Statistical and Computational Methods in Data Analysis

Second, revised edition

SIEGMUND BRANDT

Physics Department, Siegen University
Siegen, Germany

1976

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM · NEW YORK · OXFORD

Library of Congress Catalog Card Number: 77-113749

ISBN North-Holland 0 7204 0334 0

1st edition 1970
2nd printing 1973
2nd, revised edition 1976

Publishers:

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM · NEW YORK · OXFORD

Sole distributors for the U.S.A. and Canada:

ELSEVIER/NORTH-HOLLAND INC.
52 VANDERBILT AVENUE
NEW YORK, N.Y. 10017

Library of Congress Cataloging in Publication Data

Brandt, Siegmund.
Statistical and computational methods in data analysis.

Translation of Statistische Methoden der Datenanalyse.
Bibliography: p.
Includes index.
1. Probabilities. 2. Mathematical statistics.
I. Title.
[QA273.B86213 1976] 519.5 77-113749
ISBN 0-444-10893-9 (American Elsevier)

PRINTED IN THE NETHERLANDS

FROM THE PREFACE TO THE FIRST EDITION

The present book is based on lectures given in 1967/68 to physics students and particle physicists at Heidelberg University. It discusses those parts of mathematical statistics most relevant to data analysis. The book is intended for students and research workers in science, medicine, engineering and economics, who are faced with the problem of evaluating experimental data.

It is written from the standpoint of a user of mathematical statistics. Mathematical rigour is not overstressed. On the other hand, it does not merely list a number of recipes for different practical applications but attempts to explain the concepts and principles of the statistical methods discussed. Part of the presentation is influenced by lecture notes and review articles written by, and for, physicists [BÖCK, 1960; OREAR, 1958; SOLMITZ, 1964].

A good basic knowledge of calculus is assumed. Other necessary mathematical tools, especially probability theory, are briefly reviewed. An essential factor of the presentation is the use of matrix notation. It provides a very compact presentation of many problems such as the least squares method. An introduction to matrix calculus is given in the appendix.

Since most complex data analysis problems are now tackled with the help of computers, FORTRAN programs are presented for several such cases. The essential features of the FORTRAN language are discussed in the appendix. It also contains a short library of matrix handling subprograms, based on a similar set of programs originally written at CERN, Geneva, by R. Böck.

It is hoped that the book will not only serve as an introduction to statistical data analysis but will also be used in everyday work. It therefore contains a few statistical tables and a short collection of the more important formulae for quick reference.

I am indebted to several of my Heidelberg colleagues for discussions, in particular to Dr. T. P. Shah who read the manuscript and made many valuable suggestions for improvements. My thanks are also due to Dr. A. G. C. Tenner (Amsterdam) for a very fruitful discussion on the organization of

the book. Dr. H. Immich (Heidelberg) has kindly provided the examples 8-2, 11-1 and 11-2. Dr. H. Frenk (Wetzlar) made available a copy of the woodcut reproduced in front.

S. BRANDT

Heidelberg, January 1970

PREFACE TO THE SECOND EDITION

In the present edition the general concept of the book – short but sufficiently rigorous mathematical treatment, main emphasis on applications – has been left unchanged. However substantial additions have been made to the sector of statistical methods for direct application, in particular with computer programs. The main new sections are

- Elements of the Monte Carlo method (ch. 5, § 5.3)
- Rough numerical and graphical analysis of sampled data (ch. 6, § 8)
- FORTRAN program for linear regression (ch. 12, § 5)
- Time series analysis (ch. 13).

All of them contain FORTRAN programs. (The number of programs for direct application to statistical problems has been tripled.)

Furthermore the question of convolution of several distributions which can be rather cumbersome in practice, has been dealt with in greater detail and examples of convolution with the normal distribution are given. The chapter on sampling now contains a short section on very small samples.

Exercises are now given at the end of the chapters. Their solutions are outlined in a special section.

I should like to thank several of my collegues in Siegen for valuable discussions and suggestions, in particular Dr. W. Heinrich who read the manuscript of the new sections. On this occasion it is a pleasure for me also to acknowledge the excellent work of Dr. W. Wojcik and Prof. H. Yoshiki who made the translations for the Polish and Japanese editions of the book.

S. BRANDT

Siegen, June 1976

CONTENTS

Preface v
Contents vii
List of examples xi
List of FORTRAN programs xv
List of frequently used symbols xvii

1. Introduction 1

2. Probabilities 4

2-1. Experiments, events, sample space 4
2-2. The concept of probability 5
2-3. Rules of probability calculus; conditional probability . . . 7
Exercises 9

3. Random variables; distributions of a random variable 10

3-1. Random variables 10
3-2. Distributions of one random variable 10
3-3. Functions of one random variable, expectation value, variance, moments 12
3-4. Chebychev's inequality 18
Exercises 19

4. Distributions of several random variables 20

4-1. Distribution function and probability density of two variables; conditional probability 20
4-2. Expectation values, variances, covariances and correlation coefficient 22
4-3. More than two variables; vector and matrix notation 25
4-4. Transformation of variables 28
4-5. Linear and orthogonal transformations; propagation of errors 32
Exercises 37

5. Some important distributions and theorems 39

5-1. Binomial and multinomial distributions 39

5-2. Frequency; the law of large numbers 42
5-3. Hypergeometric distribution 43
5-4. Poisson distribution 47
5-5. Uniform distribution and an application: the Monte Carlo method 52
5-5.1. Probability density, expectation value, variance . . . 52
5-5.2. Generation of uniformly distributed random numbers by computers 53
5-5.3. Generation of any distribution by transformation of the uniform distribution 54
5-6. The characteristic function of a distribution 62
5-7. The Laplace model of errors 64
5-8. Normal distribution. 67
5-9. Quantitative properties of the normal distribution. 69
5-10. Multivariate normal distribution 72
5-11. The central limit theorem 78
5-12. Experimental errors and normal distribution; Herschel's model 79
5-13. Convolution of distributions 82
5-13.1. Folding integrals. 82
5-13.2. Convolution with the normal distribution 85
Exercises. 89

6. Sampling 92

6-1. Random sampling; distribution of a sample; estimates . . . 92
6-2. Sampling from continuous populations 95
6-3. Sampling from partitioned distributions. 97
6-4. Sampling without replacement from finite populations; mean square deviation; degrees of freedom 102
6-5. Sampling from normal distributions; χ^2-distribution 107
6-6. χ^2 and empirical variance 111
6-7. Sampling by counting. Small samples. 113
6-8. Numerical and graphical analysis of sampled data with computer programs 118
6-8.1. Scatter diagram and histogram of a one-dimensional sample 118
6-8.2. Scatter diagram of a two-dimensional sample 125
Exercises 131

7. The method of "maximum likelihood". 133

7-1. Likelihood quotient; likelihood function 133
7-2. The concept of maximum likelihood 135
7-3. Information inequality; minimum variance and sufficient estimates 138
7-4. Asymptotic properties of the likelihood function and of maximum likelihood estimators 145

7-5. Solution of the likelihood equation by iteration 147
7-6. Simultaneous estimation of several parameters. 148
7-7. Uniqueness of the method; confidence interval. 152
7-8. Bartlett's S-function 154
Exercises 157

8. Testing of statistical hypotheses 158

8-1. F-test on equality of variances 160
8-2. Student's test; comparison of means 164
8-3. Some aspects of a general theory of tests 169
8-4. Neyman–Pearson theorem and applications 174
8-5. The likelihood ratio method. 177
8-6. The χ^2-test on goodness of fit 182
Exercises 189

9. The method of least squares 191

9-1. Direct measurements with equal or different accuracy. . . . 191
9-2. Indirect measurements 196
9-2.1. The linear case. 196
9-2.2. The non-linear case. 204
9-2.3. Properties of the least squares solution; χ^2-test. . . . 210
9-3. Constrained measurements 213
9-3.1. The method of elements. 214
9-3.2. The method of Lagrangian multipliers 216
9-4. The general case of least squares fitting 221
9-5. A FORTRAN program for general least squares fitting; examples 224
Exercises 239

10. Some remarks on minimization 242

10-1. Parameter estimation and minimization 242
10-2. Different minimization procedures. 243

11. Analysis of variance 250

11-1. One-way classification 250
11-2. Some aspects of two-way classification. 255
11-3. A FORTRAN program for two-way classification. 264
Exercises. 269

12. Linear regression 270

12-1. Linear regression as a simple case of least squares. 270
12-2. Confidence intervals. 274
12-3. Testing of hypotheses 275
12-4. Linear regression and analysis of variance 276

12-5. A general FORTRAN program for linear regression . . . 277
12-6. Interpretation of results from linear regression 284
Exercises 289

13. Time series analysis 291

13-1. Time series. Trend 291
13-2. Moving averages 292
13-3. End effects 296
13-4. Confidence interval 296
13-5. A FORTRAN program for time series analysis 298
13-6. A word of caution 301
Exercise 305

Solution and discussion of the exercises 306

APPENDICES

A. Some elements of the FORTRAN programming language 329

B. Short review of matrix calculus 340

B-1. Definitions of matrices and vectors 340
B-2. Equality, addition, subtraction and multiplication of matrices 343
B-3. Determinant and inverse of a square matrix; solution of matrix equations 347
B-4. FORTRAN programs for matrix handling 354

C. Elements of combinatorial analysis 364

D. Euler's gamma-function 367

E. Collection of important formulae 369

F. Statistical tables 389

F-1. Poisson distribution 389
F-2. Normal distribution function 392
F-3. Fractiles of the normal distribution 395
F-4. χ^2-distribution function 398
F-5. Fractiles of χ^2-distribution 400
F-6. F-test 401
F-7. Fractiles for Student's test 406
F-8. Random numbers 407

References and bibliography 408
Subject index 411
Index to FORTRAN statements and FORTRAN programs used in this book 415

CHAPTER 1

INTRODUCTION

Every branch of experimental science, after passing through an early stage of qualitative description, concerns itself with quantitative studies of the phenomena of interest, i.e. measurements. Next to the design and the performance of the experiment, an important task is the accurate evaluation and the complete exploitation of the data obtained. Let us list a few typical problems.

1. The increase in the weight of test animals under the influence of various drugs is studied. After the application of drug A to 25 animals an average increase of 5% is observed. Drug B, used on 10 animals, yields 3%. Is drug A more effective? The averages 5% and 3% give practically no answer to this question, since the lower value may have been caused by a single animal that, for some reason, lost weight. One has therefore to study the *distribution* of individual weights and their dispersion around the average value. Moreover one has to decide whether the number of test animals used will enable one to differentiate between the effects of the two drugs with a certain accuracy.

2. In experiments on crystal growth the exact maintenance of the ratio of different components is essential. From a total of 500 crystals 20 are selected and analyzed. What conclusions can be drawn as to the composition of the remaining 480? This problem of *sampling* occurs for example in production control, reliability tests of automatic measuring devices and opinion polls.

3. A certain experimental result has been obtained. It has to be decided whether it contradicts some predicted theoretical value or previous experiments. The experiment is used for *hypothesis testing*.

4. A general law is known to describe the dependence of measured variables, but parameters of this law have to be obtained from experiment. In radioactive decay, for example, the number N of atoms that decay per second decreases exponentially with time: $N(t) = \text{const.} \times \exp(-\lambda t)$. The decay con-

stant λ and its measurement error are to be determined using a number of observations $N_1(t_1)$, $N_2(t_2)$ .... This problem of *parameter estimation* is perhaps the most interesting for many experimentalists.

From these examples some of the features of data analysis become apparent. We see in particular that the outcome of an experiment is not uniquely determined by the experimental procedure but is also subject to chance: it is a *random variable*. This stochastic tendency is either rooted in the nature of the experiment (test animals are necessarily different, radioactivity is a stochastic phenomenon), or it is a consequence of the inevitable uncertainties of the experimental equipment, i.e. the measurement errors. The next chapter is therefore devoted to reviewing the most important concepts of the theory of probability.

In chapters 3 and 4 random variables are introduced. The distribution of random variables is discussed and parameters, such as mean and variance are found to characterize these distributions. Special attention is given to the interdependence of several random variables. In chapter 5 a number of distributions is studied which are of special interest in applications, in particular the properties of the normal or Gaussian distribution are discussed in detail.

In practice a distribution has to be determined from a finite number of observations, i.e. a *sample*. Different cases of sampling are considered in chapter 6. FORTRAN programs are presented for a first rough numerical treatment and graphical display of empirical data. Functions of the sample, i.e. functions containing the individual observations, can be used to estimate the characteristic parameters of the distribution. The requirements that a good estimate should satisfy are derived. At this stage the quantity χ^2 is introduced. It is the sum of the squares of the deviation between observed and expected values and is therefore a suitable indicator of the quality of observation.

The *maximum likelihood method*, discussed in chapter 7, is the core of modern statistical analysis. It allows one to construct estimators with optimum properties. The method is discussed for the single- and multi-parameter cases and illustrated in a number of examples.

Chapter 8 is devoted to hypothesis testing. It contains the most commonly used *F*-, *t*- and χ^2-tests and outlines the general theory.

The *method of least squares*, which is perhaps the most widely used statistical procedure, is the subject of chapter 9. The special cases of direct, indirect and constrained measurements, often encountered in applications, are developed in detail before the general case is discussed. A FORTRAN

program for general least squares problems is presented and its use is demonstrated in different examples. Every least squares problem can be expressed as the task of determining the minimum of a function of several variables. This is true of all parameter estimation based on the idea of maximum likelihood. In chapter 10 several computational methods are sketched to obtain such minima.

The *analysis of variance* (chapter 11) can be considered as an extension of the *F*-test. It is widely used in biological and medical research to study the dependence, or rather to test the independence, of a measured variable of experimental conditions expressed by other variables. For several variables rather complex situations can arise. Some simple numerical examples are calculated using a FORTRAN program.

Linear *regression*, the subject of chapter 12, is a special case of the least squares method and therefore already dealt with in chapter 9. Before the advent of computers usually only linear least squares problems were tractable. A special terminology, still used, has developed for this case. It seems therefore justified to devote a special chapter to this subject. At the same time it extends the treatment of chapter 9. For example the determination of confidence intervals for a solution and the relation between regression and analysis of variance are studied. A general FORTRAN program for linear regression is given and its use is shown in examples.

In the last chapter the elements of *time series analysis* are introduced. This method is used if data are given as a function of a controlled variable (usually time) and no theoretical prediction for the behaviour of the data as a function of the controlled variable is known. It is used to try to reduce the statistical fluctuation of the data without destroying the genuine dependence on the controlled variable. Since the computational work in time series analysis is rather awkward a FORTRAN program is also given.