convert it into information which may be used for action or decision. This justifies the resources expended on it.

ETYMOLOGY

The word statistics is derived from modern Latin phrase *statisticum collegium* meaning 'lecture about state affairs' or *status* meaning a "Political State" or a Government. Shakespeare used the word *Statist* in Hamlet in 1602. The Latin word gave rise to the Italian word 'statista' meaning statesman or politician and the German word 'Statistik' was used to describe the analysis of data about the state. In those days, statistics was used by rulers who needed information about land, agriculture, commerce, state population, etc. to assess their military potential, wealth and taxation. W Hooper used the word statistics in the translation of Baron BF Bieford's "Elements of Universal Erudition" in 1771. In this treatise, statistics refers to the science that teaches us the political arrangement of all the modern states of the known world.

HISTORY

In the early 19th century, the word statistics was formally used to mean collection and classification of data. In India, administrative data collection was seen around 300 BC, during the reign of Chandragupta Maurya. Abul Fazl recounts of surveys in *Ain-i-Akbari* in the late 16th Century.

The theoretical modern statistics came around the mid-17th century with the introduction of *Theory of Probability* and *Theory of Games and Chance*. These

JAYPEE BROTHERS

introductions were supposedly driven by the mathematicians and gamblers in France, Germany and England. Theory of Probability is one of the most critical concepts and can be considered as the backbone of modern statistics.

Before this, around early 17th century, 'Vital Statistics', dealing with the demographics, originated. Captain John Grant (1620–1674) in England was the first one to study the birth and death statistics and is called as the Father of Vital Statistics. His works and subsequent contributions from Casper Newman, Sir W Petty, Dr Price, etc. culminated in introduction of life insurance.

NOTABLE CONTRIBUTORS

Few main contributors in the development of statistics have been (In random order, does not signify importance of the contribution/ contributor and is not in any chronological order):

Contributor	*Contribution Field*
John Graunt	Vital statistics
B Pascal	'Problem of points' coefficients of binomial expansion
James Bernoulli	First treatise on 'Theory of Probability', Binomial distribution
Abraham De'Moivre	'The Doctrine of Chances', probability and annuities, pioneer work on normal distribution
Laplace	Probability (Laplace and Gauss rediscovered normal distribution while working on their respective theories on law of errors)
Gauss	Gave 'Principle of Least Squares' and 'Normal Law of Errors'
Thomas Bayes	Inverse probability
Lagrange	Interpolation
Sir Francis Galton	Regression, pioneered use of statistics in biometry
Karl Pearson	Standard Deviation, Skewness, Correlation, Introduced 1st test of significance- 'Chi- square Test'
WS Gosset (used pseudonym 'Student')	Student's t-test, introduced exact sample test for small samples
Sir Ronald Fisher	Known as 'Father of Statistics', applied statistics theory to diverse scientific fields such as agriculture, genetics, etc. Introduced concepts of point estimates, exact sampling distributions, variance, analysis of variance (ANOVA) and design of experiments, amongst others.
Jacques Quetelet	'Consistency of Great Numbers' which is basis of sampling theory
Laspeyres	Index numbers
Irving Fisher	Index numbers

JAYPEE BROTHERS

AN Kolmogorov	'Foundations of Probability'—Axiomatic approach to probability
Simeon Poisson	Poisson distribution
WA Shewhart	Statistical Quality Control
Wilcoxon	Non-parametric test

INDIA IN MODERN CONTEXT

In India, PC Mahalanobis, eminent scientist and applied statistician, is regarded as 'Father of Indian Statistics.' Apart from his pioneering studies in anthropometry, he founded the 'Indian Statistical Institute, Kolkata' in 1932. In 1933, the acclaimed journal "*Sankhya*" (the Hindi word for 'Number') was founded by him, after getting inspired by Karl Pearson's *'Biometrika.'* His other important contributions relate to his works on large scale sample surveys and introduction of concept of pilot surveys. These were acknowledged by international academia including Sir RA Fisher.

APPLICATIONS

Statistics has application in practically every field, and which is probably the reason why it does not have a clear-cut definition. Due to the vastness of its application, no definition is able to capture its true essence. Therefore, the simplest definition of statistics, "the science which deals with the collection, analysis and interpretation of numerical data" provided by 'Croxton and Cowden', appears to be the most suitable as it describes statistics in the simplistic form and does not limit its scope of application. Statistics does not have the true and exact characteristics like other pure sciences, as it has to deal with multiplicity of factors, cause and effect, etc. which cannot be captured in data always. It is more like a systematic technique and scientific method for obtaining meaningful information and knowledge, thereby providing assistance to all diverse fields or subjects, for their own technical advancements.

Statistics through its application, would possibly be classified as one of the most used subjects, irrespective of it being used- knowingly or unknowingly; by literates or illiterates, and in this lies its biggest disadvantage. Due to so much common use, statistics are quoted at every instance, without understanding the basis and its implication. It is generally assumed that quoting statistics tends to lend weight to the arguments. Possibly this is the reason of use and misuse of statistics.

Any data, by itself is neither correct nor incorrect. It is just that a datum or a 'data' set—an observation or set of observations 'present' or 'available' at a given point of time. How it is analyzed and inferences drawn from it, is not in the control of 'data' itself, but the entity who is handling this data. So much data is projected, used, available in print and non-print mediums, that it is difficult to trust all of it and more so, the validity of the inferences drawn, unless the source has high consistency and reliability. In general public life, percentage and inferences drawn from it are commonly misused. For example: a statement—highest number of travelers die, (say 61%) in road accidents, hence travelling by other modes should be preferred. Such a statement is incomplete and may have been wrongly inferred. It is true that 61% of deaths have

JAYPEE BROTHERS

been in road accidents, but considering roads are used by (say) 90% of the population, then as a ratio of users and deaths the inference may not hold true.

Misrepresentation and misinterpretation of data is seen and is not appropriate, but manipulation of data is another issue which leads to misuse of statistics. Misuse can happen through the use of inappropriate measures, altering definitions to suit outcomes, insufficient or inappropriate selection of samples, incomplete data or using specific period data and use of improper statistical significance tests, etc. Irrespective of whether these occur due to ignorance or deliberateness, these surely cast doubt.

In scientific community, such misquoting of statistics is not prevalent, but the challenges differ. The reason being that at research level, advanced statistics is used for analysis. Therefore, some of the key issues faced here include defining research problem, sampling design and methods utilized, establishing proper hypothesis, using appropriate test of significance, and reporting appropriately the results, including the limitations of the research study.

Types of Statistical Analysis

Statistical analysis is generally divided in two parts, namely:

1. Descriptive statistics, and
2. Inferential statistics

This demarcation is appropriate from research perspective, but cannot be termed as 'two broad divisions of statistics as a subject,' the reason being that such division ignores many aspects of statistics like vital statistics, index numbers, etc. Irrespective of suitability of such a division, these are commonly used in the terminology. In basic terms, these can be stated as:

- Descriptive statistics summarizes and describes data using indexes such as mean, standard deviation, skewness and kurtosis.
- Inferential statistics draws conclusion from data taking into account random variation and probability.

SUMMING UP

Statistics is a collection of methods and techniques to analyze numerical data or non-numerical data, which can be converted into numerical form without loss of interpretation. Therefore, it finds use in diverse fields like economics, biology, psychology, business management, etc. Most of these specialized fields do not require indepth knowledge of all aspects of statistics, but require thorough understanding of certain aspects, whereas for other aspects only basic familiarity and awareness would suffice. For example, for business management knowledge of concepts like Index number, modeling, simulation, regression, etc. is essential, but thorough knowledge of actuarial statistics would not be required, unless the person works in the insurance field.

BIOSTATISTICS

Biostatistics is the branch of statistics which deals with presentation and interpretation of scientific data in biology, public health and other biomedical sciences. This is a

JAYPEE BROTHERS

special branch because subjects (patients, mice, cells, etc.) exhibit variation in their response to various factors due to constitutional make-up/ genotype or the physical factors interacting with it/phenotype. Therefore, differences observed could be due to different treatments or they may be attributable to chance, measurement error, or other characteristics of the individual subjects. Biostatistics studies enunciate and elucidate different sources of variation. It is an interdisciplinary collaboration to advance knowledge by methodological use of statistical methods for biomedical applications. It applies statistical theory to real-world problems; helps design and conduct biomedical experiments and clinical trials; study computational algorithms; display clinical data. Biostatistics is an integral element in advancements of knowledge in disciplines of biology, public health policy, clinical medicine, proteomics, genomics, healthcare economics, etc.

Broadly stating, the concepts in statistics do not change, but only the terminology to which concepts are being applied, change. Method of testing the significance or calculating the Arithmetic Mean will remain same, irrespective of whether the data represents the monthly sales or weight of the person. Therefore, the critical aspect is—which is the best tool of statistics for the given field and the problem under the study, so that actionable information can be derived from it.

JAYPEE BROTHERS

for detailed information on medical statistics books, visit our website www.jaypeebrothers.com, for detailed information on medical statistics books, vis

Basics of Biostatistics

A Manual for Medical Practitioners

Salient Features

- A reference manual in a non-intimidating language in the field of basic statistical concepts related to medicine
- The prime purpose of the treatise is to remove the fear of statistics as an extra-terrestrial object and to foster the scientific way of thinking—accepting only the proof and not agreeing until the concept is clearly proved or disproved
- Pure mathematical approach has been limited to minimum possible with focus on the concepts
- Provides a complete overview of the statistical methods in a concise and precise manner to build a strong foundation for understanding the advanced tools and techniques
- Compares the concepts wherever possible so that the fine differences are captured and conveyed
- The chapters *Back to School*—to review the elementary concepts such as fractions, logs, set theory, permutation and combination, etc. and *Measurement and Error Analysis*—to focus on research papers and clinical trials, are added
- Includes the additional topics such as life table, Kaplan-Meier curves, interpolation, time series, etc. and subjects such as statistical quality control, decision theory, index numbers which find use in medical administration

Jatinder Bali is the Member-Secretary, Institutional Ethics Committee, Hindu Rao Hospital and North Delhi Municipal Corporation (NDMC) Medical College; and Scientific Research Committee, Hindu Rao Hospital, New Delhi, India. He is an ophthalmologist by training. He has been the Chief International Instructor, the World Ophthalmology Congress and an invited international Faculty for the Asia Pacific Academy of Ophthalmology and the American Academy of Ophthalmology. He has been a Faculty for the Delhi Ophthalmological Society and has chaired the sessions in seven All India Ophthalmological Society conferences on research, statistics and computer-related issues. He has been a peer-reviewer for British Journal of Ophthalmology, Journal of Venomous Animals and Toxins, Canadian Medical Journal, Singapore Medical Journal, Indian Journal of Ophthalmology, Annals of Health and Medical Research, Journal of Clinical Research and Ophthalmology and Saudi Journal of Ophthalmology.

As a member of the Scientific Research Committee and the Assistant DNB coordinator, he has worked closely with the students and so has a good understanding of the common problems faced by the students and practitioners in using statistics.

Anil Kant is currently the Head (Sales and Marketing), Max Speciality Films Ltd, New Delhi, India. He has a vast experience in various multinational companies such as the Director, Wifag-Polytype India, Noida; and the AVP, Sakata Inx India Ltd, Gurgaon, Haryana, India. A graduate in statistics from Kurukshetra University with an MBA in International Business from Faculty of Management Studies from Banaras Hindu University, he has widely relied on statistics in making business and the allied decisions. As a statistics enthusiast, he has always believed in demystifying the subject concepts for all and sundry. He understands the statistical need of the people with limited exposure to mathematics, especially, while conducting systematic research and in decision making.

Available at all medical bookstores
or buy online at www.jaypeebrothers.com

JAYPEE BROTHERS
Medical Publishers (P) Ltd.
www.jaypeebrothers.com

Join us on facebook.com/JaypeeMedicalPublishers

Shelving Recommendation
MEDICAL STATISTICS

ISBN 978-93-86150-71-4
9 789386 150714