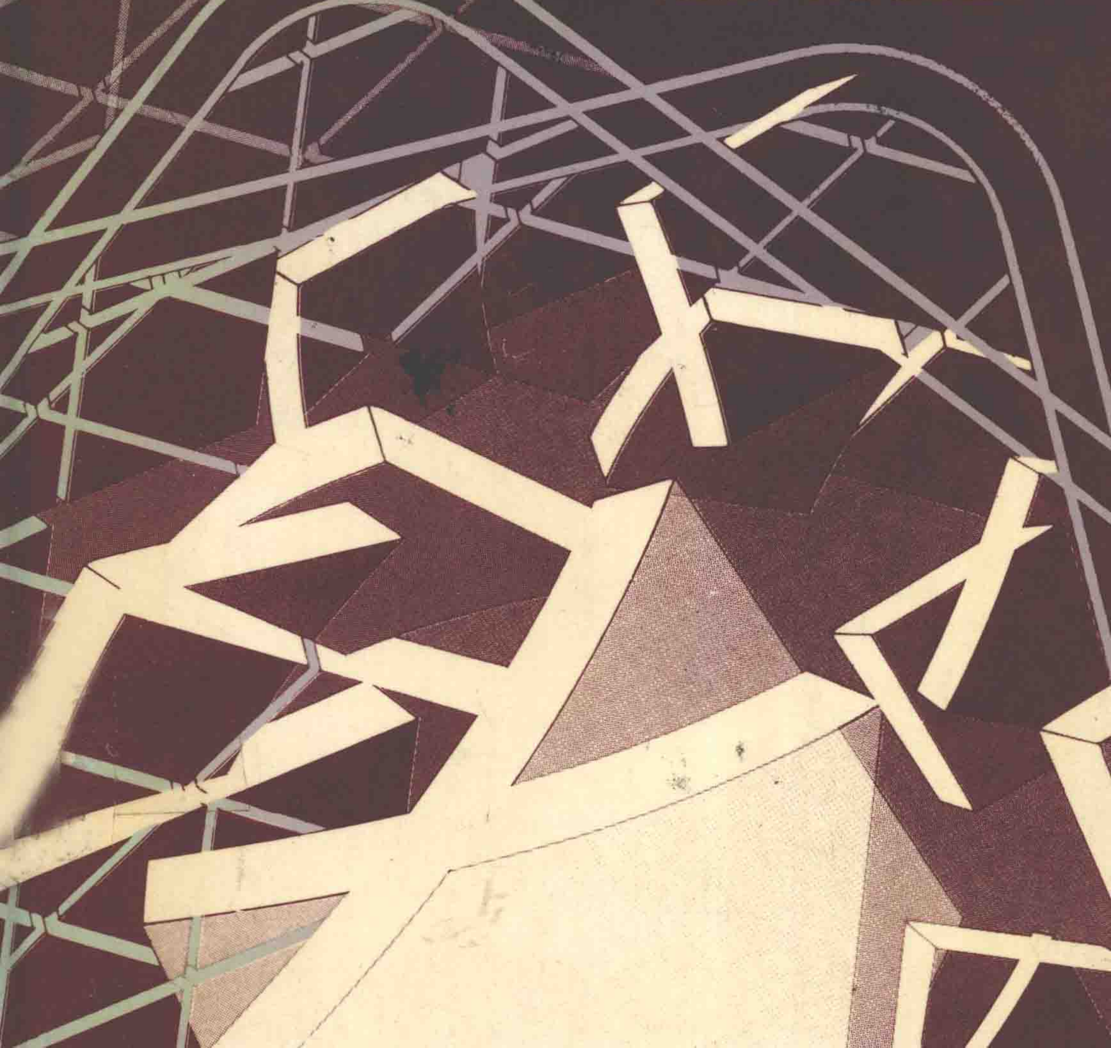
Ellis Horwood Series
MATHEMATICS AND ITS APPLICATIONS

APPLIED AND COMPUTATIONAL STATISTICS

a first course

KEITH STOODLEY



APPLIED AND COMPUTATIONAL STATISTICS
A First Course

ELLIS HORWOOD SERIES IN MATHEMATICS AND ITS APPLICATIONS

Series Editor: Professor G. M. BELL, Chelsea College, University of London

Statistics and Operational Research

Editor: B. W. CONOLLY, Chelsea College, University of London

Beaumont, G. P.	**Introductory Applied Probability**
Beaumont, G. P.	**Basic Probability and Random Variables***
Conolly, B. W.	**Techniques in Operational Research: Vol. 1, Queueing Systems**
Conolly, B. W.	**Techniques in Operational Research: Vol. 2, Models, Search, Randomization**
French, S.	**Sequencing and Scheduling: Mathematics of the Job Shop**
French, S.	**Decision Theory**
Hartley R.	**Linear Methods of Mathematical Programming***
Jones, A. J.	**Game Theory**
Kemp, K. W.	**Dice, Data and Decisions: Introductory Statistics**
Oliveira-Pinto, F.	**Simulation Concepts in Mathematical Modelling**
Oliveira-Pinto, F. & Conolly, B. W.	**Applicable Mathematics of Non-physical Phenomena**
Schendel, U.	**Introduction to Numerical Methods for Parallel Computers**
Stoodley, K. D. C.	**Applied and Computational Statistics: A First Course**
Stoodley, K. D. C., Lewis, T. & Stainton, C. L. S.	**Applied Statistical Techniques**
Thomas, L. C.	**Games, Theory and Applications**
Whitehead, J. R.	**The Design and Analysis of Sequential Clinical Trials**

**In preparation*

APPLIED AND COMPUTATIONAL STATISTICS

A First Course

DLEY, B.Sc., Ph.D.
Mathematical Sciences
University of Bradford

RWOOD LIMITED
Publishers · Chichester
Halsted Press: a division of
JOHN WILEY & SONS
New York · Chichester · Brisbane · Toronto

First published in 1984 by

ELLIS HORWOOD LIMITED
Market Cross House, Cooper Street, Chichester, West Sussex, PO19 1EB, England

The publisher's colophon is reproduced from James Gillison's drawing of the ancient Market Cross, Chichester.

Distributors:

Australia, New Zealand, South-east Asia:
Jacaranda-Wiley Ltd., Jacaranda Press,
JOHN WILEY & SONS INC.,
G.P.O. Box 859, Brisbane, Queensland 40001, Australia

Canada:
JOHN WILEY & SONS CANADA LIMITED
22 Worcester Road, Rexdale, Ontario, Canada.

Europe, Africa:
JOHN WILEY & SONS LIMITED
Baffins Lane, Chichester, West Sussex, England.

North and South America and the rest of the world:
Halsted Press: a division of
JOHN WILEY & SONS
605 Third Avenue, New York, N.Y. 10016, U.S.A.

British Library Cataloguing in Publication Data
Stoodley, Keith D.C.
Applied and computational statistics: a first course. –
(Ellis Horwood series in mathematics and its applications)
1. Mathematical statistics
I. Title
519.5 QA276

Library of Congress Card No. 84-12840

ISBN 0-85312-572-4 (Ellis Horwood Limited – Library Edn.)
ISBN 0-85312-805-7 (Ellis Horwood Limited – Student Edn.)
ISBN 0-470-20114-2 (Halsted Press)

Typeset by Ellis Horwood Limited.
Printed in Great Britain by R.J. Acford, Chichester.

Table of Contents

Chapter 3 Elementary Probability Theory

Chapter 4 Probability Distributions

Author's Preface

The material in this book constitutes a first course in computational and applied statistics equivalent to approximately twenty-five hours of lectures. Mathematical derivations have been omitted from the main body of the text, although the derivation of some of the more straightforward results is carried out in chapter appendices. Thus, the book is aimed principally at students who study statistics as an ancillary subject in order to apply statistical techniques in their own field. It should, nevertheless, be of use as background reading to those following more specialist mathematical or statistical courses.

The final section of each chapter contains programs developed on desk-top computers to illustrate the material introduced in that chapter. For these sections, it is assumed that the reader is able to program in BASIC. However, although the computing section refers back to the remainder of the chapter, the converse is not true. Thus, the reader who cannot program in BASIC can still make good use of the book. His loss will be his inability to appreciate the impact of the microcomputer in reducing the arithmetical rigours of statistical calculations, and the facilities provided by such machines for simulating random data and experimental situations.

Throughout the text, emphasis is placed on the concepts underlying the techniques discussed, the conditions under which the techniques are valid and the models assumed for the various experimental situations under investigation. The text is liberally illustrated with worked examples. Stress is placed on the use of calculators.

Each chapter is concluded with a set of exercises, to which answers are given. The exercises are divided into three sections; those in Section A are drawn from applications in the life sciences, while those in Section B relate to the physical sciences and engineering applications. The exercises in Section C are computer-orientated.

The book serves as an introductory text to *Applied Statistical Techniques* (Ellis Horwood, 1980) by Stoodley, Lewis and Stainton, and also to other more advanced statistical texts.

In conclusion, I should like to thank Mrs V. M. Hunter for her skilful typing of the manuscript and her equal skill in interpreting my handwriting; Mrs Jennifer Braithwaite for preparing the figures in the text; the University of Bradford for permission to use material from past examination papers; and the Biometrika Trustees for permission to make use of material from *Biometrika Tables for Statisticians,* Volume I, in preparing Tables A1, A2, A3 and A4 of the Appendix.

The computer programs listed in the text are available in substantially the same form on disk. Copies of the disk may be obtained from the publisher, Ellis Horwood Limited, Market Cross Hous, Cooper Street, Chichester, West Sussex, PO19 1EB.

Although the programs in this book have been carefully tested, it is the user's responsibility to ensure that normal checks are made in any application to which the programs are put.

Glossary of Greek characters used in the text

Greek character	English equivalent	Read as	Usage	Chapter
α	a	alpha	provisional mean	2.6
			Poisson distribution	4.4
			significance level	5
			line intercept	6
β	b	beta	line gradient	6
δ	d	delta	small increment	4.5.1
			term in model	5.8.2
ϵ	e	epsilon	experimental error	5.8.2, 6
λ	l	lambda	Poisson mean	4.4
μ	m	mu	population mean	all
			negative exponential distribution	4.5.2
ν	n	nu	degrees of freedom	5, 6
ρ	r	rho	population correlation coefficient	6.3
σ	s	sigma	population standard deviation	all
ξ	x	xi	coded variable	2.6
χ	ch	chi	chi-square distribution	5.5, 5.9
Σ	S	sigma	summation sign	all

To S.K.S.

1

Introduction

1.1 INFERENCE AND DECISION MAKING IN THE FACE OF UNCERTAINTY

Statistics is a highly logical and precise way of saying half-truths inaccurately.

Statistics is like dynamite – correctly used by experts it can be a useful and constructive tool; in careless or unscrupulous hands it can be a dangerous weapon. Used properly, statistics will allow an experimenter to quantify his concepts and conclusions, and help him to design his experiments so as to take into account sources of systematic variation and to minimise the effect of random error. It will draw his attention to the accuracy of his data and to the type and quality of the inferences which can be made from them. It will enable him to make rational decisions on the basis of a well established range of techniques, in spite of any random component of variation which may be present in his data. In this book the most important of the basic techniques used in modern statistics will be introduced; at the same time the limitations of the methods and the conditions under which they are valid will be stressed. It is hoped that in this way the reader will not be tempted to apply the methods in inappropriate circumstances.

The word 'statistics' is used in several distinct senses which may be summarised as follows:

(a) 'Statistics' may refer to a collection of data or observations, for example, accident statistics or vital statistics.

(b) 'Statistics' may refer to a body of techniques which have been developed in order to analyse data. The general object of such an analysis is to extract the maximum amount of relevant information from the data.

(c) A 'statistic' (where the word is used here in the singular) refers to a function of the observations (as, for example, the mean of the observations (Chapter 2)) which summarises some aspect of the information contained by the data.

Usually the particular interpretation to be placed on the word statistics will be obvious from the context in which it is used.

The first step in the analysis of a large body of data (for example the weights of 100 aspirin tablets illustrated in Table 2.1) is to summarise the information in the data by the calculation of appropriate statistics. In Chapter 2 it will be shown how to calculate statistics to illustrate the central tendency of the data and the dispersion of the data about this central value. Various methods of graphical presentation of the data, which help to elucidate their underlying structure, will also be described.

In addition to calculating statistics which will satisfactorily summarise the data we often require to make inferences, using the information in a sample, about the population from which the sample is drawn.

In quantitative experiments carried out in physics, engineering, chemistry, biology and many other disciplines, the responses obtained are often subject to random, as well as systematic, experimental errors. Such errors can arise in either, or both, of two different ways

(a) because of limitations in the accuracy of the apparatus being used and/or in the abilities of the experimenter,
(b) because of the nature of the law under investigation.

As an illustration of the situation (a) consider an experiment which involves the measurement of a time interval using a stop clock. Then, if the experimenter measured the interval several times as accurately as possible using the same stop clock, his results would vary over a range of values because of variation in his own response, limitations in the accuracy to which the clock can be read and possible sources of inaccuracy in the mechanism of the clock. Some of the errors may be systematic; for example a particular experimenter may consistently be a fraction late in starting the clock or the clock may be running fast; other errors will be random. The first step in the elimination of error in an experiment is to use the best methods and apparatus available and to perform the experiment carefully. Statistics will then allow us to cope in the most efficient manner with sources of error which are left.

As an example of situation (b) consider an experiment to investigate the relationship between the blood pressure and age of a person. If we take a random sample of adult males and plot their blood pressure against their ages it will be evident that there is an approximate linear relationship between these two variables. In this situation the approximate nature of the relationship does not arise to any appreciable extent from experimental errors in the measurement of blood pressure (although these may well exist) or age, but rather from the statistical nature of the underlying 'law' relating blood pressure to age.

The techniques of *statistical inference* enable deductions to be made concerning the parameters of the population from which a sample is drawn, when the observations are subject to random experimental error. Examples of situations in which such techniques may be applied are

(a) The weights of a random sample of 100 aspirin tablets from a large batch are known. We would like to estimate the mean weight of the tablets in the batch and to give some indication of the accuracy of our estimate.
(b) A large batch of bottles has been filled with fluid by an automatic process. On the basis of a random sample of the filled bottles we would like to test whether the mean volume of the fluid in the bottles from the batch differs significantly from the nominal content of the bottles.
(c) Patients suffering from a complaint are split into pairs matched as far as possible according to factors such as age, sex and medical history. One member of each pair is selected at random and is given a control drug, and the other member is given a new drug. We would like to know if the new drug is more effective than the control drug in the treatment of the complaint.
(d) A number of operators measure the percentage of a component in a mixed fluid, each operator using two methods. We would like to know whether there is a bias between the methods.
(e) The blood pressures of a random sample of adult males are measured. We would like to establish a relationship between blood pressure and age.

Various techniques for making statistical inferences are introduced in Chapter 5, where they are applied to situations such as (a) to (d) above. These techniques are based on the distributions associated with the random variables being measured, and the properties of several important distributions are described in Chapter 4. The consideration of these properties in turn requires the concept of probability and laws for the combination of probabilities of events. These topics are discussed in Chapter 3.

In situation (e) above we are faced with the problem of finding the 'best' straight line through a series of experimental points. This topic will be discussed in Chapter 6.

1.2 STATISTICAL COMPUTING

To err is human – it needs a computer to really foul things up.

The modern digital computer has become an invaluable aid to the statistician. Indeed many statistical techniques developed theoretically in the 1930s have only recently become of practical use with the advent of the computer. The desk-top computer, made possible by microprocessors which compress complex circuitry into an exceedingly small space, is more than adequate for carrying out the calculations associated with the techniques discussed in this book. Each chapter is illustrated by a final section which contains programs, developed on a desk-top computer, relevant to the material in the chapter. To make use of these programs, and the corresponding section of the exercises for the chapter (section C), the reader must have a knowledge of the programming language BASIC. Other readers may omit the computing sections, as the remainder of the book does not depend on them; however, they will miss the opportunity of seeing how the computer can ease the numerical work involved in statistical calculations, and also of generating their own data in order to illustrate and test the techniques developed in the book.

1.3 SCALES OF MEASUREMENT

We conclude this introductory chapter with a note on the types of data which may be subjected to statistical analysis. Data may be classified according to their scales of measurement as follows

(a) Nominal scale
On a *nominal scale* numbers are used as labels and no ranking is possible; for example groups of people could be labelled 1, 2 and 3 corresponding to the classification Liberal, Socialist and Conservative.

(b) Ordinal scale
On an *ordinal scale* scores can be assigned to the observations in such a way that the rankings obtained are meaningful, but the intervals between the rankings need not be defined; for example in a survey people's opinion of a product may be ranked according to the scale (1) poor, (2) fairly good, (3) good, (4) very good, (5) excellent.

(c) Interval scale
On an *interval scale* both the ranking and intervals are meaningful, but there is no meaningful zero; an example is the measurement of temperature in degrees Celsius.

(d) Ratio scale
On a *ratio scale* the measurements may be ranked, the intervals are meaningful and there is also a meaningful zero. Thus the ratios of measurements are also meaningful if they are made on a ratio scale. Measurements of quantities such as length, area and time are made on ratio scales, as are counts, such as the number of arrivals of cars at a service station in a given period of time.

In the following chapters we shall be concerned with the statistical analysis of data measured on interval and ratio scales. Statistical inferences may also be made on the basis of nominal or ordinal data, but the necessary techniques are not considered in this text.