

STATISTICS: textbooks and monographs

volume 33

STATISTICAL COMPUTING

WILLIAM J. KENNEDY, Jr.
JAMES E. GENTLE

STATISTICAL COMPUTING

William J. Kennedy, Jr.

*Department of Statistics
Iowa State University
Ames, Iowa*

James E. Gentle

*ISML, Inc.
Houston, Texas*

MARCEL DEKKER, INC. New York and Basel

Library of Congress Cataloging in Publication Data

Kennedy, William J., Jr., [Date]
Statistical computing.

(Statistics, textbooks and monographs ; v. 33)
Includes index.

1. Mathematical statistics--Data processing.

I. Gentle, James E., [Date] joint author.

II. Title.

QA276.4.K46 519.5'028'54 80-10976

ISBN 0-8247-6898-1

COPYRIGHT © 1980 by MARCEL DEKKER, INC.

ALL RIGHTS RESERVED

Neither this book nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage and retrieval system, without permission in writing from the publisher.

MARCEL DEKKER, INC.

270 Madison Avenue, New York, New York 10016

Current printing (last digit):

10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

The purpose of this book is to present material that is generally considered to be in the area called *statistical computing*. Many numerical methods and algorithms are discussed from a computational viewpoint, and techniques for implementing given algorithms in a computer are often considered. The diversity of statistical computer applications is reflected in the book through the various chapters, each of which deals with a relatively general but unique application area. This book is designed to serve both as a textbook at the beginning graduate level, and as a reference text for people interested in computer applications of statistics. Exercises are provided at the end of most chapters for the student, and an extensive list of references is given on each subject to benefit all readers.

Much of the material in this text forms the basis for the Statistics 580 and 581 courses that have been presented at Iowa State University for the past several years. This two-quarter course sequence is taken by graduate students in statistics, computer science, mathematics, and related disciplines. Emphasis in these courses, and in this book, is on numerical methods in statistical computing. Semi-numerical and nonnumerical topics, such as computer graphics for example, are not discussed in detail.

The authors are indebted to many people for assistance which led to the completion of this text. All cannot be mentioned in this short space, but especially we wish to thank Professors H. O. Hartley, W. J. Hemmerle, and C. E. Gates for major contributions to our

education in statistical computing. We also thank Dr. T. A. Bancroft and Dr. H. A. David for their continuing energetic support of statistical computing which has provided the atmosphere and resources needed for growth in this subject area at ISU. Thanks also go to Professors V. A. Sposito, T. J. Boardman, and W. J. Hemmerle who read parts of the manuscript and made valuable suggestions. Finally, we thank Valerie Engeltjes for the excellent typing support that she provided through several revisions of the original manuscript; and Joyce Johnson, Marlene Sposito, and Darlene Wicks for preparation of the final camera-ready copy.

William J. Kennedy, Jr.

James E. Gentle

CONTENTS

Preface	v
1/ INTRODUCTION	1
1.1 Orientation	1
1.2 Purpose	2
1.3 Prerequisites	4
1.4 Presentation of Algorithms	5
2 COMPUTER ORGANIZATION	7
2.1 Introduction	7
2.2 Components of the Digital Computer System	7
2.3 Representation of Numeric Values	11
2.3.1 Integer Mode Representation	11
2.3.2 Representation in Floating-Point Mode	12
2.4 Floating- and Fixed-Point Arithmetic	14
2.4.1 Floating-Point Arithmetic Operations	14
2.4.2 Fixed-Point Arithmetic Operations	18
Exercises	19
References	21
3/ ERROR IN FLOATING-POINT COMPUTATION	23
3.1 Introduction	23
3.2 Types of Error	24
3.3 Error Due to Approximation Imposed by the Computer	25
3.4 Analyzing Error in a Finite Process	26
3.5 Rounding Error in Floating-Point Computations	27
3.6 Rounding Error in Two Common Floating-Point Calculations	31
3.7 Condition and Numerical Stability	33
3.8 Other Methods of Assessing Error in Computation	37
3.9 Summary	38
Exercises	39
References	41
4/ PROGRAMMING AND STATISTICAL SOFTWARE	45
4.1 Programming Languages: Introduction	46

4.2	Components of Programming Languages	48
4.2.1	Data Types	48
4.2.2	Data Structures	49
4.2.3	Syntax	50
4.2.4	Control Structures	51
4.3	Program Development	51
4.4	Statistical Software	56
	References and Further Readings	58
5/	APPROXIMATING PROBABILITIES AND PERCENTAGE POINTS IN SELECTED PROBABILITY DISTRIBUTIONS	61
5.1	Notation and General Considerations	63
5.1.1	Probability Distributions	63
5.1.2	Accuracy Considerations	64
5.2	General Methods in Approximation	66
5.2.1	Approximate Transformation of Random Variables	67
5.2.2	Closed Form Approximations	68
5.2.3	General Series Expansion	69
5.2.4	Exact Relationship Between Distributions	72
5.2.5	Numerical Root Finding	72
5.2.6	Continued Fractions	75
5.2.7	Gaussian Quadrature	80
5.2.8	Newton-Cotes Quadrature	86
5.3	The Normal Distribution	89
5.3.1	Normal Probabilities	90
5.3.2	Normal Percentage Points	93
5.4	Student's t Distribution	96
5.4.1	t Probabilities	96
5.4.2	t-Percentage Points	100
5.5	The Beta Distribution	103
5.5.1	Evaluating the Incomplete Beta Function	104
5.5.2	Inverting the Incomplete Beta Function	110
5.6	F Distribution	112
5.6.1	F Probabilities	112
5.6.2	F Percentage Points	116
5.7	Chi-Square Distribution	116
5.7.1	Chi-Square Probabilities	116
5.7.2	Chi-Square Percentage Points	118
	Exercises	119
	References and Further Readings	121
6/	RANDOM NUMBERS: GENERATION, TESTS AND APPLICATIONS	133
6.1	Introduction	133
6.2	Generation of Uniform Random Numbers	136
6.2.1	Congruential Methods	136
6.2.2	Feedback Shift Register Methods	150
6.2.3	Coupled Generators	162
6.2.4	Portable Generators	165
6.3	Tests of Random Number Generators	166
6.3.1	Theoretical Tests	166

6.3.2	Empirical Tests	169
6.3.3	Selecting a Random Number Generator	175
6.4	General Techniques for Generation of Nonuniform Random Deviates	176
6.4.1	Use of the Cumulative Distribution Function	176
6.4.2	Use of Mixtures of Distributions	179
6.4.3	Rejection Methods	184
6.4.4	Table Sampling Methods for Discrete Distributions	192
6.4.5	The Alias Method for Discrete Distributions	197
6.5	Generation of Variates from Specific Distributions	200
6.5.1	The Normal Distribution	201
6.5.2	The Gamma Distribution	209
6.5.3	The Beta Distribution	216
6.5.4	The F, t, and Chi-Square Distributions	219
6.5.5	The Binomial Distribution	221
6.5.6	The Poisson Distribution	223
6.5.7	Distribution of Order Statistics	225
6.5.8	Some Other Univariate Distributions	228
6.5.9	The Multivariate Normal Distribution	228
6.5.10	Some Other Multivariate Distributions	231
6.6	Applications	233
6.6.1	The Monte Carlo Method	233
6.6.2	Sampling and Randomization	236
	Exercises	241
	References and Further Readings	245
7/	SELECTED COMPUTATIONAL METHODS IN LINEAR ALGEBRA	265
7.1	Introduction	265
7.2	Methods Based on Orthogonal Transformations	266
7.2.1	Householder Transformations	266
7.2.2	Givens Transformations	273
7.2.3	The Modified Gram-Schmidt Method	276
7.2.4	Singular-value Decomposition	278
7.3	Gaussian Elimination and the Sweep Operator	286
7.4	Cholesky Decomposition and Rank-One Update	294
	Exercises	302
	References and Further Readings	308
8/	COMPUTATIONAL METHODS FOR MULTIPLE LINEAR REGRESSION ANALYSIS	313
8.1	Basic Computational Methods	313
8.1.1	Methods Using Orthogonal Triangularization of X	314
8.1.2	Sweep Operations and Normal Equations	321
8.1.3	Checking Programs, Computed Results and Improving Solutions Iteratively	326
8.2	Regression Model Building	331
8.2.1	All Possible Regressions	332
8.2.2	Stepwise Regression	335
8.2.3	Other Methods	340
8.2.4	A Special Case -- Polynomial Models	342

8.3	Multiple Regression Under Linear Restrictions	347
8.3.1	Linear Equality Restrictions	348
8.3.2	Linear Inequality Restrictions	352
	Exercises	361
	References and Further Readings	364
9/	COMPUTATIONAL METHODS FOR CLASSIFICATION MODELS	373
9.1	Introduction	373
9.1.1	Fixed-effects Models	374
9.1.2	Restrictions on Models and Constraints on Solutions	376
9.1.3	Reductions in Sums of Squares	378
9.1.4	An Example	379
9.2	The Special Case of Balance and Completeness for Fixed-Effects Models	382
9.2.1	Basic Definitions and Considerations	383
9.2.2	Computer-related Considerations in the Special Case	387
9.2.3	Analysis of Covariance	391
9.3	The General Problem for Fixed-Effects Models	395
9.3.1	Estimable Functions	396
9.3.2	Selection Criterion 1	400
9.3.3	Selection Criterion 2	404
9.3.4	Summary	406
9.4	Computing Expected Mean Squares and Estimates of Variance Components	406
9.4.1	Computing Expected Mean Squares	407
9.4.2	Variance Component Estimation	410
	Exercises	415
	References and Further Readings	416
10/	UNCONSTRAINED OPTIMIZATION AND NONLINEAR REGRESSION	425
10.1	Preliminaries	426
10.1.1	Iteration	427
10.1.2	Function Minima	427
10.1.3	Step Direction	429
10.1.4	Step Size	431
10.1.5	Convergence of the Iterative Methods	435
10.1.6	Termination of Iteration	436
10.2	Methods for Unconstrained Minimization	437
10.2.1	Method of Steepest Descent	438
10.2.2	Newton's Method and Some Modifications	442
10.2.3	Quasi-Newton Methods	451
10.2.4	Conjugate Gradient Method	460
10.2.5	Conjugate Direction Method	469
10.2.6	Other Derivative-Free Methods	473
10.3	Computational Methods in Nonlinear Regression	475
10.3.1	Newton's Method for the Nonlinear Regression Problem	478

10.3.2	The Modified Gauss-Newton Method	480
10.3.3	The Levenberg-Marquardt Modification of Gauss-Newton	483
10.3.4	Alternative Gradient Methods	485
10.3.5	Minimization Without Derivatives	487
10.3.6	Summary	491
10.4	Test Problems	492
	Exercises	498
	References and Further Readings	502
11/	MODEL FITTING BASED ON CRITERIA OTHER THAN LEAST SQUARES	513
11.1	Introduction	513
11.2	Minimum L_p Norm Estimators	514
11.2.1	L_1 Estimation	515
11.2.2	L_∞ Estimation	525
11.2.3	Other L_p Estimators	528
11.3	Other Robust Estimators	532
11.4	Biased Estimation	539
11.5	Robust Nonlinear Regression	544
	Exercises	547
	References and Further Readings	548
12/	SELECTED MULTIVARIATE METHODS	561
12.1	Introduction	561
12.2	Canonical Correlations	561
12.3	Principal Components	566
12.4	Factor Analysis	567
12.5	Multivariate Analysis of Variance	574
	Exercises	578
	References and Further Readings	579
	Index	583

1 / INTRODUCTION

1.1 ORIENTATION

During the past few years a tremendous amount of computer software has been developed to support statistical computing requirements. One significant feature of the software which has evolved over the years is its ever-improving quality in terms of capability, efficiency, and reliability. Much of this improvement can be traced to the development of better numerical methods, algorithms, and programming techniques along with an awareness of the characteristics of computer floating-point arithmetic.

Top-quality computer software for statistical applications is not easily produced. Development begins with a careful choice of the best among competing numerical methods. Selected methods are then specified in detail, with an eye toward computer implementation, in the form of algorithms. Each algorithm must subsequently be carefully coded in a computer language to form the nucleus of the finished software package. Additional considerations such as ease of use and flexibility are important to the overall makeup of the software. Decisions in this area will usually lead to the use of other methods and associated algorithms, many of which may be semi- or nonnumerical in nature. These latter algorithms must also be carefully coded to insure optimal performance of the final product. Thus both numerical and nonnumerical methods and associated algorithms form a basis for computer software, and careful implementation of selected algorithms is a vitally important aspect of the developmental process.

The need for new and improved methods and algorithms is never-ending. Research workers in statistics and related sciences are continually producing new results which require computational support. In many cases an algorithm to support a new computing requirement can be formed by modifying and/or specializing some existing algorithms. In other cases an extension of some numerical method may lead to a desirable algorithm. The most difficult situation is one in which a totally new numerical method must be derived to satisfy the computing requirement. Fortunately, this is not the usual case.

1.2 PURPOSE

The purpose of this book is to present selected computational methods, algorithms, and other subject matter which is important to the effective use of the given methods and algorithms in the computer. Most of the computational methods and algorithms contained in this book have been used in one or more software systems. There is usually more than one "good" method, and possibly several algorithms for each method, for handling a given problem, and the reader will find this fact reflected frequently in the chapters of this book.

Since the literature in statistical computing is extensive, and since it is not possible in the space of this text to treat every subject area and all important aspects within the area, the references listed at the end of each chapter are an important part of the chapter. The authors have attempted to include an extensive list of publications (including many that are not cited in the body of the chapter) which will allow the reader to move quickly to an in-depth study in any of the major areas defined in this text. The literature in statistical computing is to be found primarily in journals on statistics, computer science, and numerical analyses. A list of the most frequently referenced journals in the field and the abbreviations used for each is given in Table 1.1. Proceedings of conferences on statistical computing also contain many relevant articles. The annual Symposium on the Interface of Statistics and

Table 1.1 Journal Titles and Abbreviations

Abbreviation	Journal Title
JACM	<i>Journal of the Association for Computing Machinery (ACM)</i>
CACM	<i>Communications of the ACM</i>
TOMS	<i>ACM Transactions on Mathematical Software</i>
Comp. Rev.	<i>ACM Computing Reviews</i>
Comp. Sur.	<i>ACM Computing Surveys</i>
Numer. Math.	<i>Numerische Mathematik</i>
SIAM J. ^a	<i>Journal of the Society for Industrial and Applied Mathematics (SIAM)</i>
SIAM J. Appl. Math.	<i>SIAM Journal on Applied Mathematics</i>
SIAM Rev.	<i>SIAM Review</i>
SIAM J. Num. Anal.	<i>SIAM Journal on Numerical Analysis</i>
MTAC ^b	<i>Mathematical Tables and Other Aids to Computation</i>
Math. Comp.	<i>Mathematics of Computation</i>
JSCS	<i>Journal of Statistical Computation and Simulation</i>
Ann. Stat.	<i>Annals of Statistics</i>
AMS	<i>Annals of Mathematical Statistics</i>
JASA	<i>Journal of the American Statistical Association</i>
Amer. Stat.	<i>The American Statistician</i>
Techno.	<i>Technometrics</i>
Comm. Stat. (A)	<i>Communications in Statistics, Part A</i>
Comm. Stat. (B)	<i>Communications in Statistics, Part B</i>
JRSS (A)	<i>Journal of the Royal Statistical Society, Series A</i>
JRSS (B)	<i>Journal of the Royal Statistical Society, Series B</i>
Appl. Stat.	<i>Applied Statistics (Also called JRSS, Series C)</i>
Comp. J.	<i>Computer Journal</i>

Table 1.1 (Continued)

Abbreviation	Journal Title
<i>BIT</i>	<i>Nordisk Tidskrift for Information-behandling</i>
<i>Computing</i>	<i>Computing</i>
<i>J.O.T.A.</i>	<i>Journal of Optimization Theory and Applications</i>
<i>J. Inst. Maths. Applies.</i>	<i>Journal of the Institute of Mathematics and its Applications</i>

^aThe journal title *SIAM Journal* was used until 1966 when the name was changed to *SIAM Journal on Applied Mathematics*.

^bThis title was used from 1943 through 1959. In 1960 the title *Mathematics of Computation* was adopted.

Computer Science, the proceedings of which are published, has been a major factor in determining the growth of the field of statistical computing. The Statistical Computing Section of the American Statistical Association sponsors sessions at the annual meetings of the ASA, and the proceedings of these sessions are also published.

Exercises are given at the end of most chapters. The purpose of each exercise will be either to reinforce or to extend the material given in the chapter. None of the exercises is intended to be extremely difficult to solve.

1.3 PREREQUISITES

This book is designed for use by readers who have background in statistics equivalent to undergraduate training in statistical methods and theory; mathematical training in elementary calculus, matrix algebra, and numerical analysis; and a familiarity with the FORTRAN computer-programming language.

A complete discussion of an algorithm and its implementation on the computer often requires consideration of details which are dependent on computer hardware characteristics, the computer language used, and in some cases the language compiler. In situations in which computer hardware specifications are pertinent

to the discussion, this text assumes use of the IBM 370 series computer. Needed hardware specifications for this machine are given in Chapter 2. This reduces what would otherwise be a much larger prerequisite in applied computer science, and it gives the user of machines other than the IBM 370 (or equivalents) a basis for translating the text material to another machine-compatible form. There are only a few cases where the discussion is affected by the specific machine, the major ones being pseudorandom number generation and error analysis.

1.4 PRESENTATION OF ALGORITHMS

Many different computational methods will be described and discussed in succeeding chapters. In some instances an algorithm associated with some given computational method will be provided. When such is the case, and the algorithm is not stated in the form of a FORTRAN program, then the general form of the statement of the algorithm will be as follows. The first line will contain the word "ALGORITHM" followed by the name given to the algorithm and a FORTRAN-like argument list, enclosed in parentheses, containing input and output quantities. For example, a first line might be

```
ALGORITHM  RAND(I,J,X)
```

Following the first line, in some cases, will be a short set of comments designed to facilitate overall understanding of the algorithm. Then one or more numbered steps will be given along with a description of the operations to be performed in each step. These will take the form

Step	Description
1.	Set $I = I * J$
:	:
:	:

Comments will also appear between the steps in the algorithm as often as they are needed to explain the process.

When algorithms are stated in a programming language, such as FORTRAN, the meaning of the "=" symbol as an assignment of value is clear. However, there is some possibility of confusion when the "=" sign is used in the statement of algorithms using other than a programming language. The question may arise, in a specific situation, as to whether the "=" symbol is specifying an assignment of value or whether a statement of equality is intended. Sometimes the symbol " \leftarrow " or "!=" is used to denote assignment of value. This will not be done in this text. The context in which the symbol "=" is used will serve to define whether an assignment of value or a statement of equality is intended. If there is room for doubt, a comment will be made to clarify the situation.

2 / COMPUTER ORGANIZATION

2.1 INTRODUCTION

The purpose of this chapter is to describe the computer in general terms and consider some of the internal computer operations which are of special interest to the user in scientific applications. The computer is a tool to be used in performing calculations, and a basic understanding of how this tool operates is essential.

2.2 COMPONENTS OF THE DIGITAL COMPUTER SYSTEM

The computer system may be thought of as being made up of five components: (1) the main storage, (2) a control component, (3) an arithmetic and logic component, (4) an input component, and (5) an output component. Figure 2.1 depicts the way in which these components are interconnected.

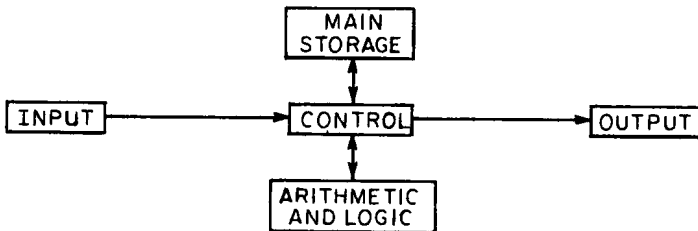


Figure 2.1 Components of the computer system.