V. BARNETT
T. LEWIS

OUTLIERS IN STATISTICAL DATA

3RD EDITION

WILEY SERIES IN PROBABILITY AND MATHEMATICAL STATISTICS

WILEY

Outliers in Statistical Data

Third Edition

VIC BARNETT
Rothamsted Experimental Station, UK

and

TOBY LEWIS
University of East Anglia, UK

JOHN WILEY & SONS
Chichester • New York • Brisbane • Toronto • Singapore

Other Wiley Editorial Offices

John Wiley & Sons, Inc., 605 Third Avenue,
New York, NY 10158-0012, USA

Jacaranda Wiley Ltd, 33 Park Road, Milton,
Queensland 4064, Australia

John Wiley & Sons (Canada) Ltd, 22 Worcester Road,
Rexdale, Ontario M9W 1L1, Canada

John Wiley & Sons (SEA) Pte Ltd, 37 Jalan Pemimpin #05-04,
Block B, Union Industrial Building, Singapore 2057

Library of Congress Cataloging-in-Publication Data

Barnett, Vic.
Outliers in statistical data / Vic Barnett and Toby Lewis. — 3rd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0 471 93094 6
1. Outliers (Statistics) I. Lewis, Toby. II. Title.
QA276.B2849 1994 93–29289
519.5′2—dc20 CIP

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 471 93094 6

Typeset in 10/12pt Times by Pure Tech Corporation, Pondicherry, India
Printed and bound in Great Britain by Biddles Ltd,
Guildford and King's Lynn

Outliers in Statistical Data

Third Edition

Preface to First Edition

The concept of an outlier has fascinated experimentalists since the earliest attempts to interpret data. Even before the formal development of statistical method, argument raged over whether, and on what basis, we should discard observations from a set of data on the grounds that they are 'unrepresentative', 'spurious', or 'mavericks' or 'rogues'. The early emphasis stressed the contamination of the data by unanticipated and unwelcome errors or mistakes affecting some of the observations. Attitudes varied from one extreme to another: from the view that we should never sully the sanctity of the data by daring to adjudge its propriety, to an ultimate pragmatism expressing 'if in doubt, throw it out'.

The present views are more sophisticated. A wider variety of aims are recognized in the handling of outliers, outlier- generating models have been proposed, and there is now available a vast array of specific statistical techniques for processing outliers. The work is scattered throughout the literature of the present century, shows no sign of any abatement, but has not previously been drawn together in a comprehensive review. Our purpose in writing this book is to attempt to provide such a review, at two levels. On the one hand we seek to survey the existing state of knowledge in the outlier field and to present the details of selected procedures for different situations. On the other hand we attempt to categorize differences in attitude, aim, and model in the study of outliers, and to follow the implications of such distinctions for the development of new research approaches. In offering such a comprehensive overview of the principles and methods associated with outliers we hope that we may help the practitioner in the analysis of data and the researcher in opening up possible new avenues of enquiry.

Early work on outliers was (inevitably) characterized by lack of attention to the modelling of the outlier-generating mechanism, by informality of technique with no backing in terms of a study of the statistical properties of proposed procedures, and by a leaning towards the hardline view that outliers should be either rejected or retained with full import. Even today sufficient attention is not always paid to the form of the outlier model, or to the practical purpose of investigating outliers, in the presentation of methods for processing outliers. Many procedures have an *ad hoc*,

intuitively justified, basis with little external reference in the sense of the relative statistical merits of different possibilities. In reviewing such techniques we will attempt to set them, as far as possible, within a wider framework of model, statistical principle, and practical aim, and we shall also consider the extent to which such basic considerations have begun to formally permeate outlier study over recent years.

Such an emphasis is reflected in the structure of the book.* The opening two chapters are designed respectively to motivate examination of outliers and to pose basic questions about the nature of an outlier. Chapter 1 gives a general survey of the field. In Chapter 2 we consider the various ways in which we can model the presence of outliers in a set of data. We examine the different interests (from *rejection* of unacceptable contamination, through the *accommodation* of outliers with reduced influence in robust procedures applied to the whole set of data, to specific *identification* of outliers as the facets of principal interest in the data). We discuss the statistical respectability of distinct methods of study, and the special problems that arise from the dimensionality of the data set or from the purpose of its analysis (single-sample estimation or testing, regression, analysis of data from designed experiments, examination of slippage in multisample data, and so on).

Chapter 3 examines at length the assessment of discordancy of outliers in single univariate samples. It discusses basic considerations and also presents a battery of techniques for practical use with comment on the circumstances supporting one method rather than another.

Chapter 4, on the accommodation of outliers in single univariate samples, deals with inference procedures which are robust in the sense of providing protection against the effect of outliers. Chapter 5 is concerned with processing several univariate samples both with regard to the relative slippage of the distributions from which they arise and (to a lesser extent) in relation to the accommodation of outliers in robust analysis of the whole set of data.

Chapters 6 and 7 extend the ideas and methods (in relation to the three interests: rejection, accommodation, identification) to single multivariate samples and to the analysis of data in regression, designed experiments, or time series situations. Chapter 8 gives fuller and more specific attention to the implications of adopting a Bayesian, or a non-parametric, approach to the study of outliers. The concluding Chapter 9 poses a few issues for further consideration or investigation.

The book aims to bring together in a logical framework the vast amount of work on outliers which has been scattered over the years in the various professional journals and texts, and which appears to have acquired a new

* Chapter numbers have been changed in the Second and Third Editions (see *Preface* and *Contents* to each Edition).

lease of life over the last decade or so. It is directed to more than one kind of reader: to the student (to inform him of the range of ideas and techniques), to the experimentalist (to assist him in the judicious choice of methods for handling outliers), and to the professional statistician (as a guide to the present state of knowledge and a springboard for further research).

The level of treatment assumes a knowledge of elementary probability theory and statistical method such as would be acquired in an introductory university-level course. The methodological exposition leans on an understanding of the principles and practical implications of testing and estimation. Where basic modelling and demonstration of statistical propriety are discussed, a more mathematical appreciation of basic principles is assumed, including some familiarity with optimality properties of methods of constructing tests and estimators and some knowledge of the properties of order statistics. Proofs of results are formally presented where appropriate, but at a heuristic rather than highly mathematical level.

Extensive tables of appropriate statistical functions are presented in an Appendix, to aid the practical worker in the use of the different procedures. Many of these tables are extracted from existing published tables; we are grateful to all the authors and publishers concerned, and have made individual acknowledgement at the appropriate places in our text. Other tables have been specially produced by us. The whole set of tables has been presented in as compact and consistent a style as possible. This has involved a good deal of selection and re-ordering of the previously published material; we have aimed as far as possible to standardize the ranges of tabulated values of sample size, percentage point, etc.

Copious references are given throughout the text to source material and to further work on the various topics. They are gathered together in the section entitled 'References and Bibliography' with appropriate page references to places of principal relevance in the text. Additional references augment those which have been discussed in the text. These will of course appear without any page reference, but will carry an indication of the main area to which they are relevant.

It is, of course, a privilege and pleasure to acknowledge the help of others. We thank Dave Collett, Nick Fieller, Agnes Herzberg, and David Kendall for helpful comments on early drafts of some of the material. We are particularly grateful to Kim Malafant who carried out the extensive calculations of the new statistical tables in Chapter 3. Our grateful thanks go also to Hazel Howard who coped nobly with the typing of a difficult manuscript.

We are solely responsible for any imperfections in the book and should be glad to be informed of them.

VIC BARNETT
TOBY LEWIS
July, 1977

Preface to Second Edition

The first edition of this book took the opportunity of drawing together for the first time the vast and assorted literature of over a century on the topic of outliers. In offering a combination of specific methods for the experimentalist and a 'state-of-the-art' review for the professional statistician it was designed to provide a logical framework for, and comprehensive coverage of, the variety of topics and emphases in outlier study. About four hundred references were used as basic source material, in a climate where the subject seemed clearly 'to have acquired a new lease of life over the last decade or so'. What could not be anticipated was the sheer force and momentum of this revitalization of interest. In the six years since the appearance of the first edition about three hundred more published articles have appeared: in crude numerical terms, a seventy-five per cent expansion over the whole previous history of the subject. It is inevitable therefore that we should seek to reassess the current situation in the light of so many new ideas and refinements. The second edition aims to do this.

The modifications in the revised edition are of three types. Firstly, there are areas of enquiry which represent new topics of outlier study or which now need to be described within the context of outlier methodology. These are reflected in the new chapters (7 and 11) on outliers in directional data, and in time series, respectively. Secondly, there are the many contributions which refine, reassess, or extend our knowledge on specific aspects of outlier investigation. Such developments are incorporated by means of substantial expansion, and some judicious pruning, of the discussion of almost all the topics of the earlier edition of the book. Particular attention has been given to discordancy tests for univariate and multivariate samples and for data from structured models (linear models generally, and specific aspects of regression and designed experiment situations). At a more general level, there has been a welcome growth of emphasis on methods of accommodation (robust inference in the face of outliers) and on informal (often graphical) descriptive procedures. These latter areas of development have contributed to the stimulus for a final type of modification—a reordering or reemphasis of some of the basic ideas and principles. In particular this has prompted a separation of general approach from specific results for the study of univariate samples with regard both to tests of

discordancy and methods of accommodation, and a reversal of order in the treatment of these two aspects. Thus we now have Chapter 3 and 4 on accommodation and Chapters 5 and 6 on discordancy testing, for univariate samples, distinguishing between general principle (Chapters 3 and 5) and specific method (Chapter 4 and 6) in each case. Accordingly the chapters dealing with slippage, multivariate outliers, outliers in linear models (regression, designed experiments), and Bayesian methods are now renumbered as Chapters 8, 9, 10, and 12, respectively.

There has been much activity in recent years on the theme of 'influential observations': as sample values which have disproportionate effect on estimates or tests of parameters in a model (particularly in the field of regression). Whilst not entirely coincident concepts, outliers and influential observations are clearly related in important aspects. Similarities and distinctions of aim and method are highlighted at appropriate stages in our revised study of outlier methods, especially in Chapter 10.

We are grateful to those who have pointed out misprints and ambiguities in the first edition; particularly to Dr N. A. Campbell. We have taken the opportunities of remedying such matters.

Whilst the new edition reflects the many developments and extensions mentioned above, it is important to stress that the basic emphasis and aim is unchanged. We continue to eschew the two extremes of pre-digested recipes for instant outlier management on the one hand and indulgence in mathematical formality or sophistication for its own sake on the other. Our aim remains one of explaining basic principle and developing associated method to a level where outlier techniques can be soundly and sanely applied (with relevant illustration and tabulation) and the researcher can be provided with a springboard for further exploration of a fascinating statistical topic.

VIC BARNETT
TOBY LEWIS
December, 1983

Preface to Third Edition

Since the publication of the second edition of this book, work in outlier methodology has continued unabated. Indeed, over 1000 new refereed publications have appeared in the literature. These not only extend previous knowledge in existing fields, but have opened up whole new areas of enquiry. Clearly, a new edition was needed to do justice to these many developments and to maintain our objective of providing a comprehensive and up to date coverage of the subject.

We have thoroughly revised and updated the material on the range of topics that were covered in the earlier editions. At the same time, extra material has been introduced to reflect new themes and changing emphases.

In contrast, a small amount of material has been omitted, in particular the chapter on the peripheral topic of *outlying subsamples*.

Topics on which the coverage *is new or has been substantially changed or extended* include:

- **basic principles**; distribution theory under contamination models, measures of efficiency and performance for multiple outliers, assessment of masking and swamping, allocation of outliers.
- **univariate data**; new tests (including extreme value and Weibull), wider study of robustness and accommodation (including logistic and double exponential distributions), additional tables.
- **multivariate and structured data**; estimation of individual components of vector parameters, use of correlation estimators, deletion methods, elliptically symmetric distributions, graphical methods, least median of squares and L_1-norm methods, multivariate linear model, multiple outliers, non-linear regression (including logistic and generalized linear models).
- **special topics**; Bayesian methods, time series (ARIMA model, distinction of AO and IO outliers, model specification, new accommodation methods, diagnostics, multiple time series), directional data (methods for axial and vectorial spherical data, accommodation).

Further special topics are dealt with in *new chapters*—these include new methods for outliers in contingency tables, problems of sample surveys, statistical software and international standards and regulations. Practical

illustrations remain important for reinforcement of ideas—new examples are included as also is discussion of data studies in the literature.

To cope with the vast expansion of material, it has been advantageous to restructure the book, which is now divided into the following four distinct parts: *Basic Principles, Univariate Data, Multivariate and Structured Data, Special Topics.* These reflect the headings used above to categorize the major areas of change. The *References and Bibliography* section has been substantially expanded to cover the new material.

We would like to thank Eileen Stoydin, Marian Joyce and Sharon Wilson for their most efficient assistance in the preparation of the typescript.

We hope that this new edition will prove to be valuable both to research statisticians and to practitioners in various fields.

VIC BARNETT
TOBY LEWIS
April, 1993

Contents

PART I

Basic Principles