Statistical Algorithms

Alpha Science

Rajan Chattamvelli

STA CAL ALGORITHMS

Rajan Chattamvelli

Alpha Science International Ltd.
Oxford, U.K.

Rajan Chattamvelli
Associate Professor of IT,
Periyar Maniammai University,
Thanjavur, Tamil Nadu 613 403
India

ALPHA SCIENCE INTERNATIONAL LTD.
7200 The Quorum, Oxford Business Park North
Garsington Road, Oxford OX4 2JZ, U.K.

www.alphasci.com

Printed from the camera-ready copy provided by the Author.

ISBN 978-1-84265-727-0

Printed in India

Preface

This is an intermediate level textbook on statistics with emphasis on theory and algorithms. It integrates the multitude of algorithms available for common data analysis tasks. Numerical examples illustrate most of the algorithms. Pseudo code for the algorithms is given to benefit the students and researchers.

This book can be used for senior undergraduate or graduate level courses in statistical computing, statistical algorithms, computational methods in statistics, computer intensive methods, and engineering statistics. Some of the chapters can also be used for courses on mathematical statistics (ch.1-4,6-10,12), advanced statistical methods (ch.1, 5-12), actuarial statistics (ch.2-4,6,7), biostatistics, economics (ch.2-4,8), information technology (ch.1-5,12), numerical methods (ch.1,5,12), industrial statistics (ch.2-5,8-10,12), econometrics (ch.1-7,9-11), management(ch.2-4, 8-10), medical statistics, commerce (ch.2-4,6) and computer science (ch.1-5,8,11,12). This is also an ideal book for researchers in various fields. The study of statistical algorithms is also important for scientists and professionals. A brief overview of the chapters is given below.

A general discussion of algorithms appears in Chapter 1. Complexity of algorithms is clearly discussed and types of complexities are described. Complexity notations are briefly reviewed in §1.2 (1-3). A recursive algorithm for sample mean appears in page 1-3. A summary table of complexity of algorithms can be found in page 1-5. Algorithm design techniques are discussed next. Several examples of recursive and iterative algorithms are given. Efficiency of algorithms is discussed in §1.4(p.1-14), followed by data dependency on algorithms. Algorithms for minimum and maximum (§1.5), powers, combinations, k^{th} smallest or largest element in an unsorted list (§1.7) are discussed.

Chapter 2 introduces popular measures of central tendency. Topics covered include arithmetic, geometric, harmonic and quadratic means, median and mode. Means for frequency data and weighted means are also discussed. Several novel algorithms to update these measures when data arrive afresh are given (p.2-5). Algorithms for trimmed-mean (page 2-7), sample median (page 2-12), mode (page 2-14) etc are discussed.

Chapter 3 introduces dispersion measures. Sample range (p.3-1), inter-quartile range, average absolute deviation (p.3-4), variance and standard deviation (p.3-5), etc are discussed. This is followed by a discussion of outliers and their detection (p.3-9). The chapter ends with a section on data transformations (p.3-11).

The chapter 4 discusses various measures of skewness and kurtosis. Absolute and relative measures are discussed in p.4-2. Popular skewness measures like Pearson's measure(p.4-3), Galton's measure (p.4-4), Kelly's measure and Bowley's measure (p.4-5) are introduced, and a new measure is proposed (p.4-5). This is followed by a discussion of kurtosis measures (p.4-8). The chapter ends with a review of L-kurtosis (p.4-10).

Chapter 5 is on matrix algorithms. Several statistical procedures like multiple regression, principal component analysis etc use matrix algorithms. This chapter starts with a general discussion of matrices. These are grouped under various categories for ease of comprehension. Quadratic forms are briefly introduced in section §5.5. Some of the popular matrix algorithms are introduced and illustrated with numerical examples. These include matrix inverse (p.5-10), matrix decompositions (p.5-15), eigen values (p.5-20), iterative matrix algorithms (p.5-28), etc. Linear equation algorithms are illustrated in §5.9.

Chapter 6 on Simple Linear Regression (SLR) gives algorithms for fitting linear regression model. Some variants of SLR are also discussed. Multiple Linear Regression (MLR) is described in §6.3. Multi-collinearity and its detection techniques are introduced in §6.4 and §6.5. This is followed by outlier detection (§6.6), variable selection methods (§6.7), and partial least squares (§6.8).

Chapter 7 gives a thorough discussion of measures of association. Sample covariance and correlation are introduced in §7.2 and §7.3. This is followed by other types of correlation like rank-correlation, Spearman's ρ, Kendall's τ, point biserial correlation, and ϕ-coefficient.

Chapter 8 is on distance-based methods. These include clustering, nearest neighbor algorithms, etc. Distance metrics that are extensively used in nearest neighbor algorithms, clustering algorithms, etc are discussed in the first few sections (p. 8-3 to 8-6). This is followed by similarity metrics. Popular clustering algorithms are discussed next. The chapter ends with a discussion of Nearest Neighbour Algorithms (NNA).

Chapter 9 discusses important discrete distributions. Some simple algorithms for the most popular distributions like Bernoulli, binomial, Poisson, geometric, hypergeometric, negative binomial and discrete uniform distributions are discussed in this chapter. Important properties are briefly summarised, and used in chapter 12 on random numbers.

The tenth chapter is on continuous univariate distributions. It discusses exponential, normal, gamma, beta, Cauchy, chi-square, t and F

distributions. A new family of distributions is introduced in §10.4 (p.10-4). Other distributions like skew-normal, log-normal, inverse Gaussian and arc-sin are briefly mentioned. Their properties and tail areas are summarised and used in chapter 12.

Chapter 11 is on EM Algorithm (EMA). This is a popular alternative to estimate parameters of a statistical population in the presence of missing data. A discussion on convexity of functions can be found in the beginning of this chapter. This is not needed from a practitioners view point, but is used in the derivation of the EM algorithm. The EMA is derived in page 11-8. Several examples of its use appear next. Variants of EM algorithms are also briefly reviewed.

The last chapter is on generating random numbers for simulation studies. Topics covered include random shuffling, congruential generators (linear and additive), functional composition method, etc. Generating random numbers for most of the important discrete and continuous distributions are discussed, and pseudocode for some of the algorithms are given. The last section discusses correlated random numbers. Chapters 9, 10 and 12 can be used for a course on simulation.

Chapter 2, 3, 7, 9 and 10 can be used for undergraduate courses in statistics, mathematics, econometrics, and business management. Other chapters may be used as 'supplementary reading' for various courses mentioned above. Professionals and practitioners in various fields can also use the book for self-study.

A large number of useful code-snippets and pseudo-code of algorithms appear throughout the book[1]. Educational use of this code is permitted, provided that it is duly acknowledged. Every software that uses the code in this book must document it as "Source: RAJAN CHATTAMVELLI (STATISTICAL ALGORITHMS)" without fail. Please contact the author at the following email for commercial use, or for porting to specific languages. Any misuse must be reported to the author. The author or publisher is not liable for any damage, loss, malfunction or hazard resulting from the use of the code.

A carefully selected set of exercises has been provided to benefit the students and self-study professionals. Answers and hints are provided for selected exercises. Any suggestions and comments for improvement are welcome. All suggestions should be sent to **dmmbook@gmail.com**. Up-to-date errata will be made available upon request.

Rajan Chattamvelli
Tanjore, Tamil Nadu.

[1]The semicolons that terminate each statement (in C/C++,Java) are omitted in all pseudocodes given in subsequent chapters to make it easy to port to VB, FORTRAN, and other programming languages without statement terminators.

Table of Contents

List of Figures

List of Tables

List of Algorithms