



1011163528

VIC BARNETT  
TOBY LEWIS  
OUTLIERS IN  
STATISTICAL DATA  
Second Edition

WILEY SERIES IN PROBABILITY  
AND MATHEMATICAL STATISTICS



*Outliers in  
Statistical Data  
Second Edition*

VIC BARNETT

*University of Sheffield*

and

TOBY LEWIS

*The Open University*

*John Wiley & Sons*

*Chichester · New York · Brisbane · Toronto · Singapore*

Copyright © 1978, 1984, by John Wiley & Sons Ltd.

All rights reserved.

No part of this book may be reproduced by any means, nor transmitted, nor translated into a machine language without the written permission of the publisher.

***Library of Congress Cataloging in Publication Data:***

Barnett, Vic.

Outliers in statistical data.

(Wiley series in probability and mathematical statistics. Applied probability and statistics)

Includes bibliographical references and index.

1. Outliers (Statistics) I. Lewis, Toby. II. Title.

III. Series.

QA276.B2849 1984 519.5 84-7537

ISBN 0 471 90507 0

***British Library Cataloguing in Publication Data:***

Barnett, Vic

Outliers in statistical data.—2nd ed.—

(Wiley series in probability and mathematical statistics)

1. Outliers (Statistics)

I. Title II. Lewis, Tobias III Series

519.5 QA276

ISBN 0 471 90507 0

Photoset at Thomson Press (India) Limited, New Delhi  
Printed in Great Britain by Page Bros., Norwich

QA  
276  
.B2849  
1984

## *Preface to First*

The concept of an outlier has attempts to interpret data. Even method, argument raged over observations from a set of data of 'spurious', or 'mavericks' or 'rotation of the data by unan affecting some of the observati another: from the view that we s daring to adjudge its propriety, doubt, throw it out'.

The present views are more recognized in the handling of o proposed, and there is now availa for processing outliers. The work present century, shows no sign of drawn together in a comprehensive attempt to provide such a view, a the existing state of knowledge in selected procedures for different categorize differences in attitude, follow the implications of such dis approaches. In offering such a o methods associated with outliers w analysis of data and the research enquiry.

Early work on outliers was (inc the modelling of the outlier-gener with no backing in terms of a st procedures, and by a leaning tow either rejected or retained with full

9  
UNIV  
NEW MEXIC

QA  
276  
.B2849  
1984

## *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

923727  
UNIVERSITY LIBRARY  
NEW MEXICO STATE UNIVERSITY

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 lease

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

of life over the last decade or so. It is for the student (to inform him of the work of the experimentalist (to assist him in the study of outliers), and to the professional statistician (to provide a knowledge and a springboard for further work).

The level of treatment assumes a level of knowledge of statistics and statistical method such as would be expected of a first level course. The methodological principles and practical implications are given, with modelling and demonstration of mathematical appreciation of basic principles, familiarity with optimality properties of estimators and some knowledge of the way in which results are formally presented when they are at a highly mathematical level.

Extensive tables of appropriate statistical distributions, Appendix, to aid the practical worker. Many of these tables are extracted from the literature to all the authors and publishers. Acknowledgement at the appropriate places. The book is specially produced by us. The style is compact and consistent as far as possible to standardize the presentation of the percentage point, etc.

Copious references are given throughout the book to further work on the various topics. A list of references entitled 'References and Bibliography' is given at the end of the book. Places of principal relevance in the literature are indicated by page reference, but will carry an index of relevance.

It is, of course, a privilege and pleasure to thank Dave Collett, Nick Fieller, and others for their helpful comments on early drafts of the book. Grateful to Kim Malafant who carried out the statistical tables in Chapter 3. Our grateful thanks to the typists who coped nobly with the typing of a book of this size.

We are solely responsible for any errors and omissions to be informed of them.

July, 1977

model, or to the practical purpose of of methods for processing outliers. ely justified, basis with little external ical merits of different possibilities. In to set them, as far as possible, within a ciple, and practical aim, and we shall basic considerations have begun to nt years.

icture of the book.\* The opening two ivate examination of outliers and to an outlier. Chapter 1 gives a general der the various ways in which we can ata. We examine the different interests ation, through the accommodation of rocedures applied to the whole set of s the facets of principal interest in the ility of distinct methods of study, and ensionality of the data set or from the mation or testing, regression, analysis ation of slippage in multisample data,

essment of discordancy of outliers in sic considerations and also presents a with comment on the circumstances ber.

outliers in single univariate samples, are robust in the sense of providing Chapter 5 is concerned with processing egard to the relative slippage of the (to a lesser extent) in relation to the ysis of the whole set of data.

ethods (in relation to the three interests: to single multivariate samples and to le experiments, or time series situations. tion to the implications of adopting roach to the study of outliers. The issues for further consideration or

logical framework the vast amount of tered over the years in the various h appears to have acquired a new lease

and Edition (see *Preface to Second Edition* and

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.

July, 1977

VIC BARNETT  
TOBY LEWIS

## Contents

|                        |     |
|------------------------|-----|
| ds                     | 292 |
| ds                     | 292 |
| ds                     | 301 |
| s for the linear model | 306 |
|                        | 309 |
| del outliers           | 318 |
|                        | 3   |
|                        | 324 |
| ntial observations     | 325 |

## E SERIES: A LITTLE-EXPLORED

|                |     |
|----------------|-----|
| in time series | 330 |
| ne series      | 333 |
| ics            | 334 |
| teristics      | 337 |

## ACHES TO OUTLIERS 343

|        |     |
|--------|-----|
| iers   | 346 |
| nation | 353 |

S 364

424

## CHAPTER 1

### *Introduction*

From the earliest attempts of man to harness and employ the information implicit in collected data as an aid to understanding the world he lives in, there has been a concern for 'unrepresentative', 'rogue', or 'outlying' observations in sets of data. These are often seen as contaminating the data: reducing and distorting the information it provides about its source or generating mechanism. It is natural to seek means of interpreting or categorizing *outliers*—of sometimes rejecting them to restore the propriety of the data, or at least of taking their presence properly into account in any statistical analysis.

What are outliers and what is the outlier problem? To quote from Ferguson (1961a),

the general problem... is a very old and common one. In its simplest form it may be stated as follows. In a sample of moderate size taken from a certain population it appears that one or two values are surprisingly far away from the main group. The experimenter is tempted to throw away the apparently erroneous values, and not because he is certain that the values are spurious. On the contrary, he will undoubtedly admit that even if the population has a normal distribution there is a positive although extremely small probability that such values will occur in an experiment. It is rather because he feels that other explanations are more plausible, and that the loss in the accuracy of the experiment caused by throwing away a couple of good values is small compared to the loss caused by keeping even one bad value. The problem, then, is to introduce some degree of objectivity into the rejection of the outlying observations.\*

In the light of developments in outlier methodology over the last 20 years or so, Ferguson's formulation is unduly restrictive in various ways, as we shall see; for example, outlying observations are not necessarily 'bad' or 'erroneous', and the experimenter may be tempted in some situations not to 'throw away' the 'spurious' observation but to welcome it as an indication of, say, some unexpectedly useful industrial treatment or surprisingly successful agricultural

\* Copyright © 1961 by Regents of the University of California; reprinted by permission of the University of California Press.

variety. Then again, it may not be necessary to take either of the risks: of the loss arising from 'throwing away a couple of good values', or from 'keeping even one bad one'. In appropriate circumstances we can use 'robust' methods of inference which employ all the data but minimize the influence of any outliers. Such methods will be said to *accommodate* the outliers. However, the passage brings out many of the essentials of the outlier situation, and provides a useful basis for discussion.

The first point to make is that the problem, as well as being a 'very old and common one', is an *unavoidable* one. It is all very well to say, as some statisticians do, that one should not consider dealing with outlying observations unless furnished with information on their prior probabilities. Some will not even admit the concept of an outlier unless there is some obvious physical explanation of its presence! But the fact is that experimental scientists and other people who have to deal with data and take decisions are forced to make judgements about outliers—whether or not to include them, whether to make allowances for them on some compromise basis, and so on. Sometimes this is done in what appears, by modern standards, to be an unnecessarily naive or inefficient way. For example, a chemistry text book in current use (Calvin *et al.*, 1949, reprinted in 1960) advises its readers to use Chauvenet's method: 'Any result of a series containing  $n$ ... observations shall be rejected when the magnitude of its deviation from the mean of all measurements is such that the probability of occurrence of all deviations as large or larger is less than  $1/2n$ '. This rather strange proposal was one of the earliest for dealing with outliers, dating from the middle of the nineteenth century (Chauvenet, 1863). We shall say more about this, and about other early ideas for rejecting outliers, in Section 2.1.

Some of the basic characteristics of outliers can be illustrated by examples in English marital law: a context in which it is (unfortunately) rather unusual to encounter the application of statistical method.

In 1949, in the case of *Hadlum v. Hadlum*, Mr Hadlum appealed against the failure of an earlier petition for divorce. His claim was based on alleged adultery by Mrs Hadlum, the evidence for which consisted of the fact that Mrs Hadlum gave birth to a child on 12 August 1945 which was 349 days later than when Mr Hadlum had left the country to serve King and nation! The average gestation period for the human female is 280 days: 349 days seemed to Mr Hadlum to be surprisingly long. He adjudged it an *outlier*. He did not want his outlying observation to be *rejected* (discarded, ignored). On the contrary he wanted it to be identified (acknowledged) with appropriate consequences. (Such distinctions of aim in the handling of outliers will become a major factor in our later study). Mr Hadlum was unsuccessful. The appeal judges agreed that some limit of credibility must be entertained but ruled that 349 days, whilst improbable, was not beyond the range of scientific possibility. In other similar cases conflicting views had prevailed. In *M.-T. v. M.-T.* also in 1949 the court had ruled that 340 days was *impossible* in the light of modern gynaecological experience, whilst for *Preston-Jones v. Preston-Jones* in 1951 the House of Lords had ruled that the 'limit' should be set at 360 days. It is interesting that in none of these cases did the

## Introduction

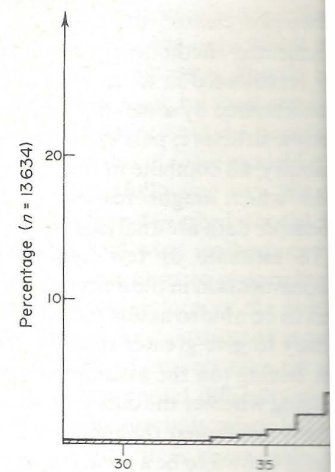


Figure 1.1 Histogram of human gestation periods, 1978, by permission of the Royal Society

distribution of human gestation periods (see Barnett, 1978a, based on results estimated gestation periods for a series of outliers superimposed. The vertical axis represents the percentage of observations born within a one-week period.

A much earlier case resurfaced in 1978 when a man failed in a petition for divorce on the grounds of separation from the marital home of 331 days. The court ruled that the divorce on grounds of separation was not another type!

Let us note the features of the distribution of gestation periods. Surprisingly so, and we designate it as a 'contaminant' of the status of the observation, 349 days is an outlier in the distribution of gestation periods. This is essentially what the court ruled in *Hadlum v. Hadlum*: an observation from a different origin). This is precisely what Mr Hadlum wanted: a prospect which would not be of interest to the court. A relatively unsurprising appeal still is a contaminant. The actual gestation date 20 days later than the expected date is a contaminant will not show up as an outlier. This is reiterated in a more formal manner.

With the development during the

to take either of the risks: of the loss of values', or from 'keeping even one can use 'robust' methods of inference the influence of any outliers. Such outliers. However, the passage brings attention, and provides a useful basis for

them, as well as being a 'very old and very well to say, as some statisticians with outlying observations unless probabilities. Some will not even admit the obvious physical explanation of its scientists and other people who have forced to make judgements about whether to make allowances for them sometimes this is done in what appears, by a naive or inefficient way. For example, a *et al.*, 1949, reprinted in 1960) advises any result of a series containing  $n...$  magnitude of its deviation from the mean ability of occurrence of all deviations as her strange proposal was one of the in the middle of the nineteenth century at this, and about other early ideas for

ers can be illustrated by examples in is (unfortunately) rather unusual to method.

Mr Hadlum appealed against the claim was based on alleged adultery consisted of the fact that Mrs Hadlum which was 349 days later than when Mr g and nation! The average gestation 49 days seemed to Mr Hadlum to be outlier. He did not want his outlying ed). On the contrary he wanted it to be e consequences. (Such distinctions of a major factor in our later study). Mr es agreed that some limit of credibility s, whilst improbable, was not beyond r similar cases conflicting views had he court had ruled that 340 days was logical experience, whilst for *Preston-* e of Lords had ruled that the 'limit' g that in none of these cases did the

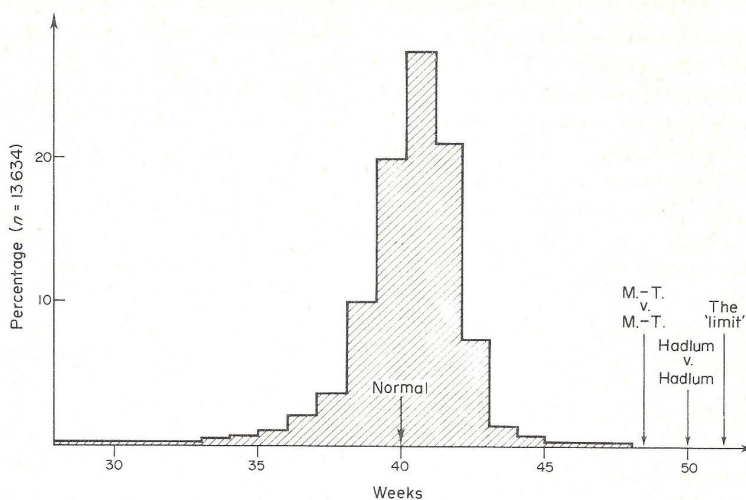


Figure 1.1 Histogram of human gestation periods (reproduced from Barnett, 1978, by permission of the Royal Statistical Society)

distribution of human gestation periods seem to figure in the evidence. Figure 1.1 (see Barnett, 1978a, based on results in Chamberlain, 1975) shows a histogram of estimated gestation periods for a sample of 13 634 British births, with the above outliers superimposed. The vertical scale represents percentage of the sample born within a one-week period.

A much earlier case resurfaced at about the same time. In 1921, Mr Gaskill failed in a petition for divorce on the grounds of adultery based on an absence from the marital home of 331 days. In 1960, Mrs Gaskill succeeded in obtaining a divorce on grounds of separation for a period of 39 years: an outlier of quite another type!

Let us note the features of the Hadlum case. The period of 349 days is *extreme*: surprisingly so, and we designate it an *outlier*. But this does not resolve the matter of the status of the observation, 349. It could be a genuine observation from the distribution of gestation periods which just happens to be surprisingly extreme. This is essentially what the court ruled! On the other hand it might be a *contaminant*: an observation from some other distribution (in this case with later origin). This is precisely what Mr Hadlum was claiming! Note yet a further prospect which would not be of interest to Mr Hadlum, but might be to someone else. A relatively unsurprising *apparent* gestation period of, say, 280 days could still be a contaminant. The *actual* gestation period could be, say, 260 days with a conception date 20 days later than had been assumed. But such a non-extreme contaminant will not show up as an outlier. These crucial distinctions will be reiterated in a more formal manner later in this section.

With the development during this century of more formal approaches to the

statistical analysis of data, objectives have become clearer, principles more rigorously defined, and a vast array of sophisticated methodology has been constructed. Practical situations are commonly represented in terms of different possible families of probability models often characterized by some small number of parameters. General considerations of situation structure, past experience of similar circumstances, and mathematical tractability, all combine to suggest one particular family of probability distributions which might reasonably be expected to represent the prevailing situation. Sample data are analysed to assess the validity of the prescribed model, and to estimate or test hypotheses concerning relevant parameters. This greater sophistication in the design and use of statistical methods makes it no less important to be able to assess the integrity of a set of data. However, there is some tendency to give greater regard to the processing of data for parameter estimation or testing (on the assumption that such and such a model applies) than to investigating whether the data give added support to the general considerations which have promoted the model.

This is a somewhat dangerous principle. What is known to be a good statistical procedure for estimating the mean of a normal distribution may be most inefficient if the distribution is not normal. The actual data being analysed can sound a warning for us! Perhaps one or more observations look suspicious when the data are considered as a sample from a normal distribution: they may have been incorrectly recorded (or measured), of course, or they may be a genuine reflection of a basic error in assuming an underlying normal distribution.

Clearly such considerations are vitally important for proper statistical practice. We need a battery of techniques for assessing the integrity of a set of data with respect to an assumed model. We also need, and should welcome the recent revitalization of interest in, statistical methods which are *robust* (i.e. retain reasonable validity) over a range of possible probability models. As a particular aspect of these interests we require methods for assessing, rejecting, making allowances for, or minimizing the influence of, outlying observations. Such methods do exist, but they tend to appear in a scattered form throughout the statistical literature. The aim of this book is to bring them together and to present a unified discussion of ways of handling outliers in statistical data, in relation to the nature of the outliers and to the aims of the investigation.

At this stage we must make clear what we mean by an **outlier**. (But note that, as explained on page 30, some writers use the word *outlier* in a different sense.) We shall define an outlier in a set of data to be *an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data*. The phrase 'appears to be inconsistent' is crucial. It is a matter of subjective judgement on the part of the observer whether or not he picks out some observation (or set of observations) for scrutiny. What really worries him is whether or not some observations are genuine members of the main population. If they are not, but are *contaminants*, they may frustrate his attempts to draw inferences about that population. Of course, any contaminants which occur in the midst of the data set will not be conspicuous; also, they are unlikely to seriously

## Introduction

distort the inference process. But the observer (not only will it appear surprisingly extreme). Should such a situation they may, by their very nature, attempt to represent the population of parameters in some model for the data. It starts in the following way. We examine (in the sense described above). Not only outliers, and what methods can be used to find their values, prior to processing the data depends on the form of the population specific to, any postulated model. The processing of outliers take on an

One conceptual difficulty needs to be divided on precisely when it is just a dispute that it is reasonable to assume an observation, or mis-recording: these considerations such as the sheer possibility of human error. It is sometimes claimed that 'genuine' outliers and that if no observation appears unreasonable observation by special treatment, is invalid.

However, two factors lead us to consider a variety of methods have been proposed. These are widely used by statisticians in a classified manner for their better understanding. More fundamentally, the concept of an outlier. Suppose we think that a sample observation seems intuitively unreasonable. A statistical test confirms its unacceptability. An unreasonable observation should be rejected on grounds that it would not have appeared if the data were from a log-normal distribution, as a more reasonable observation *did appear* as a more reasonable observation. It presumably had some basis as an outlier. It allows a more appropriate model to be fitted. The dangers that may arise from basing inferences on a very much the way in which outliers are handled in the literature, and it seems a fruitful area for research.

We shall consider in subsequent chapters dealing with outliers in different situations. Outliers inevitably arise. But it is now time to consider some of the basic distinctions between observations, outliers, contaminants

mean by an **outlier**. (But note that, as the word **outlier** in a different sense.) We have to be an *observation (or subset of data) consistent with the remainder of that set of data*. This is crucial. It is a matter of subjective judgment whether or not he picks out some observations for scrutiny. What really worries him is *genuine members of the main population*. He may frustrate his attempts to draw conclusions about any contaminants which occur in the population; also, they are unlikely to seriously

We shall consider in subsequent chapters the various methods available for dealing with outliers in different situations, including some of the difficulties that inevitably arise. But it is now time to define, and explain, a little more formally some of the basic distinctions such as those which distinguish **extreme observations**, **outliers**, **contaminants**, etc.

Suppose we have a random univariate sample of size  $n$

$$x_1, x_2, \dots, x_n$$

from a distribution whose form we will denote by  $F$ .

We will write the ordered sample (from smallest to largest value) as

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}.$$

The observations  $x_{(1)}$  and  $x_{(n)}$  are the sample **extremes**. Whether we declare either of them to be an **outlier** depends on (at least an informal) consideration of how they appear in relation to the postulated model,  $F$ . In Figure 1.2(a), neither  $x_{(1)}$  nor  $x_{(n)}$  appears to be outlying. In Figure 1.2(b),  $x_{(n)}$  is an *upper outlier* and  $x_{(1)}$  also gives some cause for concern. We *might* declare  $x_{(1)}$  to be a *lower outlier* (so that  $x_{(1)}, x_{(n)}$  is an *outlier pair*). So we see that extreme values may or may not be outliers. Any outliers, however, are always extreme (or relatively extreme) values in the sample! (Note that we might declare  $x_{(n-1)}, x_{(n)}$  to be an *upper outlier pair*.)

Suppose now that not all of the observations come from the distribution  $F$ , but one or two come from a distribution  $G$  which has 'slipped' upward relative to  $F$  (i.e. it has a larger mean). The observations from  $G$  are termed **contaminants**. Such contaminants may appear as extremes, but *need* not do so. Figure 1.2(c) shows two contaminants (indicated  $\bullet$ ), one of which is the upper extreme, the other is in the midst of the sample. But  $x_{(n)}$ , whilst an extreme and a contaminant, is not an outlier! In Figure 1.2(d), however, we see in contrast a *non-extreme* contaminant which is nonetheless outlying: as one of the outlier pair  $x_{(n-1)}, x_{(n)}$ .

So we see that outliers may or may not be contaminants; contaminants may or may not be outliers! Of course, we have no way of knowing whether or not any

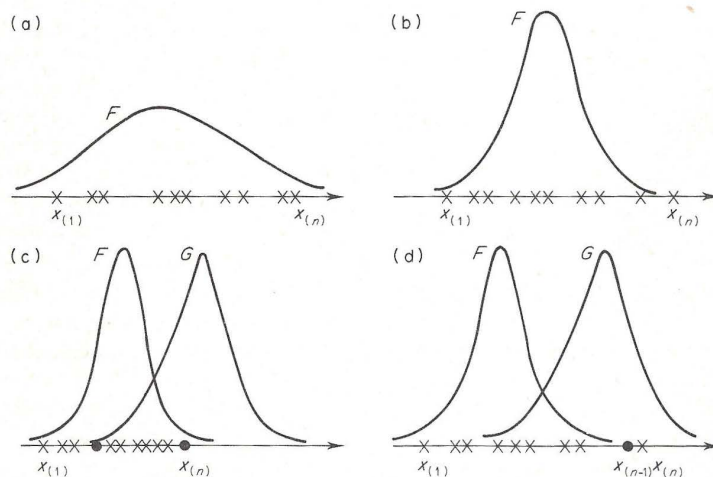


Figure 1.2 Extremes, outliers, and contaminants

## Introduction

observation is a contaminant, as the possible manifestation of methods for examining outliers.

In the following sections of the book, we will discuss briefly, to illustrate the ways in which outliers can arise. These also serve to motivate the book in detail throughout the book.

## 1.1 HUMAN ERROR AND

There is a class of situations where the possibility of dealing with them is obvious. In many cases, human errors lead to blatant outliers, which are regarded to practical factors resulting from human error.

In a study of low-temperature weather in the British Isles, Barnett and his colleagues recorded temperatures over several years. In 1963, temperatures were recorded in Wick in northern Scotland the late evening of 31 December 1963.

43, 43, 41,

The values 58, 58 for midnight and 59 for 1.00 are outliers in this time-series section. Whether or not they are genuine outliers is not clear. On further investigation, it is perfectly reasonable! At midnight, the measuring unit from degrees Fahrenheit to degrees Celsius, the values appear as follows (to one decimal place):

43, 43, 41,

These are much more satisfactory. In his Presidential address to the Royal Society, an interesting example of the way in which an outlier in a set of data. He reported on poultry. For one bird, the weights at intervals were shown as:

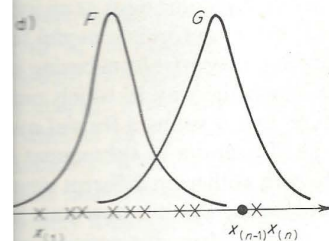
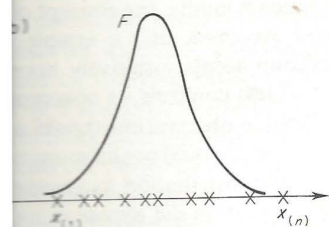
1.20,

From the manner in which the

sample of size  $n$

note by  $F$ .  
smallest to largest value) as

multiple extremes. Whether we declare (least an informal) consideration of a model,  $F$ . In Figure 1.2(a), neither in 1.2(b),  $x_{(n)}$  is an upper outlier and  $x_{(1)}$  declare  $x_{(1)}$  to be a lower outlier (so extreme values may or may not be extreme (or relatively extreme) values  $x_{(n-1)}, x_{(n)}$  to be an upper outlier pair.)  $x_{(1)}$  comes from the distribution  $F$ , but  $x_{(n)}$  has 'slipped' upward relative to  $F$  from  $G$  are termed **contaminants**. Such need not do so. Figure 1.2(c) shows  $x_{(1)}$  is the upper extreme, the other is in extreme and a contaminant, is not an contrast a non-extreme contaminant the outlier pair  $x_{(n-1)}, x_{(n)}$ . contaminants; contaminants may or way of knowing whether or not any



extremes, and contaminants

observation is a contaminant. All we can do is concentrate attention on outliers as the possible manifestation of contamination and we shall see how statistical methods for examining outliers aim specifically at this prospect.

In the following sections of this chapter some practical examples are discussed briefly, to illustrate the ways in which the outlier problem may present itself. These also serve to motivate the different forms of statistical analysis considered in detail throughout the book.

## 1.1 HUMAN ERROR AND IGNORANCE

There is a class of situations where outliers are readily handled: where the manner of dealing with them is obvious and non-controversial. Such is the situation when human errors lead to blatantly incorrect recording of data, or where lack of regard to practical factors results in serious misinterpretation.

In a study of low-temperature probabilities throughout the winter months in the British Isles, Barnett and Lewis (1967) analysed extensive data on hourly temperatures over several years and various geographical sites. In the main, temperatures were recorded in degrees Fahrenheit. Among extensive data for Wick in northern Scotland the following hourly temperatures were found for the late evening of 31 December 1960 and early morning of 1 January 1961:

43, 43, 41, 41, 41, 42, 43, 58, 58, 41, 41

↑  
midnight

The values 58, 58 for midnight and 1.00 a.m. stand out in severe contrast to the others in this time-series section, and initially give one grounds for concern as to whether or not they are genuine—it seemed very warm at midnight for New Year in northern Scotland. On further enquiry, however, they were found to be perfectly reasonable! At midnight the Meteorological Office changed its measuring unit from degrees Fahrenheit to  $\frac{1}{10}^{\circ}\text{C}$ , so that in degrees Fahrenheit the values appear as follows (to the nearest degree):

43, 43, 41, 41, 41, 42, 43, 42, 42, 39, 39

↑  
midnight

These are much more satisfactory; so much for the 'outliers' 58 and 58!

In his Presidential address to the Royal Statistical Society, Finney (1974) gives an interesting example of the way in which a recording error may appear as an outlier in a set of data. He reports on measurements taken on the growth of poultry. For one bird, the weights (in kg) for successive weighings at regular intervals were shown as:

1.20, 1.60, 1.90, 1.55, 2.20, 2.25

From the manner in which the weights were determined and recorded it was

clearly possible to commit recording errors of 0.50 kg or 1.00 kg. It seems highly likely that the fourth reading is a mis-recording of what should have been 2.05 kg. This conclusion is supported by biological considerations, and by the overall pattern of results for a large sample of birds.

Whilst one cannot be one hundred per cent certain of the interpretation of 1.55 in this last example, some instances arise where recorded values are absolutely impossible. In a recent student exercise, records were kept of the numbers of times a six occurred in ten throws of ten dice. One student returned the results:

2, 0, 3, 12, 2, 0, 1, 1, 3

Clearly the value 12 cannot be genuine. Furthermore, since ten observations were asked for, it would seem merely that a comma has been omitted in a sequence of numbers which should have read:

2, 0, 3, 1, 2, 2, 0, 1, 1, 3

These examples all illustrate the effect of non-statistical factors attributable, to a greater or lesser extent, to lack of care in the recording or presentation of data. Processing of outliers (or spurious values of any sort) in such cases is not a matter of statistical analysis, but of native wit! Correct action also raises no difficulties in most cases. In the examples which follow, however, we will be very much concerned with statistical factors which affect the occurrence and treatment of outliers.

## 1.2 OUTLIERS IN RELATION TO PROBABILITY MODELS

An interesting area of statistical enquiry is the manner in which foodstuffs and general household products are purchased. Table 1.1 shows a frequency distribution for the number of packets of a particular brand and packet size of breakfast cereal (Chatfield, 1974) bought over a given period of time.

One aspect which stands out in these data are the two single instances of purchases of 39 and 52 packets, which seem somewhat out of line with other observations. How should we react to these outliers? There are various possibilities.

Table 1.1 Frequency distribution for numbers of packets of cereal purchased over 13 weeks by 2000 customers

|                |      |     |     |    |    |    |    |    |    |    |    |    |    |    |    |    |      |    |    |
|----------------|------|-----|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|------|----|----|
| No. of packets | 0    | 1   | 2   | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16   | 17 | 18 |
| Frequency      | 1149 | 199 | 129 | 87 | 71 | 43 | 49 | 46 | 44 | 24 | 45 | 22 | 23 | 33 | 8  | 2  | 7    | 2  | 3  |
| No. of packets | 19   | 20  | 21  | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35   | 36 |    |
| Frequency      | 1    | 2   | 0   | 0  | 1  | 0  | 1  | 3  | 2  | 0  | 1  | 1  | 0  | 0  | 0  | 0  | 0    | 0  |    |
| No. of packets | 37   | 38  | 39  | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | ≥ 53 |    |    |
| Frequency      | 0    | 0   | 1   | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0    |    |    |

## Introduction

(i) We might be interested in purchasing behaviour—perhaps shopping, or indicating hoarding expected price rises. The observation is of prime interest. We might try to find the majority of the data, relating to the purchase of cereal, to consider the outliers relative to the main body of the data, then lead us to conclude that the interest is in studying atypical purchases. These outliers constitute a set of special importance. The 'outlying' individuals may not be representative, with a strong influence in social policy, for holding reasonable policies for holding cereal, accumulated with others in alternative information on this special interest.

(ii) If, in contrast with (i), we are interested in purchasing behaviour, any outlier considerations suggest a possibility that certain observations which arise for purely technical reasons, detection and rejection then aid the fit; we may be better able to estimate the unrepresentative, values can series. For example, as a first approximation, modified Poisson distribution, in the case of non-purchasers and potential purchasers of packets of cereal according to a Poisson distribution, we would have a function

$$p(0) = \theta + (1 - \theta)e^{-\lambda}$$

$$p(x) = (1 - \theta)e^{-\lambda} \frac{\lambda^x}{x!}$$

If this model is reasonable, estimates about purchasing behaviour. Outliers and the estimates, and need to be consistent with the data, its adoption is reinforced (and perhaps 39), we must doubt the reason spurious, and should be replaced by a more sophisticated model, where  $\lambda$  varies from individual to individual. Ehrenberg, and Goodhardt (1966) model this policy promotes. What proper account of values which

0.50 kg or 1.00 kg. It seems highly of what should have been 2.05 kg. considerations, and by the overall

certain of the interpretation of 1.55 are recorded values are absolutely were kept of the numbers of times student returned the results:

1, 3

more, since ten observations were has been omitted in a sequence of

1, 3

n-statistical factors attributable, to recording or presentation of data. (any sort) in such cases is not a matter of action also raises no difficulties in however, we will be very much at the occurrence and treatment of

## PROBABILITY MODELS

the manner in which foodstuffs and ed. Table 1.1 shows a frequency particular brand and packet size of er a given period of time.

ata are the two single instances of a somewhat out of line with other these outliers? There are various

ers of packets of cereal purchased 0 customers

|    |    |    |    |    |    |    |    |      |    |    |
|----|----|----|----|----|----|----|----|------|----|----|
| 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16   | 17 | 18 |
| 44 | 24 | 45 | 22 | 23 | 33 | 8  | 2  | 7    | 2  | 3  |
| 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35   | 36 |    |
| 2  | 0  | 1  | 1  | 0  | 0  | 0  | 0  | 0    | 0  |    |
| 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | ≥ 53 |    |    |
| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0    |    |    |

(i) We might be interested in identifying out-of-the-ordinary patterns of purchasing behaviour—perhaps due to institutional rather than personal shopping, or indicating hoarding of products in times of potential shortage or expected price rises. The observations 39 and 52 (possibly others) then become of prime interest. We might try to fit some probability distribution to represent the majority of the data, relating to 'reasonable private purchase', and then attempt to consider the outliers relative to that distribution. An appropriate model might then lead us to conclude that the observations 39 and 52 are anomalous. If our interest is in studying atypical purchasing behaviour, the identified individuals constitute a set of special importance in their own right. Follow-up enquiries for the 'outlying' individuals may reveal special attitudes or patterns of behaviour with a strong influence in sociological terms or in terms of the choice of reasonable policies for holding stocks of the product. The outliers may be accumulated with others in alternative sets of data, to build up more comprehensive information on this special group.

(ii) If, in contrast with (i), prime interest rests on the *overall* pattern of purchasing behaviour, any outliers play a subsidiary role. It may be that general considerations suggest a possible model for the distribution of purchases, and that certain observations which appear as outliers merely cloud the issue; they arise for purely technical reasons unimportant in the nature of our enquiry. Their detection and rejection then aid the study of the basic model. We can better assess its fit; we may be better able to estimate relevant parameters. One or two extreme, unrepresentative, values can seriously distort the fitting or estimation process. For example, as a first approximation we might set up a model in the form of a modified Poisson distribution, in which the population is divided into two groups (non-purchasers and potential purchasers) where potential purchasers buy packets of cereal according to a Poisson process of rate  $\lambda$ . If  $\theta$  is the proportion of non-purchasers, we would have a probability distribution for  $X$ , the number of packets purchased over the observation period of length  $T$ , with probability function

$$p(0) = \theta + (1 - \theta)e^{-\lambda T}$$

$$p(x) = (1 - \theta)(\lambda T)^x e^{-\lambda T} / x! \quad x = 1, 2, \dots$$

If this model is reasonable, estimation of  $\theta$  and  $\lambda$  provides useful information about purchasing behaviour. Outlying observations can seriously affect the fit, and the estimates, and need to be carefully examined. If the model fits well to *all* the data, its adoption is reinforced. If it fits well apart from the observation 52 (and perhaps 39), we must decide whether the outlier 52 (and 39) is for some reason spurious, and should be omitted in further study, or whether we might need a more sophisticated model. For example, a compound Poisson distribution where  $\lambda$  varies from individual to individual might be warranted. See Chatfield, Ehrenberg, and Goodhardt (1966) for evidence supporting the negative binomial model this policy promotes. Whatever our conclusion, its basis rests on taking proper account of values which are unrepresentative of the original model.

(iii) Again we must be on the look-out for non-statistical factors influencing the occurrence of outliers. In Table 1.1 we note that the data refer to numbers of purchases over a *thirteen-week* period. Might it be more than coincidence that the two outliers, 52 and 39, are both multiples of 13? There seem to be further 'blips' in the frequency distribution at 26 (3 purchases) and 13 (33 purchases). The outliers call our attention to another possible ingredient in the purchasing model, corresponding to automatic regular purchases of 4, 3, 2, or 1 packets per week. A better model might be one with three components reflecting

- (a) lack of interest in the product,
- (b) a distribution of regular purchases of 1, 2, ... packets per week, and
- (c) a Poisson (or mixed Poisson) process pattern of casual purchases.

In (ii) above we remarked on the way in which outliers may influence the propriety of different methods of estimating parameters in the basic model. Let us consider a more specific example. Suppose the following random observations were obtained for some variable of interest:

1.74, 1.46, -1.28, -0.02, -0.40, 0.02, **3.89**, 1.35, -0.10, 1.71

We wish to estimate the 'centre' of the parent population. Initial considerations suggest that the population may be normal,  $N(\theta, 1)$ , with mean  $\theta$  and variance 1, so the sample mean would clearly be a sensible form of estimator. But the value 3.89 makes us suspicious of the  $N(\theta, 1)$  assumption! *In fact*, these data were generated as a random sample from a Cauchy distribution, with probability density function

$$f(x) = \frac{1}{\pi} (1 + x^2)^{-1}.$$

The sample mean here is not even *consistent*, let alone of reasonable efficiency, and we should have made very poor use of our data in the estimation procedure had we used it as an estimator of location.

Observations far removed from the main body of the sample arise naturally in sampling from a Cauchy distribution. This illustrates very clearly the relative nature of the element of 'surprise' in our declaration of an outlier. An outlier is not only an extreme, but is surprisingly extreme. But what is surprising in relation to the  $N(\theta, 1)$  model may not be so in relation to the Cauchy model. Observations would need to be an order of magnitude 'more extreme' for us to declare them outliers relative to a Cauchy model. Such a phenomenon occurs not infrequently in biological contexts. For example, the distribution of the number of cones on a fir tree for trees in a given area of forest, or the distribution of the number of lepidoptera (butterflies or moths) of the same species present and observed in a particular location, are both characterized by high skewness. A typical sample from this latter type of distribution is given below; it refers to the number of individuals of a given species in a random sample of nocturnal *macrolepidoptera* caught in a light-trap at Rothamsted (Fisher, Corbet and Williams, 1943):

11, 54, 5, 7, 4, 15, **560**, 18, 120, 24, 3, 51, 3, 12, 84

Here we have a situation in which the outlier is an inherent feature of the natural

### 1.3 OUTLIERS IN MORE STRUCTURED SITUATIONS

The examination of univariate parameters, whilst an important aim, is not the only aim of statistical analysis. In many structured situations. For example, the study of a variable of principal interest, leads to the study of regression. Or again, concern for the influence of outliers on the principal variable leads to the study of robust techniques. These various models are parts in the study of multivariate analysis.

In all of these more structured situations, the time to time, unrepresentative observations may be important as in the simple univariate case and make allowance for outliers in the estimation techniques. Outliers may, as before, be indicative of incorrect specification of the model, with consequential implications for the procedures. With such more structured situations, the methods tend to be less intuitively apparent and more formal methods for their rejection or treatment.

The outlier problem is a complex one. The apparently simple formulation of the problem has many ramifications begin to appear. It is not surprising that perceiving them to be 'surprisingly surprising'. From the examples given, it is clear that the treatment of outliers depends on the nature of the distribution. The relative nature of the outlier is stressed. Again, in what way does the outlier affect the relationship of the outlying values to the main body of the data? This is the situation in which inspection? This is the situation in which inspection? For a univariate sample the problem is to detect the outlier value in a designed experiment - square experiment. A simple visual inspection may reveal the outlier. Its presence in the data may be model are fitted and the deviate values are tabulated. Another situation is the study of outlying values in a regression analysis nowadays which advises that the outlier is an essential part of any regression analysis.