

# *Outliers in Statistical Data*

VIC BARNETT

*University of Sheffield*

and

TOBY LEWIS

*University of Hull*

*John Wiley & Sons*

*Chichester · New York · Brisbane · Toronto*

Copyright © 1978 by John Wiley & Sons Ltd.

Reprinted February 1979

Reprinted June 1980

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. Barnett/Lewis  
(Wiley series in probability and mathematical  
statistics)

Bibliography: p.  
Includes index.  
1. Outliers (Statistics) I. Lewis, Tobias, joint author. II. Title.  
QA276.B2849 519.5 77-21024

ISBN 0 471 99599 1

Typeset by The Universities Press, Belfast, Northern Ireland  
Printed and bound in Great Britain  
at The Pitman Press, Bath



## *Preface*

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

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.

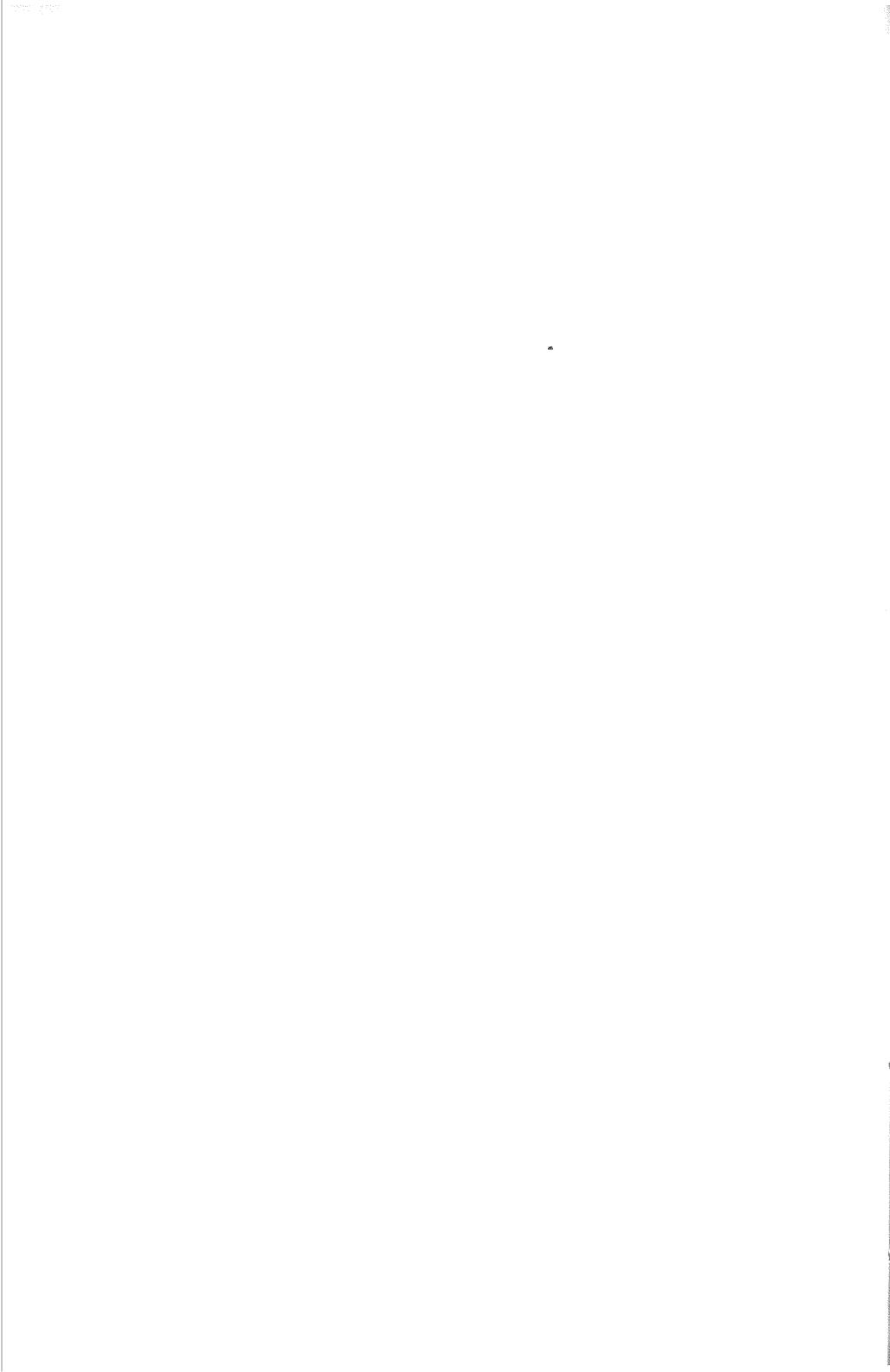
Copious references are given throughout the text to source material and to further work on the various topics. They are gathered together in the section entitled 'References and Bibliography' with appropriate page references to places of principal relevance in the text. Additional references augment 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.

*July, 1977*

VIC BARNETT  
TOBY LEWIS



# *Contents*

## CHAPTER 1 INTRODUCTION 1

- 1.1 *Human error and ignorance* 6
- 1.2 *Outliers in relation to probability models* 7
- 1.3 *Outliers in more structured situations* 10
- 1.4 *Bayesian and non-parametric methods* 15
- 1.5 *Survey of outlier problems* 16

## CHAPTER 2 WHAT SHOULD ONE DO ABOUT OUTLYING OBSERVATIONS? 18

- 2.1 *Early informal approaches* 18
- 2.2 *Various aims* 22
- 2.3 *Models for discordancy* 28
- 2.4 *Test statistics* 38
- 2.5 *Statistical principles underlying tests of discordancy* 41
- 2.6 *Accommodation of outliers: robust estimation and testing* 46

## CHAPTER 3 DISCORDANCY TESTS FOR OUTLIERS IN UNIVARIATE SAMPLES 52

- 3.1 *Statistical bases for construction of tests* 56
  - 3.1.1 *Inclusive and exclusive measures, and a recursive algorithm for the null distribution of a test statistic* 61
- 3.2 *Performance criteria of tests* 64
- 3.3 *The multiple outlier problem* 68
  - 3.3.1 *Block procedures for multiple outliers in univariate samples* 71
  - 3.3.2. *Consecutive procedures for multiple outliers in univariate samples* 73
- 3.4 *Discordancy tests for practical use* 75
  - 3.4.1 *Guide to use of the tests* 75

3.4.2	<i>Discordancy tests for gamma (including exponential) samples</i>	76
3.4.3	<i>Discordancy tests for normal samples</i>	89
3.4.4	<i>Discordancy tests for samples from other distributions</i>	115

## CHAPTER 4 ACCOMMODATION OF OUTLIERS IN UNIVARIATE SAMPLES: ROBUST ESTIMATION AND TESTING 126

4.1	<i>Performance criteria</i>	130
4.1.1	<i>Efficiency measures for estimators</i>	130
4.1.2	<i>The qualitative approach: influence curves</i>	136
4.1.3	<i>Robustness of confidence intervals</i>	141
4.1.4	<i>Robustness of significance tests</i>	142
4.2	<i>General methods of accommodation</i>	144
4.2.1	<i>Estimation of location</i>	144
4.2.2	<i>Performance characteristics of location estimators</i>	155
4.2.3	<i>Estimation of scale or dispersion</i>	158
4.2.4	<i>Studentized location estimates, tests, and confidence intervals</i>	160
4.3	<i>Accommodation of outliers in univariate normal samples</i>	163
4.4	<i>Accommodation of outliers in exponential samples</i>	171

## CHAPTER 5 OUTLYING SUB-SAMPLES: SLIPPAGE TESTS 174

5.1.	<i>Non-parametric slippage tests</i>	176
5.1.1	<i>Non-parametric tests for slippage of a single population</i>	176
5.1.2	<i>Non-parametric tests for slippage of several populations: multiple comparisons</i>	183
5.2	<i>The slippage model</i>	186
5.3	<i>Parametric slippage tests</i>	187
5.3.1	<i>Normal samples</i>	188
5.3.2	<i>General slippage tests</i>	197
5.3.3	<i>Non-normal samples</i>	201
5.3.4	<i>Group parametric slippage tests</i>	204
5.4	<i>Other slippage work</i>	205

## CHAPTER 6 OUTLIERS IN MULTIVARIATE DATA 208

6.1	<i>Outliers in multivariate normal samples</i>	209
6.2	<i>Informal detection of multivariate outliers</i>	219
6.2.1	<i>Marginal outliers</i>	220
6.2.2	<i>Linear constraints</i>	221
6.2.3	<i>Graphical methods</i>	221
6.2.4	<i>Principal component analysis method</i>	223
6.2.5	<i>Use of generalized distances</i>	224
6.2.6	<i>Fourier-type representation</i>	227

- 6.2.7 *Correlation methods* 227
- 6.2.8 *A ‘gap test’ for multivariate outliers* 229
- 6.3 *Accommodation of multivariate outliers* 231

## CHAPTER 7 OUTLIERS IN DESIGNED EXPERIMENTS, REGRESSION AND IN TIME-SERIES 234

- 7.1 *Outliers in designed experiments* 238
  - 7.1.1 *Discordancy tests based on residuals* 238
  - 7.1.2 *Residual-based accommodation procedures* 246
  - 7.1.3 *Graphical methods* 247
  - 7.1.4 *Non-residual-based methods* 249
  - 7.1.5 *Non-parametric, and Bayesian, methods* 251
- 7.2 *Outliers in regression* 252
  - 7.2.1 *Outliers in linear regression* 252
  - 7.2.2 *Multiple regression* 256
- 7.3 *Outliers with general linear models* 257
  - 7.3.1 *Residual-based methods* 257
  - 7.3.2 *Non-residual-based methods* 264
- 7.4 *Outliers in time-series* 266

## CHAPTER 8 BAYESIAN AND NON-PARAMETRIC APPROACHES 269

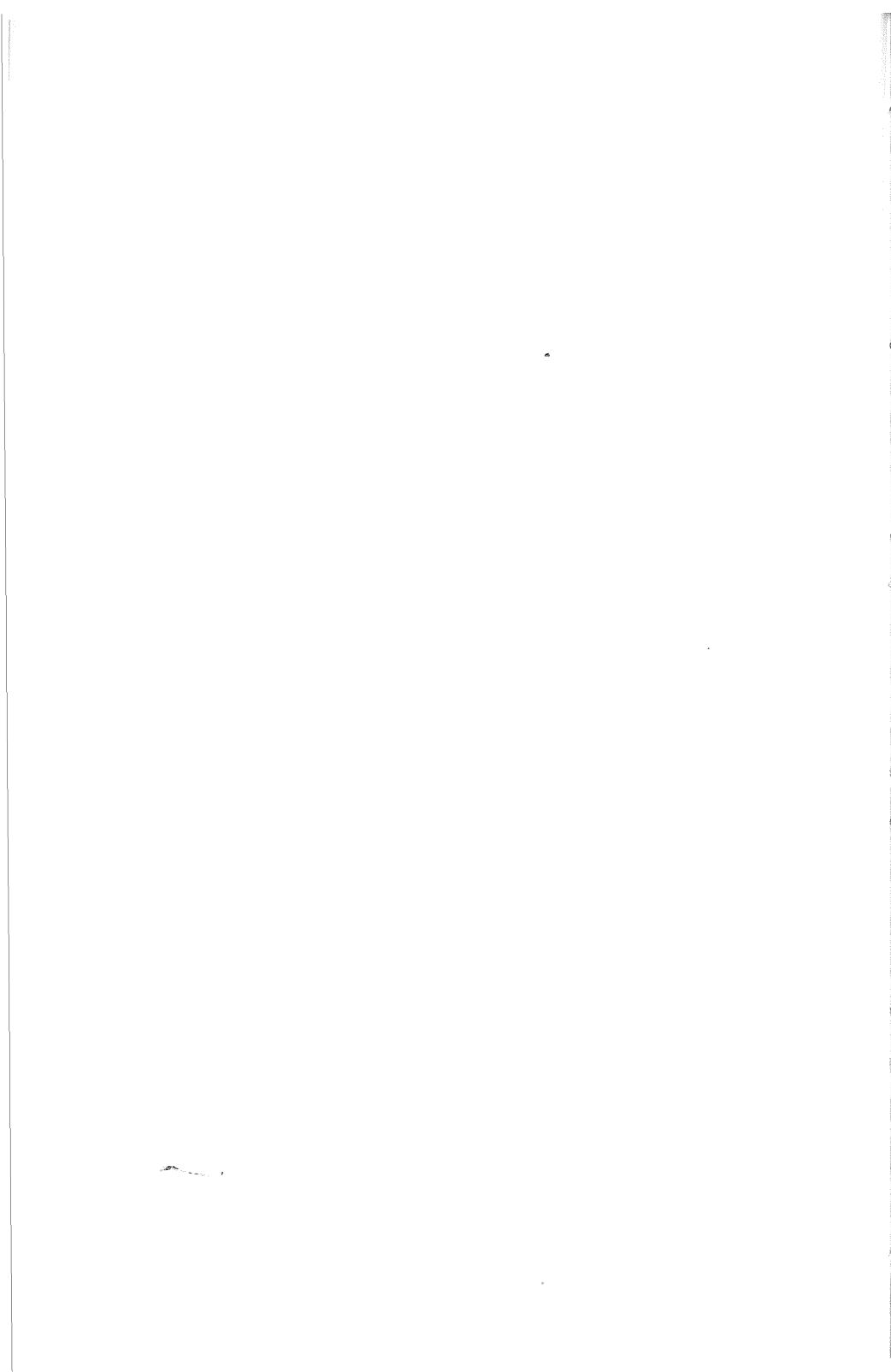
- 8.1 *Bayesian methods* 269
  - 8.1.1 *Bayesian ‘tests of discordancy’* 269
  - 8.1.2 *Bayesian accommodation of outliers* 277
- 8.2 *Non-parametric methods* 282

## CHAPTER 9 PERSPECTIVE 286

## APPENDIX: STATISTICAL TABLES 289

## REFERENCES AND BIBLIOGRAPHY 337

## INDEX 357



## CHAPTER 1

### *Introduction*

From the earliest gropings 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. (Copyright © 1961 by Regents of the University of California; reprinted by permission of the University of California Press)

In the light of developments in outlier methodology over the last 15 years, Ferguson's formulation is unduly restrictive in various ways, as we shall see; for example, outlying values are not necessary ‘bad’ or ‘erroneous’, and the experimenter may be tempted in some situations not to ‘throw away’ the discordant value but to welcome it as an indication of, say, some unexpectedly useful industrial treatment or agricultural variety. However, the passage brings out most of the essentials of the outlier situation, and will serve as a 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 judgments 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  $n \dots$  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 method is one of the earliest extant for dealing with outliers, and dates from the middle of the nineteenth century (Chauvenet, 1863). We return to it in Section 2.1.

Some data which have attracted much interest among statisticians over the years are the results given by Mercer and Hall (1912) on yields of wheat grain and straw for 500 similar-sized plots of soil planted with wheat over a rectangular field. Both grain yields, and straw yields, seem to provide a good fit to normal distributions. If one is concerned with edge effects and looks at the 25 grain yields along the southern boundary of the field, results (in lb) are found as shown in the second row from the bottom in Figure 1.1, which is part of the original table given by Mercer and Hall. (Upright numerals are grain yields; italic, straw yields.)

*On these figures alone* one might be rather worried about the value 5.09 for grain yield in the fourth plot from the western edge. Mercer and Hall were not concerned only with the edge yields, nor were they on the look out for ‘outliers’. But even at the time of their work (1912) there were available better methods than Chauvenet’s for detecting outliers. Wright’s rule (1884), for example, rejected any observation distant more than  $\pm 3.37$  estimated standard deviations from the sample mean.

What happens if we apply Chauvenet’s, or Wright’s, method to the 25 southern edge grain yields? The sample mean is  $m = 3.95$  and the estimated standard deviation  $s = 0.463$ . Neither Chauvenet nor Wright distinguished between  $s$  and the population measure  $\sigma$ . For rejecting the observation 5.09 on the Chauvenet principle we need  $|5.09 - m|/s$  to exceed 2.33. But  $(5.09 - 3.95)/0.463 = 2.46$ , so that on this basic there would be some cause for concern about the value 5.09. The opposite conclusion is reached on Wright’s principle!

With the development during this century of more formal approaches to the statistical analysis of data, objectives have become clearer, principles

**Figure 1.1** Extract from Mercer and Hall (1912) data (reproduced by permission of Cambridge University Press).

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 may be 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 the basic impropriety of 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. As a particular aspect of this we need methods for assessing, rejecting, or making allowances for, 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*. 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, they may frustrate his attempts to draw inferences about that population. Any small number of spurious observations in the midst of the data may not be conspicuous in any case: they are perhaps unlikely to seriously distort the inference process. But what characterizes the 'outlier' is its impact on the observer (it appears *extreme* in some

way). Should such observations be foreign to the main population they may, by their very nature, cause difficulties in the attempt to represent the population: they can grossly contaminate estimates (or tests) of parameters in some model for the population. Accordingly the outlier problem takes the following form. We examine the data set. We decide that outliers exist (in the sense described above). How should we react to them? What methods can be used to support rejecting the outlying observations, or adjusting their values, prior to processing the principal mass of data? Clearly, the answer depends on the form of the population; techniques will be conditioned by, and specific to, any postulated model for that population. Thus, methods for the processing of outliers take on an entirely *relative* form. It may be, of course, that we do not go beyond the rejection stage in some cases. Our interest rests in identifying foreign observations as matters of major concern: they indicate particular matters of practical interest.

One conceptual difficulty needs to be recognized at the outset. Opinion is divided on precisely when it is justifiable to scrutinize outliers. There is little dispute that it is reasonable when outliers exist in the form of errors of observation, or mis-recording, that is, when they can be substantiated by practical considerations such as the sheer impossibility of a recorded value, or an obvious human error. It is sometimes claimed (as remarked above) that these are the only genuine 'outliers' and that if no such tangible explanation can be found for apparently unreasonable observations then their rejection, or accommodation by special treatment, is invalid. However, two factors lead us to reject this nihilistic attitude. In the first place a variety of methods have been proposed for dealing with 'non-tangible' outliers; these are used by statisticians and it seems desirable to present them in a classified manner for their better understanding and application. Secondly, and more fundamentally, the examination of an outlier must have propriety if viewed in relative terms. Suppose we think that a sample arises from a normal distribution but one observation seems intuitively unreasonable (it is an outlier); an appropriate statistical test confirms its unacceptability. It seems to beg the question to say that the unreasonable observation should not have been regarded as an outlier, on the grounds that it would not have appeared unreasonable if we had had in mind, say, a log-normal distribution, as a model for explaining the data. Be this as it may, the rogue observation *did appear as an outlier relative to our original model*, which presumably had some basis as an initial specification. Examination of the outlier allows a more appropriate model to be formulated, or enables us to assess any dangers that may arise from basing inferences on the normality assumption. This is very much the way in which outliers have been discussed in the statistical literature, and seems a fruitful avenue of enquiry.

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.

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:

$$43, 43, 41, 41, 41, 42, 43, 58, 58, 41, 41 \\ \uparrow \\ \text{midnight}$$

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 recording 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):

$$43, 43, 41, 41, 41, 42, 43, 42, 42, 39, 39 \\ \uparrow \\ \text{midnight}$$

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 recording errors may appear as outliers 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, perhaps, 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

(1) 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 method might lead us to conclude that the observations 39 and 52 are anomalous and might cast doubt in terms of poorness of fit on some of the observations in the range 20–30. 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.

(2) If, in contrast with (1), 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

$$\begin{aligned} p(0) &= \theta + (1 - \theta)e^{-\lambda T} \\ p(x) &= (1 - \theta)(\lambda T)^x e^{-\lambda T}/x! \quad x = 1, 2, \dots \end{aligned}$$

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 is often warranted. Current work indeed favours the negative binomial model this policy promotes (Chatfield, Ehrenberg, and Goodhardt, 1966). Whatever our conclusion, its basis rests on assessing values which are unrepresentative of the original model.

(3) 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

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

In (2) 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,  $N(\theta, 1)$ , 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, and this contrasts with the common situation where the presence of an outlier suggests the possible inappropriateness of a model. A similar 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* 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 an outlying value (the value 560) is an inherent feature of the natural data pattern, and in no way anomalous.

### 1.3 OUTLIERS IN MORE STRUCTURED SITUATIONS

The examination of univariate samples for fitting models and estimating parameters, whilst an important part of statistical practice, has somewhat limited aims and utility. More often, and more usefully, we need to consider more structured situations. For example, an interest in the way in which observations of a variable of principal interest vary with values of other variables, or vary with time, leads to the study of *regression* models, and *time-series* models, respectively. Or again, concern for the influence of different qualitative factors on the principal variable leads to additive models analysed by *analysis of variance* techniques. These various models and techniques have, of course, their counterparts in the study of *multivariate* data.

In all these more structured cases we must also expect to encounter, from time to time, unrepresentative data in the form of outliers. Here it is just as important as in the simple univariate sample to be able to interpret and accommodate outliers by using appropriately designed statistical techniques. Outliers may, as before, be of intrinsic interest in their own right, or may be indicative of inappropriate specifications of the error structure, or of the basic model, with consequential implications for the use of appropriate inference procedures. With more structured data two complications arise: suspicious observations tend to be less intuitively apparent, more hidden, in the data mass, and formal methods for their rejection or their accommodation are less highly developed.

The outlier problem is a complex one. As soon as one starts thinking of an apparently simple formulation such as the above-quoted one by Ferguson, the ramifications begin to appear. Ferguson speaks of identifying outliers by perceiving them to be 'surprisingly far away from the main group'. But what is *surprising*? From the examples we have already considered it is clear that treatment of outliers depends in an essential way on the assumed underlying distribution. Again, in what way does the person making the judgement assess the relationship of the outlying values to the main group? Does he do it by simple inspection? This is the situation that comes first to mind and has been illustrated. But the outlier may be a value in a designed experiment—one of the observed responses, say, in a Latin-square experiment. A simple

visual inspection of the table of responses will not reveal that it is an outlier. The evidence for this only comes to light when the parameters of the model are fitted and the deviations of the observed responses from the fitted values are tabulated. Another situation of this kind arises in the context of outlying values in a regression analysis. There is a strong body of opinion nowadays which advises that examination of residuals should be carried out as an essential part of any regression analysis. In this case also we may be concerned with outliers—outlying *residuals*. Why is this different from Ferguson's outlying values in univariate distributions? For one thing, the residuals are not independent; this may make outliers difficult to judge and also complicates the methodology.

At this stage, it is useful to consider one or two practical examples of the existence of outliers in regression, time-series, and other problems. No attempt will be made here to discuss implications in any detail (but see Chapter 7).

#### (a) Regression models

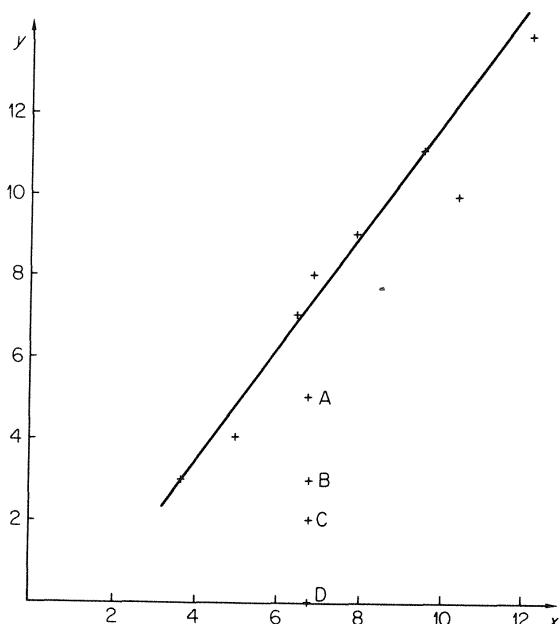
The linear regression of one random variable,  $Y$ , on a controlled variable,  $X$ , (or conditional on the observed values of a *random* variable,  $X$ ) is a model widely used for an initial study of the way in which  $Y$  varies with  $X$ . Some data by Cruikshank, reported by Quenouille (1953) show measurements ( $y_i$ ) of the coronary flow in a cat at twelve different times, with the associated auricular pressure values ( $x_i$ ). The scatter diagram of the results is given in Figure 1.2.

Rahman (1972, p. 174) suggests fitting a linear regression of coronary flow on auricular pressure. Viewing such a proposal uncritically we are nonetheless immediately struck by the apparent linearity of the relationship revealed by the data *in the absence of the observations A, B, C, and D*. In their presence the linear model has less obvious support, but the outliers have a strange consistency. They all relate to the same auricular pressure reading of 6.8, and were apparently the last four observations taken (in order: D last). Are they for some reason of quite a different basic nature to the others; was the cat *dead*? There may well be strong grounds here for omitting the outliers A, B, C, and D from *any* analysis of the relationship between coronary flow and auricular pressure.

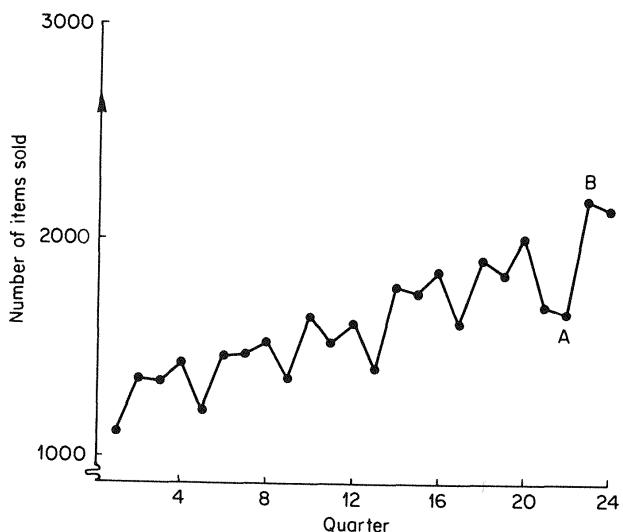
#### (b) Time-Series

Time-series data are widely studied in commercial, industrial, meteorological, and sociological processes. Outliers can again arise and cause difficulties. Consider the following examples.

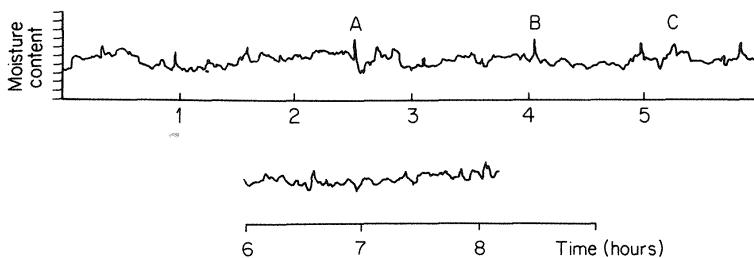
Chatfield (1975) presents some data on numbers of a particular product which are sold each quarter over a six-year period. These data are shown in Figure 1.3. Discussing the idea of outliers in the time-series data, Chatfield



**Figure 1.2** Auricular pressure ( $x$ ) and coronary flow ( $y$ ) for a cat



**Figure 1.3** Sales figures for a product over consecutive quarters for a period of six years (reproduced by permission of C. Chatfield and Chapman & Hall)



**Figure 1.4** Moisture content of Malaysian tobacco continuously monitored over eight hours

suggests that there is reason to doubt the observation A. His reasoning differs from our earlier examples, where some sort of ‘extremeness’ was the key to examining an outlying observation. Here it is the break in the ‘pattern’ of results which makes him suspect A. In previous years a relatively low sales figure in the first quarter was followed by two intermediate values and a relatively high value. In the last year, the second quarter figure (A) breaks this pattern. Perhaps economic or accounting factors have produced a spurious result, possibly compensated for by the opposingly atypical value, B. Alternatively, the results for the final year may indicate a radical change in the cyclic pattern of sales over the year.

In Figure 1.4 we see (Fenton, 1975) a continuous trace of measured moisture content of Malaysian tobacco, being automatically monitored as it flows past the recording equipment on a conveyor belt at a particular stage of the curing process. The equipment is known to suffer from occasional electronic ‘hiccoughs’ which appear as sharp spikes in the trace. An experienced observer comments that ‘A and B are clearly outliers’ in this sense, whilst ‘C is unlikely to be an outlier’.

A final example on possible outliers in time-series data is seen in the data of Table 1.2 on the percentages of road accidents each month which result in death, for the 10 years 1960–1970 in the British Isles (Chedzoy, 1973).

As a simple screening procedure for the data, which takes some account of the inevitable seasonal nature of variability in such data, Chedzoy has indicated ( / ) those observations, month by month, which are furthest from the monthly average. The standard deviations of the monthly percentages seem to be substantially reduced by omitting these extremes—see last two rows of Table 1.2. They must be reduced of course; what is relevant is whether or not they are reduced to a significantly greater extent than would be expected by chance! It is interesting to note the particular observations picked out by this process. January and February 1963 were months of particularly severe weather. We might expect an excess of minor accidents: a reduction in the death rate. Perhaps November and December 1966 ('Black

Table 1.2 Proportions of road accidents, month by month, resulting in death for the 10 years 1960–1970 in the British Isles

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Year
1961	2.31	1.94	2.24	1.85	1.96	1.71	1.82	1.65	1.91	2.09	2.13	2.25	1.975
1962	2.30	2.10	2.09	1.77	1.68	1.79	1.78	1.83	1.88	2.12	2.13	2.22	1.963
1963	1.71	1.73	1.99	1.86	1.77	1.79	1.78	1.90	1.89	2.11	2.24	2.38	1.943
1964	2.22	2.25	2.09	1.75	1.78	1.87	1.68	1.92	1.92	2.15	2.27	2.58	2.029
1965	2.16	2.24	1.91	2.08	1.81	1.89	1.70	1.86	1.97	2.06	2.10	2.31	1.998
1966	2.14	2.27	1.87	1.80	1.76	1.77	1.93	1.90	2.03	1.97	2.38	2.61	2.035
1967	2.18	2.30	2.00	1.80	1.64	1.79	1.91	1.81	2.10	1.92	2.21	2.20	1.978
1968	2.08	2.05	2.00	1.68	1.79	1.77	1.82	1.93	1.99	2.00	2.14	2.17	1.950
1969	2.22	1.89	2.02	1.98	1.80	1.84	2.11	1.95	2.23	2.27	2.22	2.47	2.090
1970	2.29	2.10	1.86	1.93	1.84	1.81	1.97	2.10	1.93	2.24	2.30	2.32	2.064
Mean of 10 values	2.16	2.09	2.01	1.85	1.78	1.80	1.85	1.89	1.99	2.09	2.21	2.35	2.003
Mean of 9 values	2.21	2.13	1.98	1.32	1.76	1.81	1.82	1.91	1.96	2.07	2.19	2.32	1.993
St. dev. of 10	0.175	0.188	0.115	0.119	0.087	0.053	0.131	0.115	0.110	0.116	0.089	0.156	0.050
St. dev. of 9	0.079	0.149	0.085	0.092	0.064	0.044	0.100	0.085	0.072	0.098	0.071	0.135	0.042

Christmas') and September and October 1969 also had tangible explanations for their unreasonably high death rates—clement weather, with unexpected fogs on motorways!

### (c) *Designed experiments, multivariate analyses*

Here, also, we may expect to encounter outliers, with interpretations both deterministic (arising from tangible, non-statistical, sources) and probabilistic (causing us to question distributional or structural assumptions). A taxonomist might well be confronted with a problem of classification of an individual on which he has taken a vector of measurements, no single one of which is 'surprising' in relation to its own marginal distribution yet the assemblage of which as a multivariate observation is in some sense 'surprisingly far away from the main group'. Limited study has been made of ways

of handling outliers in these situations, but some methods have been proposed and will be discussed later. See Chapters 6 and 7 on multivariate samples and designed experiments, respectively.

## 1.4 BAYESIAN AND NON-PARAMETRIC METHODS

During recent years there has been much interest in, and application of, *Bayesian* and *decision-theoretic* methods of statistical analysis. These differ from the more traditional methods both in terms of their interpretation of basic concepts and aims, and in their use of extra forms of relevant information (prior probabilities and consequential costs, additional to sample data). The sample data components of the information used in such approaches may well contain outlying observations. Again we must know how to deal with them. The more modern methods of statistical analysis affect the outlier problem in two ways. On the one hand we might ask how a Bayesian, or decision-theoretic, analysis should try to cope with outliers. On the other hand we might ask if such forms of analysis can be used for the detection and processing of outliers. There are some conceptual difficulties here. A basic tenet of the Bayesian approach is that inferences, or decisions, are strictly *conditional* on the actual sample data that have been obtained. In the main, it is contrary to the Bayesian tradition to view the data within a framework of alternative sets of data *which might have arisen*. Thus the probability mechanism for generating the data is seen to be irrelevant, and it might appear to be similarly irrelevant to question the integrity of the realized data—in particular, to examine the implications of outlying observations. However, Bayesian methods revolve around the likelihood function which needs to be specified in any particular case. The likelihood does depend on an assumed probability model, and in turn it would seem that, relative to that model, certain observations might be outliers. Can the Bayesian approach afford, on principle, to ignore what these outliers might imply about incorrect specification of the likelihood?

One aspect of the Bayesian approach is its formalization of subjective impressions as an ingredient of statistical analysis. We have remarked above how subjective factors arise in judging whether or not outliers exist in a set of data. In the example on weights of birds in Section 1.1, for instance, the reading 1.55 looked suspicious and could well represent a recording error for 2.05. In the time-series example on quarterly sales in Section 1.3, the pattern of sales over the last year appeared inconsistent with earlier experience and made us question the values A and B. In both these cases subjective judgement was involved. Might it not be better to use Bayesian methods directly in trying to reach a conclusion about the outliers, or in taking them into account in processing the data? For example, we might attempt to assign prior probabilities to different possible explanations of the

outliers. Some efforts in this direction have been made and will be considered in Chapter 8.

Another avenue for statistical analysis is in the use of *non-parametric methods*. We shall need to consider whether the notion of an outlier makes any sense in the context of standard non-parametric inference procedures. At first sight the relative nature of outliers (on page 5 we defined outliers in relation to a provisional probability model) seems to deny such a prospect. However, in non-parametric tests of location or dispersion for two samples we are essentially asking if one sample is an outlier relative to the other. Here, of course, observations are ‘labelled’ as belonging to one or the other sample. Such labelling is crucial: it enables us still to operate within a ‘relative’ framework and to extend the outlier concept to the non-parametric area. The discussion of slippage tests in Chapter 5 is part of such an extended view. It also seems sensible to enquire whether we could use a non-parametric approach for the analysis of outliers where a probability model *has* been prescribed.

## 1.5 SURVEY OF OUTLIER PROBLEMS

The informal discussion of ideas and examples in the earlier parts of this chapter enables us to draw up a broad classification of the types of enquiry we need to make in the study of outliers. No single factor classification will do since we must consider the distinctions

- (i) between deterministic and statistical causes of outliers,
- (ii) between univariate and multivariate data sets,
- (iii) between different specific probability models,
- (iv) between different forms of statistical analysis in which the outliers are encountered,
- (v) between single or multiple outliers (including outlying samples and sets of samples), and
- (vi) most fundamentally, between the different aims and purposes we may have in studying outliers.

Such a complex array of considerations does not lead to a particularly tidy subdivision of topics, but the arrangement of the succeeding chapters and their subsections has been chosen in recognition of such distinctions in what seems to be a fairly natural progression from the simpler to the more complex considerations. Our main object has been to present a fairly full review of existing methodology in the treatment of outliers. We have aimed throughout both to provide sufficient theoretical detail to meet the interests of the mathematical statistician and to give a full enough description of practical methods to meet the needs of the data analyst. Application of the methods is of paramount importance and we have tried to present relevant illustration (within the limitations of the scale of the book).

We shall start with a general discussion of the different aims and purposes in studying outliers and proceed to what is perhaps the most elementary aspect of this, the examination of outliers in single univariate samples from a given probability distribution. Special attention is given to the effect of outliers in estimation procedures. Outlying sub-samples are then considered; also multivariate data. We progress to more structured models such as regression, designed experiments, and time-series. After some discussion of Bayesian and non-parametric procedures, we conclude with comment on the future state of the art.

## CHAPTER 2

# *What should one do about outlying observations?*

### 2.1 EARLY INFORMAL APPROACHES

The existence of the problem of doubtful or anomalous values has been recognized for a very long time, certainly since the middle of the eighteenth century. Daniel Bernoulli, writing in 1777 about the combination of astronomical observations, said:

is it right to hold that the several observations are of the same weight or moment, or equally prone to any and every error? . . . Is there everywhere the same probability? Such an assertion would be quite absurd, which is undoubtedly the reason why astronomers prefer to reject completely observations which they judge to be too wide of the truth, while retaining the rest and, indeed, assigning to them the same reliability. . . . I see no way of drawing a dividing line between those that are to be utterly rejected and those that are to be wholly retained; it may even happen that the rejected observation is the one that would have supplied the best correction to the others. Nevertheless, I do not condemn in every case the principle of rejecting one or other of the observations, indeed I approve it, whenever in the course of observation an accident occurs which in itself raises an immediate scruple in the mind of the observer, before he has considered the event and compared it with the other observations. If there is no such reason for dissatisfaction I think each and every observation should be admitted whatever its quality, as long as the observer is conscious that he has taken every care. (Allen, 1961)

To take an even earlier example, Boscovich, attempting in 1755 to determine the ellipticity of the earth by averaging ten measurements of excess of the polar degree over the equatorial, decided to discard the two extreme values of excess as outliers and recomputed the mean from the reduced sample of eight, see Maire; Boscovich (1755). From this period until the middle of the nineteenth century, the main point of discussion in the literature with regard to outlying values is whether rejection is justified. Some writers took the same view as Daniel Bernoulli, that observations should not be rejected purely on grounds of appearing inconsistent with the remaining data; Bessel and Baeuer, for example, wrote in 1838 that

they had never rejected an observation merely because of its large residual, and that all completed observations, with equal weight, ought to be allowed to contribute to the result. Others, such as Boscovich, practised rejection. However, rejection was never envisaged in these early days as being carried out according to any formal procedure, but was purely a matter of the observer's judgment. Legendre, for example, in 1805 was recommending the rejection of deviations 'adjudged too large to be admissible'. Indeed a century later Saunder could write (1903):

I believe that the practice amongst computers of experience is to rely almost entirely on their individual judgment, taking into account the conditions of the observations, and drawing the line somewhere about those observations which give residuals of five times the probable error.

The first published objective test for anomalous observations was due to the American astronomer Peirce (1852). In Peirce's procedure,  $k$  doubtful observations in a sample of  $n$  should be rejected if

the probability of the system of errors obtained by retaining them is less than that of the system of errors obtained by their rejection multiplied by the probability of making so many, and no more, abnormal observations.

This last probability Peirce took to be  $p^k(1-p)^{n-k}$ , where he began by defining  $p$  as 'the probability, supposed to be unknown, of such an abnormal observation that it is rejected upon account of its magnitude', and then assigned it the value  $k/n$ .

This bizarre test was followed in 1863 by the publication of the test for a single doubtful observation (already referred to in Chapter 1) by the American astronomer, Chauvenet. Chauvenet's test has an attractive simplicity lacking in Peirce's; despite its now evident shortcomings, it persists in print to the present day, at any rate in a number of textbooks for students in engineering and the experimental sciences. The reasoning was as follows (see Stone, 1868). If  $\theta(x)$  denotes the probability that an error is equal to or greater than  $x$ ,

then the number of errors equal to or greater than  $x$  which may fairly be expected in  $n$  observations is  $n\theta(x)$ . If therefore we find  $x$  such that  $n\theta(x)=\frac{1}{2}$ , any error greater than  $x$  will have a greater probability against it than for it, and may, therefore, be rejected.

In effect, an observation is to be rejected if it lies outside the lower and upper  $1/(4n)$  points of the null distribution. Evidently with this procedure the chance of wrongly rejecting a non-discordant value is in a large sample approximately  $1 - e^{-\frac{1}{2}}$ , i.e. about 40 per cent!

Soon after Chauvenet, Stone (1868) introduced a rejection test based on the concept of a *modulus of carelessness*,  $m$ . This concept can be expressed in the following way: a given observer in a given sampling situation makes on average one mistake in every  $m$  observations he takes. An observation is to be discarded if its deviation can be attributed with more probability to the

observer's carelessness than to random variation. This means, in effect, that an observation is rejected if it lies outside the lower and upper  $1/(2m)$  points of the null distribution, so the test is essentially similar to Chauvenet's, becoming identical with it if  $m = 2n$ .

Alternatives to outright rejection of extreme values were also being considered by a number of writers. Within a few years of Stone's rejection test, several methods were published—one by Stone himself—for the weighting of observations in calculating a sample mean. This can be regarded as a robust procedure for estimating a location parameter which secures the *accommodation* of outliers. In Rider's words (1933),

Since the object of combining observations is to obtain the best possible estimate of the true value of a magnitude, the principle underlying ... [weighting] methods is that an observation which differs widely from the rest should be retained, but assigned a smaller weight than the others in computing a weighted average. Of course retention with an exceedingly small weight amounts to virtual rejection.

Glaisher (1872–73) was perhaps the first to publish a weighting procedure, remarking, 'It will be seen that it supersedes the necessity for the *rejection* of anomalous observations'.

Glaisher's method was to assume that the  $n$  observations  $x_i$  ( $i = 1, \dots, n$ ) were normally distributed, with a common mean  $\mu$  required to be estimated, and with unknown and unequal variances  $\sigma_i^2$ . For a provisional estimate of the mean he allotted what he regarded as plausible values for the  $\sigma_i^2$ . These in turn led to a modified estimate of the mean, new estimates of the  $\sigma_i^2$ , and so on. Specifically, successive weighted means  $m_{(1)}, m_{(2)}, \dots$  were calculated,  $m_{(1)}$  being the ordinary sample mean  $\bar{x}$ , and the weights  $W_r$  for the  $r$ th weighted mean  $m_{(r)} = \sum_i W_{ri}x_i$  being defined recursively by

$$W_r \propto \exp\{-2(W_{r-1,1} + \dots + W_{r-1,n})(x_i - m_{(r-1)})^2\}. \quad (2.1.1)$$

Stone (1873) followed a few months later with a criticism of Glaisher's method and a proposal for an alternative weighting procedure, based in effect on maximizing the likelihood, which is proportional to

$$\prod_i [(1/\sigma_i) \exp -\{(x_i - \mu)^2 / 2\sigma_i^2\}], \quad (2.1.2)$$

with respect to  $\mu$  and all the  $\sigma_i$ . This leads to a weighted mean  $\tilde{\mu}$  given by the  $(n-1)$ th degree equation

$$\sum_{i=1}^n (x_i - \tilde{\mu})^{-1} = 0. \quad (2.1.3)$$

The same method was published independently by Edgeworth (1883) ten years later, though Edgeworth subsequently (1887) acknowledged Stone's priority.

Another weighting method proposed at this time was by Newcomb (1886); see Stigler (1973b). His procedure assumes the  $n$  observations to have come from a mixture of  $r$  normal distributions, and evolves a final estimate of  $\mu$  which is constructed as a weighted mean of  $r^n$  different weighted means of the  $x_i$ . Rather interestingly, Newcomb refers in his paper to the 'evil' of a value; this turns out to be the mean squared error of an estimate, and is an interesting early use of the concept of a loss function.

Apart from these and other similar references of the period to the accommodation of extreme values by weighting, it is interesting to find what we would now call *trimming* discussed in 1895 by Mendeleev, the discoverer of the periodic table of elements. Referring to the evaluation of the length of the Russian standard platinum-iridium metre from a set of eleven determinations, Mendeleev wrote:

I use . . . [the following] method to evaluate the harmony of a series of observations that must give identical numbers, namely I divide all the numbers into three, if possible equal, groups (if the number of observations is not divisible by three, the greatest number is left in the middle group): those of greatest magnitude, those of medium magnitude, and those of smallest magnitude: the mean of the middle group is considered the most probable . . . and if the mean of the remaining groups is close to it . . . the observations are considered harmonious. (See Harter, 1974–1976, Part I)

Returning to the history of ad hoc rejection tests such as Chauvenet's and Stone's, the literature of the following fifty years until the period of the First World War affords a number of further examples of such tests. In particular, we may note Wright's procedure (1884), which rejects any observation deviating from the mean by more than three times the standard deviation, or equivalently five times the probable error; the modified version by Wright and Hayford (1906), which adds to Wright's rule the further instruction:

Examine carefully each observation for which the residual exceeds 3.5 times the probable error, and reject it if any of the accompanying conditions are such as to produce lack of confidence

and Goodwin's procedure (1913), which rejects an outlying observation in a sample of  $n$  if its deviation from the mean of the remaining  $n - 1$  exceeds four times the average deviation of the  $n - 1$ .

One notes that Wright and Goodwin chose, for the critical ratios in their tests, values 5, 3.5, 4, and so on, which were independent of the sample size  $n$ . This relates to one general defect of all the test procedures proposed up to this time—they failed to distinguish between population variance and sample variance.

Perhaps the first writer to make this point explicitly with regard to outlier procedures was Irwin (1925), who pointed out the implications for outlier rejection of the unreliability of the sample standard deviation,  $s$ , as an

estimate of its population analogue,  $\sigma$ . For the case where  $\sigma$  is known, he proposed the test statistics

$$[x_{(n)} - x_{(n-1)}]/\sigma \quad \text{and} \quad [x_{(n-1)} - x_{(n-2)}]/\sigma$$

where  $x_{(i)}$  denotes the  $i$ th ordered sample value. Ten years later an exact test based on a studentized criterion,  $(x - \bar{x})/s$ , was published by Thompson (1935). This was shortly followed by the classic paper by Pearson and Chandra Sekar (1936) entitled 'The efficiency of statistical tools and a criterion for the rejection of outlying observations'. A rationale for the treatment of outliers was beginning to take shape.

An encyclopaedic survey of outlier methods from the earliest times is included in Harter (1974–1976).

## 2.2 VARIOUS AIMS

In the previous section we reviewed some of the early informal approaches which had been used for the study of outliers. In such work there was seldom any overt consideration of the purpose of examining the outlying observations, of the manner in which outliers may reflect contamination of a basic probability model, or of any optimality properties possessed by the prevailing statistical methods. Any detailed examination of the current state of theory and practice in relation to the study of outliers must consider such matters of aim, model, and principle. The remaining sections of this chapter present a general review of

- (i) the different aims in examining outliers,
- (ii) probabilistic models explaining outliers, and
- (iii) the philosophy and form of relevant statistical procedures.

Such matters are reconsidered in appropriate detail in later chapters where fuller treatments of specific topics are given.

We commence with what must be the most fundamental question: Why are we concerned about outliers? The practical examples of Chapter 1 have illustrated some grounds for concern. Let us examine these more systematically.

Let us recall what we mean by an 'outlier'. Note that, as explained below, the term 'outlier' is used by different authors in two different senses. As employed in this text, it is a subjective post-data manifestation. In observing a set of observations in some practical situation one (or more) of the observations 'jars', stands out in contrast to other observations, usually as an extreme value. As Grubbs (1969) remarks (italics inserted):

An outlying observation, or 'outlier', is one that *appears to deviate markedly* from other members of the sample in which it occurs.

Such outliers do not fit with the tidy pattern present in our mind, at the outset of our enquiry, of what constitutes a reasonable set of data. We have

subjective doubts about the propriety of the outlying values both in relation to the specific data set we have obtained and in relation to our initial views of an appropriate probability model to describe the generation of our data. Note how our feelings about the data may, in this respect, differ quite widely with different possible basic probability models. If we anticipate a normal distribution we may react quite strongly to certain observations which would arouse no specific concern if the expected model is longer-tailed, say log-normal or Cauchy. The purpose of a body of statistical method for examining outliers is, in broad terms, to provide a means of assessing whether our *subjective* declaration of the presence of outliers in a particular set of data has important *objective* implications for the further analysis of the data.

Outliers may have arisen for purely deterministic reasons: a reading, recording, or calculating error in the data. When it is obvious that this is so the remedy is clear: the offending sample values should be removed from the sample or replaced by corrected values when the method of 'correction' is unambiguously understood. See the examples in Section 1.1. In less clear-cut circumstances where we suspect, but cannot guarantee, such a tangible explanation for outliers, appropriate statistical procedures may be used to assess *discordancy*. In this text, an observation will be termed *discordant* if it is *statistically unreasonable* on the basis of some prescribed probability model. We shall later extend this definition to include an observation known to have been generated by a different probability model; such a *discordant observation* need not necessarily show up as an *outlier*.

Some writers use the word 'outlier' for an observation which is both surprising and discordant; a term such as 'suspect value' is then used by them to describe a surprising value (an outlier in our sense). To quote a typical example we read (Grubbs, 1950):

Then again, both the largest and the smallest observations may appear to be 'different' from the remaining items in the sample. Here we are interested in testing the hypothesis that both the largest and the smallest observations are truly 'outliers'.

Of course extreme values must always occur in a set of data. What is important is whether or not they are so extreme that they could not reasonably have arisen by chance from the adopted model. If so, we may feel that we now have substantiating evidence for some earlier conjecture of a 'mistake' in the data, and again would wish to reject (or correct) that mistake. Alternatively, discordancy of unreasonably extreme outliers may promote the adoption of a new probability model, with important implications for further analysis of the data.

What form does the alternative model take? There are various possibilities. It may be merely a differently shaped distribution in relation to which the complete set of data (including the outliers) appears as a homogeneous random sample, or it may need to be more structured. For example, a random mixture of two distributions may reasonably account for

the data form and, indeed, there may seem to be no alternative but to assume that the outliers reflect 'foreign' random (rather than deterministic) influences in an otherwise homogeneous set of data. Such foreign influences may, at one extreme, be matters of great interest in their own right. At the opposite extreme they may act only as obstructions in our efforts to assess the properties of the main mass of data. Some examples of these distinctions are given in Section 1.2. Statistical methods aimed at assessing discordancy have played a large part in the literature on outliers. We shall be considering a variety of *tests for discordancy* for different situations in Chapter 3, and later in this chapter we shall examine general statistical principles on which they are based.

Inevitably, a test for discordancy for outliers plays only an initial role in the analysis of the data. Leaving aside the wider statistical analysis we intend to apply to the data (and for which purpose they were presumably assembled) an assessment of discordancy of some outliers must be viewed only as a first stage of study of the outliers themselves. What action are we to take if we adjudge one or more outliers to be discordant? This will depend on a variety of factors relating to our interest in the practical situation. Obvious possibilities arise: we may decide to *reject* (or correct) the discordant outliers and proceed to analyse the residual (modified) data on the original model, we may choose to modify the model to *incorporate* the outliers in a non-discordant fashion, we may refine the way in which we analyse the whole data set to *accommodate* the outliers (render the analysis relatively impervious to their presence), or we may concentrate attention on the discordant outliers as a welcome *identification* of unsuspected factors of practical importance.

Let us consider some numerical examples, where, at first sight, these separate possibilities seem reasonable. Suppose that in each case any declared outliers prove to be *discordant* on the basis of an appropriate statistical test employing an 'initially reasonable' model.

Chauvenet (1863) declared the observations 1.01 and -1.40 as, respectively, upper and lower outliers in the following set of 15 residuals (about a simple model) of observations of the vertical semi-diameter of Venus, in seconds, made by Lt. Herndon in 1846.

-0.30	-0.24	<b>-1.40</b>	+0.18
-0.44	+0.06	-0.22	+0.39
<b>+1.01</b>	+0.63	-0.05	+0.10
+0.48	-0.13	+0.20	

If the outliers prove to be discordant on an assumed normal distribution it is quite likely (bearing in mind the possibilities of inexplicable 'gross errors') that we may choose to *reject* them before proceeding to further study of the data. We cannot, of course, be sure that this action is entirely proper. Perhaps an appropriately more sophisticated non-normal model

would *incorporate* them in a non-discordant fashion, or the purpose of the further analysis may support some form of partial *accommodation* short of complete incorporation (see comments below). Should we decide to reject the outliers then the stricture of Kruskal (1960b) makes sound sense.

As to practice, I suggest that it is of great importance to preach the doctrine that apparent outliers should *always* be reported, even when one feels that their causes are known or when one rejects them for whatever good rule or reason. The immediate pressures of practical statistical analysis are almost uniformly in the direction of suppressing announcements of observations that do not fit the pattern; we must maintain a strong sea-wall against these pressures.

The outright rejection of outliers has statistical consequences for the further analysis of the reduced sample. We may no longer have a random sample, but a censored one. The practice of replacing rejected (non-deterministically inexplicable) outliers by statistical equivalents (further simulated random observations from the assumed underlying distribution) involves similar consequences. The practices of 'Winsorization' and trimming (see below) will also have distributional implications which must be allowed for.

The following data described by Karl Pearson (1931) present the capacities (in cc) of a sample of seventeen male Moriori skulls.

1230	1318	1380	1420	<b>1630</b>	1378
1348	1380	1470	1445	1360	1410
1540	1260	1364	1410	1545	

The observation 1630 was suspected as being 'too large'; suppose it proves to be discordant. It may here be more appropriate to seek an alternative model which *incorporates* the value 1630 in a non-discordant way. Biological data often require skew distributions as models—see the example on *macro-lepidoptera* in Section 1.2. There is of course the possibility that *identification* of the outlier reflects the presence of a small number of another species in the population being studied, or (possibly less realistically) that it has arisen from a once-and-for-all error of measurement, or of recording.

Daniel (1959) reports the results of a  $2^5$  factorial experiment where the 31 contrasts arranged in order of increasing absolute value are:

0.0000	0.0281	-0.0561	-0.0842	-0.0982	0.1263	0.1684
0.1964	0.2245	-0.2526	0.2947	-0.3087	0.3929	0.4069
0.4209	0.4350	0.4630	-0.4771	0.5472	0.6595	0.7437
-0.7437	-0.7577	-0.8138	-0.8138	-0.8980	1.080	-1.305
2.147	-2.666	-3.143				

The last three observations are discordant outliers on the normal model. But this is precisely what we are seeking: important effects of the experimental factors. Thus we *identify* the outliers as indications of features of

practical importance rather than as tedious reflections of possible inadequacies in the model or measurement technique.

There is one area of enquiry where our study of outliers may *not* necessarily commence with a test of discordancy. Consider again the Herndon data on residuals of observations of the vertical semi-diameter of the planet Venus. Suppose that we wish to estimate some summary measure of the distribution of residuals, perhaps the mean or variance. The properties of different estimators will vary with the *form* of the distribution. Not knowing its form we want to use an estimator which is reasonably *robust* against different possible distributional forms. This concern for robustness in estimation (or testing) includes an interest in procedures which protect against the possibility (or presence) of outliers. The extreme prospect is that we decide to reject the outliers prior to estimation or testing, either because of clear tangible explanations of their presence or following a test of discordancy based on a confidently assumed model. But interest in robustness against outliers denies any great confidence in the appropriate model and preliminary tests of discordancy may not be feasible. Furthermore the severe act of *rejecting* outliers may be over-extravagant and over-specific. There are possibilities of partial rejection, or indeed of impartial (sic) rejection. Although we may be suspicious of the actual values -1.40 and 1.01, we may nonetheless feel that the *direction* of the residuals carries information and wish to retain this information in some form. One possibility is to employ *Winsorization* where, for example, we replace the lower and upper extremes by their nearest neighbours. For the Herndon data, -1.40 and 1.01 are replaced by -0.44 and 0.63, respectively, thus making each of these latter values appear twice in the data. Alternatively as an aid to robustness of estimation or testing we may choose to use an  $\alpha$ -*trimmed* sample, in which a fixed fraction  $\alpha$  of lower, and upper, extreme sample values are totally discarded before processing the sample. This 'old French custom' (Huber, 1972) is not specifically concerned with protecting against outliers, though these will clearly be candidates for trimming.

From the robustness standpoint, we are thus aiming to devise statistical procedures which do not directly examine the outliers, but seek to *accommodate* them and render them less serious in their influence on estimation or tests of summary measures of the underlying distribution.

Several authors of review papers on the topic of outliers (for example, Anscombe, 1960a; Grubbs, 1969) have attempted to categorize the different ways in which outliers may arise. Such ideas have been touched on in Chapter 1. It is relevant to consider them in rather more detail. In taking observations, different sources of variability can be encountered. We can distinguish three of these.

*Inherent variability.* This is the expression of the way in which observations intrinsically vary over the population; such variation is a natural feature of the population and uncontrollable. Thus, for example, measurements of

heights of men will reflect the amount of variability indigenous to that population.

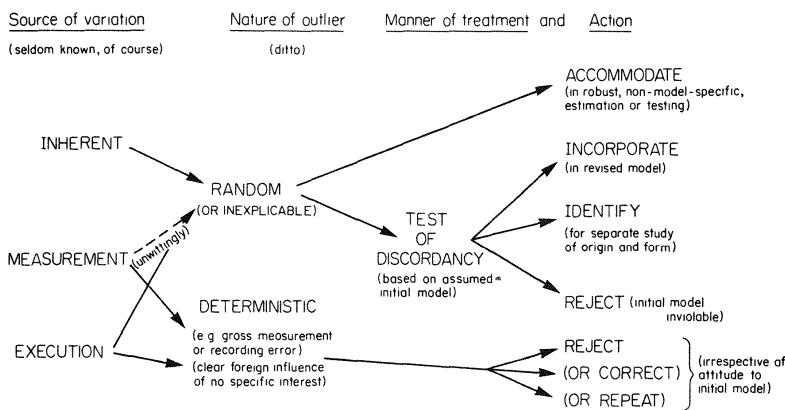
**Measurement error.** Often we must take measurements on members of a population under study. Inadequacies in the measuring instrument superimpose a further degree of variability on the inherent factor. The rounding of obtained values, or mistakes in recording, compound the measurement error: they are part of it. Some control of this type of variability is possible.

**Execution error.** A further source of variability arises in the imperfect collection of our data. We may inadvertently choose a biased sample or include individuals who are not truly representative of the population we aimed to sample. Again, sensible precautions may reduce such variability.

We can usefully attempt to classify outliers in relation to these three types of variability. An outlier in a set of data may in fact be a perfectly reasonable reflection of the natural inherent variation. If shown statistically to be discordant this reflects an inadequate *basic model* (unless of course it is merely a manifestation of Type I error). We would hope to learn from the experience of the discordant outlier and adopt a more appropriate model. But as Anscombe (1960a) points out:

In no field of observation can we entirely rule out the possibility that an observation is vitiated by a large measurement or execution error. . . . there must be a suspicion that the deviation is caused by a blunder or gross error of some kind. Several possible reasons . . . can usually be thought of without difficulty. In such cases, the reading will be checked or repeated if that is possible. If not, it may be rejected as spurious because of its big residual, even though there is no known reason for suspecting it. In sufficiently extreme cases, no one hesitates about such rejections. . . . If we could be sure that an outlier was caused by a large measurement or execution error which could not be rectified (and if we had no interest in studying such errors for their own sake), we should be justified in entirely discarding the observation and all memory of it. The act of observation would have failed; there would be nothing to report.

Certain points need to be underlined here. An obvious unrepresentative *measurement* error supports rejection of the offending observation (or occasionally we may be able to correct it, or repeat it). An outlier in the form of an excessive *execution* error may sometimes also lead to rejection, but it could on occasions warrant a modified model (perhaps of a mixture type to cope with infrequent foreign sample members—discordant values—reflected by the outlier), or it could serve to identify some factor of importance in its own right (as, for example, in an analysis of variance situation). Anscombe (1960a) distinguishes in terminology between outliers arising from large variation of the *inherent* type, and those from large *measurement* or *execution* error. He calls the former ‘outliers’, the latter ‘spurious observations’. We shall make no such distinction. The full study of statistical methods for outliers needs to encompass all derivative sources of variation; the only exceptions are outliers arising from clearly discernible deterministic mistakes of calculation, recording, etc. In this case rejection

**Figure 2.1** Treatment of outliers

(or correction) is the only remedy; otherwise we need to clearly recognize the many possibilities, other than outright rejection, for coping with outliers.

Reviewing this section we can present schematically the different interests and aims in the handling of outliers. Figure 2.1 presents a simplified diagrammatic summary.

### 2.3 MODELS FOR DISCORDANCY

In the spirit of the previous section we may regard the test of discordancy as central to much of our interest in outliers. Any statistical test must inevitably examine two hypotheses: a null hypothesis, or *working hypothesis*, which will be conserved unless significant evidence is found to support its rejection, and an *alternative hypothesis* in favour of which the working hypothesis may need to be rejected. For a test of discordancy of outliers, the working hypothesis will express some basic probability model for the generation of all the data with no contemplation of outliers; the alternative hypothesis expresses a way in which the model may be modified to incorporate or explain the outliers. We shall consider in some detail different forms of outlier-generating model which have been proposed and studied. In the following section we extend the discussion of tests of discordancy to consider the statistical nature of the tests themselves and any optimality properties they may possess.

Much of the early, or existing, work on tests of discordancy is highly intuitive in form and has little regard for the nature of the working and alternative hypotheses. Thus, for example, when concerned with a single upper outlier in a set of independent observations  $x_1, x_2, \dots, x_n$  it is natural,

and appealing, to contemplate the statistics

$$[x_{(n)} - x_{(n-1)}]/D$$

or

$$[x_{(n)} - \bar{x}']/D$$

where  $x_{(i)}$  is the  $i$ th *ordered* value in the sample (when all observations are placed in ascending order of magnitude),  $\bar{x}'$  is the sample mean excluding the outlier  $x_{(n)}$  and  $D$  is some measure of the spread of the sample (again excluding consideration of  $x_{(n)}$ ). See Chapter 3, page 53. Such statistics seem intuitively reasonable for assessing the discordancy of the outlier  $x_{(n)}$ , and were originally proposed purely on such an informal basis. Only at the more formal stage of determining the level or size of the test, or of constructing tables for assessing the statistical significance of the test statistic value, does it become unavoidable that the working hypothesis is specified. To go further and examine power characteristics of the test, or to construct tests with certain desirable statistical properties, requires the additional specification of the alternative hypothesis.

The historical development of tests of outliers mirrors such a progression in statistical sophistication. Whilst the declaration of the alternative hypothesis is the crux of the problem of defining just what we mean by outliers, it still has not been very widely discussed. Perhaps this is inevitable: outliers are not easily defined, or incorporated in a generally acceptable form of model, and a degree of controversy still surrounds their study (see the opening pages of Chapter 1).

### *The working hypothesis*

No fundamental conceptual difficulty exists in setting up the working hypothesis. We have repeatedly stressed the conditional nature of the outlier concept. An outlier is an observation which appears suspicious in the light of some provisional initial assignment of a probability model to explain the data generating process. Thus the working hypothesis is merely a statement of the *initial (basic) probability model*. In some discussions of the nature of tests of discordancy, or in attempts to accommodate outliers through robust statistical procedures, we may not need to be too specific about the probability model. It may suffice to declare, as a *working hypothesis*,  $H$ , that the data arise as independent observations from some common, but unspecified, distribution  $F$ . We can denote this as:

$$H : F.$$

But in the detailed analysis of real-world data we will often have a much more specific model in mind: for example, that the data constitute a random sample from an *exponential distribution* with scale parameter  $\theta$  (which may, or may not, be specified in value) or from a *homoscedastic linear regression*

situation with *normal error structure*. Inevitably, much of the discussion of statistical tests of discordancy for outliers (as in so many areas of statistical enquiry) assumes a *normal* working hypothesis.

Chapter 3 will be concerned with the detailed forms of discordancy tests for individual outliers, or small groups of lower and upper outliers. Test statistics, relevant tables of percentage points and special advantages or disadvantages are described for a variety of tests. Most results are available for the normal and gamma (including exponential) distributions and these are presented along with what relatively little information exists for other distributions such as the Pareto, Poisson, log-normal and uniform. More effort could usefully go into devising and examining the detailed properties of tests of discordancy for non-normal distributions.

If, on the basis of a test of discordancy, we adjudge an outlier (or small group of outliers) to be discordant we implicitly reject the working hypothesis in favour of some *alternative hypothesis*. Clearly, we must know what alternative hypothesis is being adopted! Any assessment of the power of the prevailing test of discordancy, indeed its very justification in statistical terms, depends on specifying the alternative hypothesis. All too frequently this aspect is ignored in the discussion of tests of outliers. Let us consider some of the possible forms of alternative hypothesis for tests of discordancy and examine (briefly at this stage) the extent to which they figure in the construction or application of tests prescribed in the literature.

### *Forms of alternative hypothesis*

We can readily contemplate a variety of different forms of alternative hypothesis for outlier tests of discordancy. Some of these have been discussed in the literature; others have obvious appeal but do not appear to have been considered in any detail. The formulation of the alternative hypothesis, in relation to the subjective manner in which outliers are declared in a set of data, is no easy matter and much remains to be done in seeking an entirely satisfactory form for the alternative hypothesis.

#### (i) *Deterministic alternative*

The first type of alternative hypothesis which comes to mind covers the cases of outliers caused by obvious identifiable gross errors of measurement, recording, and so on. Such an alternative hypothesis, which we term *deterministic*, is entirely specific to the actual data set, and the observed offending observations. Thus if our data  $x_1, x_2, \dots, x_n$  contain one observation, say  $x_i$ , which has clearly arisen from a mistaken reading or recording, we immediately reject any basic model,  $F$ , for the *whole* data set in favour of an alternative model which says that all  $x_j$  ( $j \neq i$ ) arise at random from  $F$  whilst  $x_i$  is quite different and requires rejection (or correction, or repeat reading). No test of discordancy is needed. Rejection of the initial model in

favour of the alternative is *deterministically* correct (even if difficult to confirm from the practical point of view).

### (ii) Inherent alternative

In the terms of the discussion of sources of variability in the previous section, we must entertain the possibility that outliers have appeared in the data as a result of a greater degree of inherent variability than we initially anticipated. Perhaps what we thought was a sample from a normal distribution was really from a 'fatter-tailed' distribution. Upper outliers may reflect, say, that an initial assumption of a gamma distribution is best replaced with a log-normal distribution. Thus where outliers reflect a larger measure (or different form) of inherent variability than is encompassed in some basic model,  $F$ , we may choose to express this by opposing the working hypothesis,

$$H:F,$$

that *all* observations arise from the distribution,  $F$ , with a suitably chosen alternative hypothesis,

$$\bar{H}:G,$$

that *all* observations arise from a distribution,  $G$ , under which the outliers no longer occasion the earlier degree of 'surprise'.  $F$  and  $G$  may be different fully specified distributions, or may be distinct general parametric families of distributions.

Of course, we would hope to be able to distinguish between  $F$  and  $G$  by appropriate statistical procedures (based on the complete sample) more powerful than the outlier-specific test of discordancy. On the other hand, in small sets of data our very motivation for considering a radically different form of model may stem entirely from the presence of outliers. Thus the *inherent* alternative hypothesis is relevant to tests of discordancy. Shapiro and Wilk have given such a test directed particularly to inherent alternatives, both for normal samples (1965) and for exponential samples (1972); detailed results on the power of their normal test against 45 different inherent alternatives are given by Shapiro, Wilk, and Chen (1968), and on the power of their exponential test against 15 different inherent alternatives by Shapiro and Wilk (1972).

### (iii) Mixture alternative

Rather than assume that outliers may reflect an unexpected degree or form of inherent variability, we might admit the possibility of 'errors of execution' allowing 'contamination' of the sample by a few members of a population other than that represented by the basic model. We assume that such 'foreign' sample members, or discordant values, show themselves as outliers. For example, in examining a sample of fossils from what is

supposed to be a homogeneous population of the same species, we might inadvertently collect one or two fossils of a different species with different size characteristics. Whilst the presence of the different species might reasonably be ascribed to execution error when we are only interested in the predominant species, the term 'error' is not in general a good label for this type of manifestation. A particle physicist measuring characteristics of paths of radioactive particles is more likely to term it 'good fortune' than 'error' if he discovers in the form of outlying observations a basically new type of particle. (This illustrates again the distinction between *rejection* and *identification* in our response to a test of discordancy.)

In these terms a sensible alternative to the working hypothesis  $H:F$  is an hypothesis of the form

$$\bar{H}:(1-\lambda)F + \lambda G$$

which declares that outliers reflect the (small) chance  $\lambda$  that observations arise from a distribution  $G$ , quite different from the initial model  $F$ . Such an alternative hypothesis will be called a *mixture* alternative, and it figures in some published work on outliers (see below).

As with the *inherent* type of alternative, we should again hope that the dichotomy

$$H:F \text{ versus } \bar{H}:(1-\lambda)F + \lambda G$$

is promoted, supported, and best analysed on a broader basis than the degree of 'surprise' engendered by the presence of one or two outliers in the data. But if the sample size, and the mixing parameter  $\lambda$ , are small, it may be that the outliers alone focus our attention on the possibility of a mixture model. Box and Tiao (1968) and Guttman (1973b) discuss the implications of a mixture alternative for the study of outliers. Box and Tiao remark that if the mixture prospect is revealed through outliers alone it will be necessary to assume that in a sample of at most 20 observations  $\lambda$  is small: possibly less than 0.05, or at very most 0.10. Otherwise, occasions will arise where we encounter more observations from the distribution  $G$  than we should be content to designate 'outliers'. (For larger samples,  $\lambda$  will need to be even smaller on this argument.) Guttman inverts this argument in support of the commonly made assumption that a set of data contains at most *one* discordant outlier! However, there seems to be a certain circularity in such arguments. If  $\lambda$  is very small and  $\bar{H}$  is a reasonable model we are unlikely to encounter more than one or two members of  $G$  which (for appropriate  $G$ ) may show up as outliers; if we encounter just one (or two) outliers it may be that  $\bar{H}$  is an appropriate model to explain the outliers but  $\lambda$  must performe be small. But how are we to adjudge the propriety of the mixture alternative  $(1-\lambda)F + \lambda G$  with no evidence other than the one (or two) outliers whose

discordancy we are to assess using this alternative model  $(1 - \lambda)F + \lambda G$ ? There is no easy answer to this—but it seems equally difficult to justify any other of the proposed forms of alternative hypothesis, although to judge from the literature there seems to be greater intuitive appeal in the *slippage* type of alternative hypothesis we shall discuss next.

At this point (though not specifically concerned with solely the mixture alternative) it is relevant to point out an apparent philosophical inconsistency in the formulation of models to explain the presence of outliers. The model  $(1 - \lambda)F + \lambda G$  merely declares that with a certain (small) probability observations might be generated by  $G$ . Yet the data have directed our attention to *specific* observations: the outliers, which appear typically as *extreme values* in the sample. Suppose we have just one (upper) outlier,  $x_{(n)}$ . The alternative hypothesis merely contemplates the possibility of *some* observations arising from  $G$ , not necessarily, specifically or solely,  $x_{(n)}$ . Statistical principles *may* lead to the conclusion that if there is only one observation from  $G$  it is best (in some sense) adjudged to be  $x_{(n)}$ . But this seems to differ from our original subjective interest in the outlier  $x_{(n)}$ , *per se*, which would favour an alternative hypothesis specifically related to  $x_{(n)}$ . There appears to be no discussion of this matter in the literature. In any case, there is no obvious way of expressing such an interest through a mixture type of alternative hypothesis. But even for the slippage type of alternative discussed below, where such a facility can in fact be found, it does not seem to have been contemplated. We discuss this in some detail in Section 3.1.

As an example of the mixture model, Box and Tiao (1968) consider a Bayesian approach to outliers in which, under the working hypothesis, we have a random sample from a normal distribution  $N(\mu, \sigma^2)$ . The alternative hypothesis declares that, independently and with probabilities  $1 - \lambda$ ,  $\lambda$ , respectively ( $\lambda < 0.1$ ), observations arise either from  $N(\mu, \sigma^2)$  or from  $N(\mu, b\sigma^2)$ , with  $b > 1$ . Since  $b > 1$  we might expect observations from the latter distribution ( $G$ ) to appear as extreme values, declared to be outliers. (We consider the results of Box and Tiao in more detail in Section 8.1.2. Their interest is in estimating  $\mu$  rather than testing discordancy of outliers, so it comes under the heading of *accommodation*.)

The mixing distribution  $G$  need not, of course, be restricted to a scale-shifted version of  $F$ ; it could express a change of location or even a radically different form of distribution. Clearly, only certain forms of mixture will be relevant to outliers: shifts of scale with  $b > 1$  may give rise to upper and lower outliers in combination; shifts of location, upper or lower outliers separately depending on the direction of shift. Various mixtures will not be expected to be reflected in the occurrence of outliers.

An early use of a mixture model for outliers is made by Dixon (1953). Tukey (1960) is interested in a mixture model with two normal distributions differing in variance.

(iv) *Slippage alternative*

By far the most common type of alternative hypothesis as a model for the generation of outliers is what we shall refer to as the *slippage* alternative. It has been widely discussed and used, and figures (sometimes only implicitly) in work by Grubbs (1950), Dixon (1950), Anscombe (1960a), Ferguson (1961a), McMillan and David (1971), McMillan (1971), Guttman (1973b), and many others. In its most usual form the slippage alternative states that all observations apart from some small number  $k$  (1 or 2, say) arise independently from the initial model  $F$  indexed by location and scale parameters,  $\mu$  and  $\sigma^2$ , whilst the remaining  $k$  are independent observations from a modified version of  $F$  in which  $\mu$  or  $\sigma^2$  have been shifted in value ( $\mu$  in either direction,  $\sigma^2$  typically increased). In most published work  $F$  is a *normal* distribution. The *models A* and *B* of Ferguson (1961a) are perhaps the most general expression of the *normal* slippage alternative, reflecting shifts of location and dispersion, respectively.

*Model A*  $x_1, x_2, \dots, x_n$  arise independently from normal distributions with common variance,  $\sigma^2$ . (Under  $H$  they have common mean  $\mu$ .) There are known constants  $a_1, a_2, \dots, a_n$  (most of which will be zero), an unknown parameter  $\Delta$  and an unknown permutation  $(\nu_1, \nu_2, \dots, \nu_n)$  of  $(1, 2, \dots, n)$  such that the normal distributions from which the  $x_i$  arise have means

$$\mu_i = \mu + \sigma \Delta a_{\nu_i} \quad (i = 1, 2, \dots, n) \quad (2.3.1)$$

$\bar{H}: \Delta \neq 0$  (or one-sided analogues, e.g.  $\Delta > 0$  when the  $a_i$  have the same sign)

*Model B*  $x_1, x_2, \dots, x_n$  arise independently from normal distributions with common mean,  $\mu$ . (Under  $H$  they have common variance  $\sigma^2$ .) There are known positive constants  $a_1, a_2, \dots, a_n$  (most of which will be zero), an unknown parameter  $\Delta$  and an unknown permutation  $(\nu_1, \nu_2, \dots, \nu_n)$  of  $(1, 2, \dots, n)$  such that the normal distributions from which the  $x_i$  arise have variances

$$\sigma_i^2 = \sigma^2 \exp(\Delta a_{\nu_i}) \quad (i = 1, 2, \dots, n) \quad (2.3.2)$$

$\bar{H}: \Delta > 0$  ( $\Delta < 0$  is irrelevant to the outlier problem).

These models are quite general with regard to the number of outliers in the data. Some particularizations, or modifications, are worth examining. For illustration we retain the normality assumption for  $F$ . Anscombe (1960a) considers the case of an outlier arising from a shift in the mean. He assumes that  $\sigma^2$  is known, that  $x_1, x_2, \dots, x_{n-1}$  is a random sample from  $N(\mu, \sigma^2)$  and that under the alternative hypothesis  $x_n$  arises independently from  $N(\mu + a\sigma, \sigma^2)$  with the value of  $a$  unknown. Guttman (1973b) declares

that under  $\bar{H}$  one of the  $x_i$  comes from  $\mathbf{N}(\mu + a, \sigma^2)$ ; which one is unknown; it may be any of the  $x_i$  with equal probabilities,  $1/n$ . In a non-Bayesian framework this extension of the slippage alternative hypothesis (which is best described as an *exchangeable* alternative hypothesis—see below) can be expressed

$$\bar{H}: \text{one of } H_i \ (i = 1, 2, \dots, n) \text{ holds,}$$

where

$$H_i: x_i \sim \mathbf{N}(\mu + a, \sigma^2). \quad (2.3.3)$$

If we are concerned with an outlier arising from a possible shift in scale (or dispersion) rather than location, we would take as the corresponding slippage-type alternative hypothesis

$$\bar{H}': \text{one of } H'_i \ (i = 1, 2, \dots, n) \text{ holds,}$$

where

$$H'_i: x_i \sim \mathbf{N}(\mu, b\sigma^2) \quad (b > 1). \quad (2.3.4)$$

Both  $\bar{H}$  and  $\bar{H}'$  are immediate analogues of the alternative hypothesis in multisample *slippage tests* in the special case of samples each of just *one* observation (hence our terminology). We shall be considering slippage tests in more detail in Chapter 5. If we wish to handle more than one outlier,  $\bar{H}$  and  $\bar{H}'$  need to be appropriately extended, of course, in the spirit of Ferguson's *models A* and *B*.

Again we encounter the anomaly that this type of alternative hypothesis is non-specific with regard to which observation corresponds with the location-shifted (or scale-shifted) distribution. Suppose we encounter a single upper outlier,  $x_{(n)}$ , in the case of a location-shifted alternative hypothesis. As in the case of the mixture alternative, we might hope to test the working hypothesis against the *specific* analogous alternative that  $x_{(n)}$  (specifically, rather than any single  $x_i$ ) has arisen from  $\mathbf{N}(\mu + a, \sigma^2)$ , with  $a > 0$ . As we have said, this prospect is not considered in the literature and it is not immediately obvious how such an alternative hypothesis should be expressed, even in the slippage context; see Section 3.1 for further discussion.

#### (v) Exchangeable alternative

A different approach to the form of the alternative hypothesis, extending the slippage formulation, is to be found in the work of Kale, Sinha, Veale, and others. Kale and Sinha (1971) and Veale and Kale (1972) were concerned respectively with estimating, and testing, the value of the mean,  $\theta$ , in an exponential distribution in a manner which is robust against the possibility of outliers. The model they employ to reflect the presence of, for example, a single outlier assumes in its general form that  $x_1, x_2, \dots, x_{i-1}$ ,

$x_{i+1}, \dots, x_n$  arise as independent observations from the distribution  $F$  of the initial model, whereas  $x_i$  is a random observation from a distribution  $G$ . It is further assumed that the index  $i$  of the aberrant (discordant) observation is equally likely to be any of  $1, 2, \dots, n$ . The random variables  $X_1, X_2, \dots, X_n$  are, on this model, not independent, but they are exchangeable. We shall call such an alternative hypothesis an *exchangeable* alternative.

The likelihood of the sample under the alternative hypothesis of a single discordant observation is

$$L\{\mathbf{x} | F, G\} = \frac{1}{n} \sum_{i=1}^n g(x_i) \prod_{j \neq i} f(x_j) \quad (2.3.5)$$

where  $f(x)$ ,  $g(x)$  are the probability (density) functions of the distributions  $F$  and  $G$ .  $F$  and  $G$  might be taken from distinct families of distributions, or they may correspond merely with different parameter values within a single-parameter family. The latter is the case in the papers referred to above, where

$$f(x) = (1/\theta)e^{-(x/\theta)} \quad (2.3.6)$$

and

$$g(x) = (b/\theta)e^{-(bx/\theta)} \quad (0 < b < 1). \quad (2.3.7)$$

Whilst the general form of the exchangeable alternative is clearly an extension of the slippage alternative, it comes very close to the latter in particular cases such as the exponential example above. Here it expresses a scale shift of the slippage alternative type and the corresponding likelihood is identical in form (if different in motivation) to that which would be employed in a *Bayesian* analysis of the slippage alternative with equal prior probabilities for the index of the observation which arises from the anomalous family,  $G$ , as described by Guttman (1973b).

A further extension of the exchangeable alternative is given by Joshi (1972b).

In applications of such a model for the study of outliers, the interest is on robust estimation or testing and hence is concerned with *accommodation* of outliers rather than tests of discordancy. Accordingly we shall return to such work for more detailed study in Section 2.6 and in Chapter 4.

When more than one outlier is observed the model is generalized as follows. For  $k$  discordant observations, we assume that  $x_{i_1}, x_{i_2}, \dots, x_{i_{n-k}}$  come from  $F$ , whilst  $x_{i_{n-k+1}}, \dots, x_{i_n}$  come from  $G_1, G_2, \dots, G_k$ . We might have some or all of the  $G_i$  identical. The association of the different observations with the different distributions is again assumed to occur at random.

In the cases of either one or several discordant observations certain relationships must exist between  $F$  and  $G$  (or  $G_1, \dots, G_k$ ) for it to be reasonable that the propriety of the model will be reflected in *outliers*.

Considering the case of just one discordant value, we need the observation from  $G$  to show up at one of the *extremes* of the sample. Specifically, if

$$P\{X_{(r)} \sim G\} = u_r \quad (2.3.8)$$

denotes the probability that the  $r$ th ordered value arises from  $G$  and  $dG/dF = \psi(x)$  is monotone (increasing or decreasing) in  $x$ , then the  $\{u_i\}$  can be shown to be monotone (increasing or decreasing). Thus the discordant observation from  $G$  is most likely to be either the smallest, or the largest, of  $x_1, x_2, \dots, x_n$ . This is surely what we would require of a model to explain a lower, or an upper, outlier. There are analogous results for *several* discordant observations. See Mount and Kale (1973).

#### (vi) Other alternative hypotheses

There seem to be no other formal proposals for modelling outlier occurrence (either as an alternative hypothesis in a test of discordancy, or as a model for robust accommodation of outliers). Some ideas relating to the investigation of multivariate outliers implicitly employ other modelling concepts. See Chapter 6.

Remarks by Kruskal (1960b) are relevant to the general question of modelling. In particular, he points the need for models which allow the occurrence of (measurement error) outliers to depend on the value that would have occurred had an error not taken place—but he makes no specific proposals.

#### (vii) Outlier proneness

The final topic in this section cannot really be claimed to constitute a model to describe the occurrence of outliers, although it is germane to model development.

Recalling the subjective manner in which sample observations may be declared outliers we stress again the element of ‘surprise’ they engender. Such surprise should ideally be conditional on the initial model we have in mind for the data, so that, for example, extreme values in a Cauchy sample would need to be even more extreme than those in a normal sample if we were to declare them ‘outliers’. A somewhat different attitude seems to be implied in recent work by Neyman and Scott (1971) and by Green (1974). Both works consider a method of distinguishing between families of distributions with regard to the differing extents to which they are likely to exhibit outliers. They define a concept of *outlier proneness*, and conclude, *inter alia*, that the families of log-normal and gamma distributions are outlier prone, whereas the family of Cauchy distributions is not.

The concept hinges on the probability  $P(\kappa, n \mid F)$  that a random sample of size  $n$  from a distribution  $F$  (in a family  $\mathcal{F}$ ) contains an extreme member  $x_{(n)}$  which exceeds  $x_{(n-1)}$  by more than an amount  $\kappa(x_{(n-1)} - x_{(1)})$ . If the

supremum of  $P(\kappa, n | F)$  over  $\mathcal{F}$  is strictly less than unity, then they call  $(\kappa, n)$ —*outlier-resistant*; otherwise it is  $(\kappa, n)$ —*outlier-prone*. If  $\mathcal{F}$  is  $(\kappa, n)$ —outlier-prone for all  $\kappa > 0$  and all  $n > 2$  it is *outlier-prone completely*. Green (1974) shows that this is so provided only that  $\mathcal{F}$  is  $(\kappa, n)$ —outlier-prone for *some*  $\kappa > 0$ ,  $n > 2$ . See also Kale (1975a, 1975b).

## 2.4 TEST STATISTICS

Whatever the form of the working and alternative hypotheses, indeed sometimes in total disregard of these, we can distinguish a small number of different forms of test statistic for tests of discordancy of outliers. These have obvious intuitive appeal and frequently have been demonstrated (often *subsequent* to their original introduction and use) to be supported by statistical test principles applied to appropriate models. For the moment we consider the qualitative nature of some test statistics; in Section 2.5 we review any wider statistical support that they enjoy and in Chapter 3 we consider in detail their precise form and application.

Augmenting the classification of Tietjen and Moore (1972) we can distinguish six basic types of test statistic. We shall consider these in turn. Some are more appropriate than others in different types of situation, for example, in examining a single upper outlier, or two lower outliers, or, perhaps, an upper and a lower outlier whilst safeguarding against the possibility of additional upper or lower outliers, and so on. We shall not consider such fine detail here, but will examine it more fully in Chapter 3. The only associated concept we will discuss at this stage is that of *masking* (see below). The basic types of test statistics are as follows.

### (a) *Excess/spread statistics*

These are ratios of differences between an outlier and its nearest or next-nearest neighbour to the range, or some other measure of spread of the sample (possibly omitting the outlier and other extreme observations). Examples are

$$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}$$

(Dixon, 1951, for examining an upper outlier  $x_{(n)}$ , avoiding  $x_{(1)}$ ) or (Irwin, 1925)

$$\frac{x_{(n)} - x_{(n-1)}}{\sigma}$$

where  $\sigma$  is the standard deviation in the basic model. Irwin's statistic assumes  $\sigma$  is known, and is particularly relevant for a normal distribution. Clearly we could replace  $\sigma$  with an estimate which might (perhaps usefully) be based on a restricted sample which excludes observations we wish to

protect against: such as the outlier  $x_{(n)}$  or other extremes. If an independent estimate of  $\sigma$  is available, this could also be used.

### (b) Range/spread statistics

Here we replace the numerator with the sample range; for example (David, Hartley, and Pearson, 1954; Pearson and Stephens, 1964)

$$\frac{x_{(n)} - x_{(1)}}{s}.$$

Again  $s$  might be replaced by a restricted sample analogue, independent estimate or known value of a measure of spread of the population. Using the range has the disadvantage that it is not clear without further investigation whether significant results represent discordancy of an upper outlier, a lower outlier, or both.

### (c) Deviation/spread statistics

This latter difficulty is partly offset by using in the numerator a measure of the distance of an outlier from some measure of central tendency in the data. An example (Grubbs, 1950) for a lower outlier is

$$\frac{\bar{x} - x_{(1)}}{s}$$

As for  $s$ ,  $\bar{x}$  might be based on a restricted sample, or replaced with an independent estimate, or population value, of some convenient measure of location. A modification uses maximized deviation in the numerator; for example,  $\max |x_i - \bar{x}|/s$  (Halperin *et al.*, 1955).

### (d) Sums of squares statistics

Somewhat different in form are test statistics expressed as ratios of sums of squares for the restricted and total samples: for example, the statistic

$$\sum_{i=1}^{n-2} (x_{(i)} - \bar{x}_{n,n-1})^2 / \sum_{i=1}^n (x_i - \bar{x})^2$$

where  $\bar{x}_{n,n-1} = \sum_{i=1}^{n-2} x_{(i)} / (n-2)$ , proposed by Grubbs (1950) for testing two upper outliers  $x_{(n-1)}, x_{(n)}$ .

### (e) High-order moment statistics

Statistics such as measures of skewness and kurtosis, not specifically designed for assessing outliers, can nonetheless be useful in this context; for example (Ferguson, 1961a)

$$\frac{n^{\frac{1}{2}} \sum (x_i - \bar{x})^3}{[\sum (x_i - \bar{x})^2]^{\frac{3}{2}}} \quad \text{and} \quad \frac{n \sum (x_i - \bar{x})^4}{[\sum (x_i - \bar{x})^2]^2}.$$

Another omnibus statistic of relevance to the testing of outliers is the W-statistic of Shapiro and Wilk (1965, 1972) and Shapiro, Wilk, and Chen (1968) already referred to. This consists for normal data of the ratio of the square of a particular type of linear combination of all the ordered sample values to the sum of squares of the individual deviations about the sample mean.

#### (f) *Extreme/location statistics*

Another class of test statistics takes the form of ratios of extreme values to measures of *location*. These are particularly relevant to examining outliers where the initial model is from the *gamma* family of distributions. As an example, a test of discordancy for an upper outlier may use

$$x_{(n)} / \bar{x}.$$

Such statistics have been examined by Epstein (1960a,b) and Likeš (1966). Some new results on the null distributions are given by Lewis and Fieller (1978).

In a review paper, Grubbs (1969) gives illustrative examples of the use of several different types of discordancy statistic, applying them to several sets of actual data.

One effect that many of these statistics must be prone to, to differing extents, is that of *masking*, that is to say, the tendency for the presence of extreme observations not declared as outliers to mask the discordancy of more extreme observations under investigation as outliers. The term *masking* is due to Murphy (1951); the phenomenon seems to have been first discussed by Pearson and Chandra Sekar (1936); some recent comments are made by McMillan (1971) and by Tietjen and Moore (1972) who give an empirical example of the masking effect. See also Section 3.2.

Some test statistics have been proposed to provide tests for several outliers simultaneously. One example is the statistic in (d) above due to Grubbs (1950) which is relevant to the case of *two* upper outliers. A different approach is to examine the outliers *in sequence*, using a hierarchical form of test. McMillan and David (1971) and McMillan (1971) describe what they call a 'sequential' test for several upper outliers (the terminology is confusing, since the test is not sequential in the usual sense of taking observations one at a time in order to draw conclusions as quickly as possible: it is in fact a fixed-sample-size test). They first examine the principal upper outlier  $x_{(n)}$  by means of a deviation/spread statistic of form  $[x_{(n)} - \bar{x}] / s$ , where  $s^2$  may be based on the sample data alone or combined with an independent estimate of variance. If  $x_{(n)}$  proves to be discordant on this basis, they proceed to apply a similar test to  $x_{(n-1)}$  in the reduced sample excluding  $x_{(n)}$  using the statistic

$$\frac{x_{(n-1)} - \bar{x}_n}{s'}$$

where  $\bar{x}_n$  and  $s'$  are the sample mean, and an estimate of spread, obtained on omission of  $x_{(n)}$ . If the result is significant,  $x_{(n-1)}$  is judged discordant and the procedure repeated for the next outlier, and so on until a non-discordant value is reached. The statistical properties of such a repeated procedure are investigated for the case where there are at most two outliers (McMillan and David, 1971; McMillan, 1971; Moran and McMillan, 1973). See Section 3.3 for fuller discussion of such 'sequential' (or as we shall call them *consecutive*) tests of outliers, also Tietjen and Moore (1972) and the earlier proposals by Dixon (1953) and Ferguson (1961a).

## 2.5 STATISTICAL PRINCIPLES UNDERLYING TESTS OF DISCORDANCY

In any context, a statistical test might be able to be constructed merely by setting up an intuitively appealing test statistic and rejecting or accepting some working hypothesis on the basis of the value of the test statistic. This is true of tests of discordancy for outliers. Indeed, the subjective basis of the outlier concept and the long history of its study has tended to encourage an informal attitude to proposals for 'outlier rejection'. At the very least, however, we need to be able to determine rejection criteria which relate to known significance levels: the distribution of the test statistic under the working hypothesis of no discordant outliers needs to be known. An initial filter thus operates: we can consider only those test statistics for which we know such *null* distributions.

But we must hope for more than this. To choose between rival tests of discordancy for a particular type of outlier manifestation we need to know something about the *power* of the rival tests. This requires both the specification of an alternative hypothesis to explain the outliers (a topic discussed at some length above) and also the ability to handle the often complicated distributional forms of the test statistic under such an alternative hypothesis. In this respect the field of viable tests of discordancy becomes even more limited.

Ideally we should wish to go even further and construct tests which at their conception seek to express optimality properties or at least to satisfy certain useful practical constraints. Thus if we cannot (as is inevitably the case for most outlier tests) obtain tests which are globally *uniformly most powerful* we can at least strive for *local optimality* or *unbiasedness* or the satisfaction of certain *invariance* properties. Alternatively we might choose to construct tests by some accredited practical method, such as the *maximum likelihood ratio principle*, in the hope that the frequently encountered useful characteristics of such a method transfer to the outlier problem.

We shall now consider various tests of discordancy which have some sounder basis than mere intuitive appeal. This matter will be considered more fully in the detailed discussion of tests of discordancy in Chapter 3.

We must stress that the attribution of any optimality properties is crucially dependent on the adopted form for the alternative hypothesis. The choice of an alternative hypothesis is problematical; this uncertainty in turn may reduce the utility of any apparent ‘optimum’ properties of a test of discordancy. We must acknowledge this dilemma in reacting to the following results.

Most study of outliers assumes an initial normal distribution, and the most common alternative hypothesis is of the *slippage* type. Ferguson (1961a) demonstrates that certain quasi-optimum tests of discordancy can be constructed in such situations. He considers, in the context of his *model A* (slippage of the mean), tests which are *invariant* with respect to the labelling of the observations and to changes of scale and location. (The change of scale must be effected by multiplication by a common *positive* quantity whilst the distribution for each individual observation may suffer a specific change of location.) We recall that *model A* declares (see 2.3.1)

$$\mu_i = \mu + \sigma \Delta a_{\nu_i}$$

Ferguson shows that a *locally best invariant* test of size  $\alpha$  exists for testing  $H: \Delta = 0$  against the one-sided alternative  $\bar{H}: \Delta > 0$ . It takes the following form: if  $\mu_3(a) \geq 0$  reject  $H$  whenever  $\sqrt{b_1} \geq K_1$  where  $K_1$  is chosen to yield a test of size  $\alpha$ , and

$$\mu_3(a) = \frac{1}{n} \sum_{i=1}^n (a_i - \bar{a})^3 \quad \left( \bar{a} = \frac{1}{n} \sum_{i=1}^n a_i \right) \quad (2.5.1)$$

$$b_1 = \frac{\sqrt{n} \sum (x_i - \bar{x})^3}{[\sum (x_i - \bar{x})^2]^{\frac{3}{2}}} \quad (2.5.2)$$

(this is just the coefficient of skewness statistic described above).

For a two-sided test, where the alternative hypothesis is  $\bar{H}' : \Delta \neq 0$ , there is a *locally best unbiased invariant test* of size  $\alpha$  which takes the form: if  $k_4(a) \geq 0$  reject  $H$  whenever  $b_2 \geq K_2$  where  $K_2$  is chosen to yield a test of size  $\alpha$ ,  $k_4(a)$  is the fourth  $k$ -statistic of  $a_1, a_2, \dots, a_n$  and

$$b_2 = \frac{n \sum (x_i - \bar{x})^4}{[\sum (x_i - \bar{x})^2]^2} \quad (2.5.3)$$

(the coefficient of kurtosis).

With *model B*, the alternative hypothesis is, with

$$\sigma_i^2 = \sigma^2 \exp(\Delta a_{\nu_i}),$$

$\bar{H}'' : \Delta > 0$ . Under the same invariance requirements a *locally best invariant test* of size  $\alpha$  exists and leads to rejection if  $b_2 > K$ , where  $K$  is chosen to yield a test of size  $\alpha$ .

Still restricting attention to single outliers and an initial normal distribution, there are results for *slippage tests* (where we have several samples from

normal distributions, one or more of which may have a mean, or variance, which differs from the others—see Chapter 5) which particularize, when samples are all of size 1, to tests of discordancy on a slippage type alternative hypothesis. Paulson (1952b) considers a multi-decision formulation where under  $\mathcal{D}_0$  we decide that the observations all come from  $N(\mu, \sigma^2)$  whereas under  $\mathcal{D}_i$  ( $i = 1, 2, \dots, n$ ) we decide that  $\mu_i$  has slipped to  $\mu + a$  ( $a > 0$ ). Under the restrictions that if all means are  $\mu$  we accept  $\mathcal{D}_0$  with probability  $1 - \alpha$  and that the decision procedure is invariant with respect to the index of the slipped distribution and to positive change of scale and arbitrary change of origin, he shows there to be an optimum procedure in the sense of *maximizing the probability of making the correct decision* when slippage to the right has occurred. With the modification of proof described by Kudo (1956a) to cope with samples of size 1, this leads to the optimum decision rule: if

$$t = \frac{x_{(n)} - \bar{x}}{\left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{\frac{1}{2}}},$$

then when  $t \leq h_\alpha$  conclude no discordant outlier, whilst if  $t > h_\alpha$  conclude that  $x_{(n)}$  is discordant, where  $h_\alpha$  is chosen to ensure that  $\mathcal{D}_0$  is adopted with probability  $1 - \alpha$  in the null situation.

In Paulson (1952b) the procedure is shown to be the Bayes solution when equal prior probabilities are assigned to  $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n$ . David (1970, p. 180) shows that it has corresponding support in non-Bayesian terms using the Neyman-Pearson lemma. Kudo (1956b) demonstrates that the optimality property remains when slippage in the mean is accompanied by decrease in variance for the slipped distribution, whilst Kapur (1957) adduces an *unbiasedness* property for the Paulson procedure in the sense that the probability of *incorrectly* taking any of the decisions  $\mathcal{D}_0, \mathcal{D}_1, \dots, \mathcal{D}_n$  never exceeds the probability of *correctly* taking any one of these decisions. David (1970, p. 182) also remarks that an obvious modification to allow for an independent estimate of  $\sigma^2$  remains optimum in the Paulson sense. Kudo (1956a) extends Paulson's results to the case of slippage either to the right or to the left. Truax (1953) presents the immediately parallel results for unidirectional slippage of the variance of one of the normal distributions.

A further extension, relating to simultaneous slippage of the means of two distributions by equal amounts, but in opposite directions, is given by Ramachandran and Khatri (1957).

(In all work relating to investigation of *single* outliers, we must remain ever vigilant to the prospect of *masking* if more than one outlier is present.)

In the work of Ferguson and of Paulson we note a basic distinction between the chosen measures of performance of the tests: power and maximum probability of correct action. This focuses attention on the thorny issue (largely unresolved to date) of what constitutes an appropriate performance criterion for a test for discordancy. Some comments on

the relative merits of five possible performance characteristics are made by David (1970, pp. 184–190). For a detailed discussion of these and related issues see Section 3.2.

A development of Paulson's ideas for coping with several outliers is given by Murphy (1951) and further studied by McMillan (1971). The development is limited: the alternative model is that  $k$  outliers all arise from a common shift in the mean (by the same amount and in the same direction). For slippage to the right, the optimum test statistic is

$$(x_{(n)} + x_{(n-1)} \dots + x_{(n-k+1)} - k\bar{x})/s.$$

There seems to be a dearth of results on optimal tests for discordancy when the initial model is *non-normal*. For the gamma family, statistics based on  $x_{(n)}/\bar{x}$  do at least have the advantage of arising from a *maximum likelihood ratio criterion* using a slippage type alternative model for a single upper outlier (this is true also of  $(x_{(n)} - \bar{x})/s$  when the initial distribution is normal). But beyond this we appear to be able to make few claims of statistical respectability in the sense of optimum (or even practically desirable) performance characteristics for tests of discordancy with non-normal initial distributions or multiple outliers (although Ferguson's statistics  $\sqrt{b_1}$  and  $b_2$  remain optimum in normal samples for multiple outliers whose number,  $k$ , merely satisfies the reasonable constraints  $k < 0.5n$  and  $k < 0.31n$ , respectively). The performance of some specific (intuitively based) tests has been examined empirically, by simulation, or theoretically, and the results will be discussed where appropriate in Chapter 3. But inter-comparisons revealing uniform superiority of one test over another are conspicuously lacking.

Discussion of the behaviour of non-parametric tests for outliers (such as those proposed by Walsh 1950, 1959, 1965) will be deferred to Chapter 8.

In concluding this section brief comment is necessary on two related general matters. Some proposed test statistics are clearly suited to one-sided tests, e.g.

$$\frac{x_{(n)} - \bar{x}}{s}$$

whilst others relate to two-sided tests, e.g.

$$\frac{\max |x_i - \bar{x}|}{s}$$

David (1970, p. 175) discusses a possible conflict of choice in relation to which of these types of test should be used in practice. He remarks as follows:

Indeed, the question may be raised whether we should not always use a two-sided test, since applying a one-sided test in the direction indicated as most promising by the sample at hand is clearly not playing fair. This criticism of what is often done in

practice is valid enough, except that sticking to a precise level of significance may not be crucial in exploratory work, frequently the purpose of tests for outliers. Strictly speaking, one-sided tests should be confined to the detection of outliers in cases where only those in a specified direction are of interest, or to situations such as the repeated determination of the melting point of a substance, where outliers due to impurities must be on the low side since impurities depress the melting point. Similar arguments show that it is equally incorrect to pick one's outlier test after inspection of the data.

This is a commonly expressed viewpoint but it is in opposition to the attitude we have adopted in this book. We define outliers in subjective terms relative to a particular set of data—the data themselves initiate our interest in outliers. If we declare there to be an upper outlier,  $x_{(n)}$ , it seems natural, therefore, to use the appropriate one-sided criterion, and to ascribe discordancy if, say

$$(x_{(n)} - \bar{x})/s > C$$

for some suitable value of  $C$ . Of course, if we wish to protect ourselves against other outliers we should reflect this in the choice of test statistic. Also, if our suspicions about  $x_{(n)}$  are well founded then use of  $\max |x_i - \bar{x}|/s$  rather than  $(x_{(n)} - \bar{x})/s$  is not going to be materially important. Surely what matters is

- (i) our declaration of the outliers;
- (ii) the alternative model we employ.

This relates to the earlier remarks about the lack of regard for the specific outlier in the formulation of the alternative hypothesis and the test of discordancy. A few results on performance characteristics of tests do relate to this matter. David and Paulson (1965) consider the behaviour of some tests of discordancy in terms of outlier-specific criteria (included in the five discussed by David, 1970, pp. 185–186) such as the probability that a particular sample member is significantly large *and* is the largest one, or the probability that a particular sample member is significantly large *given* that it is the largest one. See again Section 3.3 for more details.

Some Bayesian methods for examining outliers have been published. Those whose prime function is to provide a means of examining outliers *per se*, rather than to promote general inference procedures which are robust against the presence of possible outliers, are relevant to the current discussion. Bayesian methods for *accommodation* of outliers are briefly discussed in Section 2.6; the general Bayesian scene in relation to outliers is described in detail in Chapter 8.

As we remarked in Section 2.3 Guttman (1973b) presents a method of 'detection of spuriousity' in a highly specific situation. The data are assumed, *ab initio*, to arise as independent observations from  $N(\mu, \sigma^2)$ . To allow for the possibility of a single observation having been 'generated by a spurious source, where the spuriousity is of the mean shift type' an alternative model is

adopted in which one observation (arising at random from the  $n$  in the sample) comes from  $N(\mu + a, \sigma^2)$ . Starting with a non-informative prior distribution for  $\mu$ ,  $\sigma$ , and  $a$ , the marginal posterior distribution of  $a$  is determined and its form is exploited to construct a principle for determining whether 'spuriosity has or has not occurred'. Since our potential application of this approach is likely to be triggered by the occurrence of an outlier, the principle can be regarded as a means of assessing discordancy of a single outlier.

Other work in the Bayesian idiom, by Box and Tiao (1968), De Finetti (1961), Dempster and Rosner (1971), Guttman; and others will be discussed in Chapter 8.

## 2.6 ACCOMMODATION OF OUTLIERS: ROBUST ESTIMATION AND TESTING

There has been an increasing interest over recent years in statistical procedures which provide a measure of protection against uncertainties of knowledge of the data generating mechanism. These include *robust* methods for estimating or testing summary measures of the underlying distribution: where the estimators or tests retain desirable statistical properties over a range of different possible distributional forms. Alternatively, procedures which have been derived to suit the specific properties of a particular distribution become even more appealing if they can be shown to be robust (i.e. to retain worthwhile operating characteristics) when the distribution proves to be different from that which promoted the procedures. An informative review of 'robust statistics' is given by Huber (1972); other important references are Tukey (1960), Huber (1964), Bickel (1965), Jaekel (1971a), Hampel (1974), and Hogg (1974).

An obvious area in which we may wish to seek the protection of robust statistical methods is where we encounter, or anticipate, outliers in a set of data. As one example, extreme observations clearly have an extreme effect on the value of a sample variance! If we are interested in estimating a parameter in an initial model, but are concerned about the prospects of outliers, whether arising from random execution errors of no specific relevance to our studies, or from random measurement error, we would want to use an estimator which is not likely to be highly sensitive to such outliers. A simple (if somewhat paranoid) example is to be found in the use of the sample *median* as an estimator of location.

We must, of course, be ever conscious of the overriding importance of the alternative hypothesis. If outliers arise because our initial model does not reflect the appropriate degree of inherent variation (we really need, say, a fatter-tailed distribution rather than the ubiquitous normal distribution initially adopted) then omission of extreme values to 'protect against outliers' is hardly a robust policy for estimating some measure of dispersion, say

the variance. Rather than appropriately reducing the effect of extreme values it encourages underestimation!

If, on the other hand, a reasonable alternative hypothesis is of one of the types which expresses contamination of the initial model (perhaps expressing low-probability mixing, or slippage of one or two discordant values) the estimation or testing of parameters in the initial model may well be the matter of principal interest and it is sensible to employ robust procedures to protect against the occasional low-probability component or slipped value.

The idea that we may wish, in this spirit, to do more than *reject* outliers, that is, to devise statistically respectable means of *accommodating* them in a wider inferential scheme addressed to the initial model, takes our interest away from tests of discordancy. The outliers themselves are no longer of prime concern. We wish to proceed safely in spite of them! This is of the essence of the robustness concept.

Some interest in this alternative view of the outlier problem begins to show itself in quite early work. The ideas of Glaisher (1872), Newcomb (1886), Mendeleev (1895), Student (1927), and Jeffreys (1932) amount to reducing the weight attached to extreme values in estimation, the latter paper paying specific regard to outliers as the extreme members of the sample.

A review of later work which implicitly or explicitly attempts to accommodate outliers in the inference process conveniently divides itself into two parts. The first contains those methods of estimation which *implicitly* protect against outliers in placing less importance on extreme values than on other sample members. A variety of general methods of this type exist and we shall briefly examine some of these robust *blanket procedures*. The second part of the study of accommodation of outliers is specifically concerned with the nature of the initial model, and of the explanatory model for outliers, and derives methods of estimation or testing designed specifically to suit those models. Some examples of such *specific accommodation techniques* are also given below.

### *Blanket procedures*

To illustrate robust statistical methods which provide *en passant* some protection against outliers, whilst not specifically concerned with outliers, we consider a variety of methods of robust estimation of a location parameter,  $\mu$ . We single out four for discussion.

The aim throughout is to reduce the influence of extreme observations in the sample on the value of the estimate of  $\mu$ . Since outliers manifest themselves as extreme observations this has the effect of protecting against their presence as well as meeting the prime object of rendering less dramatic the effect of the tail behaviour of the generating distribution on the

estimation of  $\mu$ . Cox and Hinkley (1974, Section 9.4) review robust estimation of location, and Andrews *et al.* (1972) present a major sampling study of different estimators.

One obvious way of achieving the objective is to use estimators having the form of linear combinations of ordered sample values

$$\tilde{\mu} = \sum c_i x_{(i)} \quad (2.6.1)$$

where the weights  $c_i$  are lower in the extremes than in the body of the data set. The 'ultimate' example of such *linear order statistics estimators* (called '*L*-estimators' by Huber, 1972) is the sample median where  $c_i = 0$  for all but the middle, or two middle, ordered observations. In contrast with the sample mean ( $c_i = 1/n$ :  $i = 1, 2, \dots, n$ ) this can on occasions effect considerable improvement: for example, for the Cauchy distribution where an even more drastic policy of assigning *negative* weight to extreme observations can yield further improvement still (Barnett, 1966). Other examples are found in the use of  $\alpha$ -*trimmed means*, where a prescribed proportion,  $\alpha$ , of the lower and upper ordered sample values are omitted and  $\mu$  estimated by the (unweighted) average of the retained values, or in the use of specific combinations of a few sample quantiles (for example see Gastwirth, 1966).

Among other proposed robust estimators of  $\mu$  which, *en passant*, help to eliminate the effect of outliers we have the *Winsorized mean*, and estimators using the principle of '*jackknifing*'. An example of a Winsorized mean is obtained by 'collapsing' the most extreme (upper and lower) sample values to their nearest neighbours in the ordered sample and taking an unweighted average from the modified sample. This is just another *L*-estimator, with  $c_1 = c_n = 0$ ,  $c_2 = c_{n-1} = 2/n$ ,  $c_i = 1/n$  ( $i = 3, 4, \dots, n-2$ ). (See Tukey and McLaughlin, 1963, for a discussion of Winsorized, and trimmed, means.) The principle of *jackknifing* was originally proposed by Quenouille (1956) as a method of reducing bias in estimators; the term was introduced by Tukey. Suppose  $\tilde{\mu}_n$  is an estimator of  $\mu$  to be evaluated from a sample of size  $n$ , and  $\tilde{\mu}_{n-1}$  the corresponding estimator evaluated from a sample of size  $n-1$ . Considering the random sample  $x_1, x_2, \dots, x_n$  we have a single value for  $\tilde{\mu}_n$  from the whole sample but  $n$  possible values  $\tilde{\mu}_{n-1,i}$  for  $\tilde{\mu}_{n-1}$ , each obtained on omission of a single observation  $x_i$ , ( $i = 1, 2, \dots, n$ ). We define *jackknifed pseudo-values*

$$z_{ni} = n\tilde{\mu}_n - (n-1)\tilde{\mu}_{n-1,i} \quad (i = 1, 2, \dots, n) \quad (2.6.2)$$

and construct an estimator in terms of these values. The principal advantages of this approach are that it may reduce bias in estimators and can yield useful estimates of their variances. This latter facility is the major advantage of jackknifing.

Two other general principles of robust estimation yield what Huber (1972) calls maximum likelihood type estimators (*M*-estimators) and rank test estimators (*R*-estimators).

*M*-estimators are obtained by solving an equation of the form

$$\sum_{i=1}^n \psi(x_i - \tilde{\mu}) = 0 \quad (2.6.3)$$

to obtain an estimator  $\tilde{\mu}$  where  $x_1, x_2, \dots, x_n$  is a random sample and  $\psi(u)$  is some weight function with desirable features. For example if  $|\psi(u)|$  is small for large  $|u|$ ,  $\tilde{\mu}$  will discount extreme sample values and protect against outliers. The particular choice  $\psi(u) = f'(u)/f(u)$ , where the distribution has probability density function of the form  $f(x - \mu)$ , will of course yield  $\tilde{\mu}$  as a solution of the likelihood equation.

Hodges and Lehmann (1963) were the first to remark that estimators of  $\mu$  could be obtained from certain rank test procedures, such as the Wilcoxon test. Such *R*-estimators often prove to be robust. An example is given by Huber (1972) who considers the two-sample rank test for location shift. If the samples are  $x_1, x_2, \dots, x_n$  and  $y_1, y_2, \dots, y_n$  the test statistic is

$$W(x_1, x_2, \dots, x_n; y_1, y_2, \dots, y_n) = \sum_1^n J[i/(2n+1)] V_i \quad (2.6.4)$$

where  $J[i/(2n+1)]$  is some suitably chosen function of the empirical distribution function for the combined sample, and  $V_i = 1$  if the  $i$ th ordered value in the combined sample is one of the  $x$ -values (otherwise  $V_i = 0$ ). We can derive an estimator  $\tilde{\mu}$  as the solution of

$$W(x_1 - \tilde{\mu}, x_2 - \tilde{\mu}, \dots, x_n - \tilde{\mu}; -x_1 + \tilde{\mu}, -x_2 + \tilde{\mu}, \dots, -x_n + \tilde{\mu}) = 0 \quad (2.6.5)$$

and the asymptotic behaviour of  $\tilde{\mu}$  is obtained from the power function of the test. For symmetric distributions *R*-estimators can, for an appropriate choice of  $J(\cdot)$ , be asymptotically efficient and asymptotically normally distributed.

peculiarities of a particular set of data. *Adaptive estimators* for outlier protection have not been widely studied, but some possibilities have been considered. For example, the trimming factor  $\alpha$  in the  $\alpha$ -trimmed mean might be chosen to exclude outliers in the data (Jaeckel, 1971a). Takeuchi (1971), Johns (1974), consider estimators based on weighted combinations of the sums of subgroups of ordered sample values, with the weights determined empirically. Somewhat similar is the simplified version of the Hodges-Lehmann estimator due to Bickel and Hodges (1967): the median of the quasi-mid-ranges

$$\frac{1}{2}\{x_{(i)} + x_{(n+1-i)}\} \quad (i = 1, 2, \dots, [n/2]).$$

#### Specific accommodation techniques

A number of proposals have been made for taking specific account of outliers in the estimation or testing of parameters in the initial probability model. We shall consider two of these in detail.

The paper by Anscombe (1960a) expounds a basic philosophy in the matter of accommodating outliers and applies this to a range of different situations. Although he talks of 'rejection rules' for outliers these are really methods of estimating parameters in the presence of outliers, rather than tests of discordancy.

Illustrating his ideas for a random sample  $x_1, x_2, \dots, x_n$  from  $N(\mu, \sigma^2)$  where  $\sigma^2$  is known, where  $\mu$  is to be estimated, and where there may be a single outlier, he proposes a rule: *If  $x_M$  maximizes  $|x_i - \bar{x}|$  ( $i = 1, 2, \dots, n$ ) reject  $x_M$  if  $\max_i |x_i - \bar{x}| > C\sigma$  for some suitable choice of  $C$ ; otherwise reject no observations. Estimate  $\mu$  by the mean of the retained observations:*

$$\begin{aligned}\tilde{\mu} &= \bar{x} \quad \text{if } \max_i |x_i - \bar{x}| < C\sigma \\ &= \bar{x} - \max_i |x_i - \bar{x}|/(n-1) \quad \text{if } \max_i |x_i - \bar{x}| > C\sigma\end{aligned}$$

For more than one outlier he suggests repeated use of this rule. That is, apply the rejection criterion until no further sample values are rejected and estimate  $\mu$  by the mean of the remaining observations.

Anscombe goes on to consider the properties of such a rule with respect to the 'premium' payable and the 'protection' afforded: essentially the loss of efficiency under the basic model and the gain in efficiency under the alternative model (see Section 4.1). The rule is extended to deal with more complex sets of data, including results from a factorial design experiment (see Section 7.1.2).

Others have adopted a similar *premium-protection* approach to the handling of outliers. In particular Guttman and Smith (1969, 1971) examine further the problems of determining the premium and the protection levels for the Anscombe rejection rule and extend the investigation to situations where the outlier is not rejected but instead the sample is Winsorized (or *modified Winsorization*) takes place with the observation yielding the maximum value of  $|x_i - \bar{x}|$  being replaced by the closer of the two values  $\bar{x} \pm C\sigma$ .

A different approach is adopted by Kale and Sinha (1971), Veale and Kale (1972), Sinha (1973a, 1973b), and Kale (1975c). Having an interest in estimating or testing the value of the scale parameter in an exponential distribution Kale and Sinha (1971) postulate an *exchangeable* alternative hypothesis to account for a possible outlier. The working hypothesis declares that  $x_1, x_2, \dots, x_n$  arise at random from a distribution with probability density function

$$f(x, \theta) = \frac{1}{\theta} \exp\left(\frac{-x}{\theta}\right), \quad (2.6.6)$$

whilst under the alternative hypothesis one of the observations (which is equally likely to be any particular one) arises from the distribution  $f(x, \theta/b)$

where  $0 < b < 1$ . The estimators considered are  $L$ -estimators (linear combinations of ordered sample values). It is shown that the observation most likely to be the aberrant one is  $x_{(n)}$  and that an optimum estimator of  $\theta$  (minimizing the mean square error) based on the first  $m < n$  ordered sample values has the form

$$\tilde{\theta}_m = \frac{1}{m+1} \left[ \sum_1^{m-1} x_{(i)} + (n-m+1)x_{(m)} \right]. \quad (2.6.7)$$

No firm prescription is given about choice of  $m$ .

Veale and Kale (1972) consider tests of  $H: \theta = 1$  versus  $\bar{H}: \theta > 1$  based on  $\tilde{\theta}_m$  as test statistic. The case of  $\tilde{\theta}_{n-1}$  is considered in detail with regard to test power and extensions of the Anscombe concepts of premium and protection. Sinha (1973a) extends the study of the efficiency of the estimator  $\tilde{\theta}_m$  and elsewhere (1973c) he considers the implications of the *exchangeable* type model for estimation of scale and location parameters simultaneously in a *two-parameter exponential distribution*.

The estimator  $\tilde{\theta}_m$  has the form of a Winsorized mean, and was advanced on a premium-protection basis. In Kale (1975c) the maximum likelihood method is applied to the same type of situation but where the initial, and aberrant, distributions ( $F$  and  $G$  in the notation of Section 2.3) can have a more general form: as any members of the single parameter exponential family. The possibility of more than one outlier (observation from  $G$ ) is also entertained. Maximum likelihood estimators are shown to have the form of trimmed means, rather than Winsorized means.

The first attempt to specifically accommodate the prospect of outliers in estimation or testing situations seems to be that of Dixon (1953). Other aspects of the problem include the use of Bayesian methods (Gebhardt, 1964, 1966; Sinha, 1972, 1973b) and explicit study of the estimation of  $\mu$  from samples of size 3, where usually  $\mu$  is estimated from the two closest observations (Seth, 1950; Lieblein, 1952, 1962; Willke, 1966; Anscombe and Barron, 1966; Veale and Huntsberger, 1969; Guttman and Smith, 1969; Desu, Gehan, and Severo, 1974).

Fuller details and illustrations of methods for accommodating outliers in statistical analyses are given in Chapter 4.

## CHAPTER 3

### *Discordancy Tests for Outliers in Univariate Samples*

We have now discussed in some detail the meaning of the term ‘outlier’, the nature of the outlier problem, and the variety of contexts in which outliers can arise. We have also drawn the distinction between different types of action which may be called for in response to an outlier: rejection, and omission from the subsequent analysis; adjustment of its value for purposes of estimation from the whole sample; using it as a clue to the existence of some previously unsuspected and possibly interesting factor; interpreting it as a signal to find a more appropriate model for the data. As a prerequisite to all of these, and an indispensable one to all except the value adjustment procedure (which can, if so desired, be carried through automatically whatever the values of the observations), a *detection* procedure must be undertaken; a statistical test, termed here a *test of discordancy*, to decide whether or not the outlier is to be regarded as a member of the main population. Such tests are very often referred to in the literature as tests for the *rejection* of outliers, but, as we have stressed, rejection is not the only course open when an observation is detected as foreign to the main data set. In this and later chapters we deal with tests of discordancy in different situations. We start in this chapter with the simplest situation: when the data, with the possible exception of any outliers, form a sample from a univariate distribution from a prescribed family (for example, gamma of unknown parameter; exponential; normal; normal with known variance).

It is not difficult in any particular situation to propose reasonable-looking test statistics. (It may be quite another matter, of course, to ascertain the critical values or percentage points against which the value of any such statistic should be judged, to determine the distribution of the statistic on the assumption that the outlier is consistent with the rest of the data, and to assess the advantages and disadvantages of the test procedure.) Suppose, for example, that a civil engineer, wishing to find the mean crushing strength of cement made from a particular mix, makes up a set of ten test cubes from

the mix, allows a suitable hardening period, and then determines their strengths in p.s.i. to be as follows:

$$790, 750, 910, 650, 990, 630, 1290, 820, 860, 710.$$

The value 1290 seems to him to be out of line with the other nine values and he wishes to test it as an outlier. What test criteria might he use? The choice depends in the first place on the form of the distribution of crushing strengths of similar cubes from the same mix. Experience suggests that crushing strengths are normally distributed. However, the mean and variance will not be known.

As regards the test criterion, we must surely expect that units of measurement are irrelevant: that any test is invariant with respect to changes of scale and origin in the data. For example, the ten values in the sample

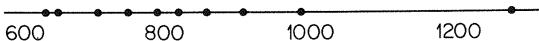
$$4.79, 4.75, 4.91, 4.65, 4.99, 4.63, 5.29, 4.82, 4.86, 4.71$$

are linear transforms of the ten values in the first sample. Any test for 1290 as an outlier in the first sample must give the same results when applied to 5.29 viewed as an outlier in the second sample. Had, say, an exponential distribution been assumed instead of a normal distribution, a test procedure would only need to remain unaltered under changes in scale, not shifts in origin, since practical support for a single (scale) parameter exponential model rests on a *natural* origin of measurement—see below.

Using the notation introduced in Section 2.3, let us arrange the ten values in ascending order and name them  $x_{(1)}, x_{(2)}, \dots, x_{(10)}$  respectively:

$$\begin{array}{cccccccccc} x_{(1)} & x_{(2)} & x_{(3)} & x_{(4)} & x_{(5)} & x_{(6)} & x_{(7)} & x_{(8)} & x_{(9)} & x_{(10)} \\ 630 & 650 & 710 & 750 & 790 & 820 & 860 & 910 & 990 & 1290 \end{array}$$

The figure shows these ten values as points on a line:



The reason the outlier  $x_{(10)}$  appears aberrant is because it is ‘widely separated’ from the remainder of the sample *in relation to the spread of the sample*. This leads one to think of test statistics of the form  $N/D$ , where the numerator  $N$  is a measure of the separation of  $x_{(10)}$  from the remainder of the sample and the denominator  $D$  is a measure of the spread of the sample. For the reason given above,  $D$  must be of the same dimensions as  $N$ , i.e. in this example  $D$  and  $N$  would both be in p.s.i. For  $N$  one might consider using the separation of  $x_{(10)}$  from its nearest neighbour  $x_{(9)}$ , i.e.  $x_{(10)} - x_{(9)} = 300$ ; or again the separation of  $x_{(10)}$  from the other nine values considered as a group, say specifically from their mean  $\bar{x}' = 790$ . For  $D$  one might use the range of this group,  $x_{(9)} - x_{(1)} = 360$ , or the spacing  $x_{(9)} - x_{(8)} = 80$  which is markedly less than  $x_{(10)} - x_{(9)}$ , or perhaps the standard deviation  $s' = 119$  of the nine values. These considerations suggest as possible test statistics such

quantities as

$$y(9, 10; 1, 9) = \frac{x_{(10)} - x_{(9)}}{x_{(9)} - x_{(1)}} \quad (\text{value here} = \frac{300}{360} = 0.83),$$

$$y(9, 10; 8, 9) = \frac{x_{(10)} - x_{(9)}}{x_{(9)} - x_{(8)}} \quad (\text{value here} = \frac{300}{80} = 3.75),$$

$$T' = \frac{x_{(10)} - \bar{x}'}{s'} \quad (\text{value here} = \frac{500}{119} = 4.20).$$

Statistics of the form

$$y(r, s; p, q) = \frac{x_{(s)} - x_{(r)}}{x_{(q)} - x_{(p)}} \quad (3.0.1)$$

—which we shall term *Dixon* statistics—have been investigated by Dixon (1950, 1951), Likeš (1966) and others, and some percentage points have been tabulated by Dixon; the  $y$ -notation is due to Likeš. An attractive alternative is to judge the outlier by the ratio of the spacing  $x_{(10)} - x_{(9)}$  to the range of all ten values including the outlier, giving

$$y(9, 10; 1, 10) = \frac{x_{(10)} - x_{(9)}}{x_{(10)} - x_{(1)}} \quad (\text{value here} = \frac{300}{660} = 0.45),$$

but this is effectively the same statistic as  $y(9, 10; 1, 9)$ , since clearly

$$\frac{1}{y(9, 10; 1, 10)} - \frac{1}{y(9, 10; 1, 9)} = 1.$$

In a similar way, the statistic

$$T = \frac{x_{(10)} - \bar{x}}{s} \quad (\text{value here} = \frac{1290 - 840}{194} = 2.32),$$

where  $\bar{x}, s$  are the mean and standard deviation of all ten values including the outlier, is equivalent to  $T'$  since the two quantities are functionally related; in fact (see Section 3.1)

$$\frac{(n-1)^2}{nT^2} - \frac{n(n-2)}{(n-1)T'^2} = 1. \quad (3.0.2)$$

Properties of the test based on  $T'$  (or  $T$ ) have been discussed by Pearson and Chandra Sekar (1936), Grubbs (1950), and others, and tables of percentage points are given by Grubbs in the same reference.

As remarked earlier, it is easy to propose other test statistics for the above outlier example, for instance

$$\frac{x_{(10)} - \bar{x}}{x_{(10)} - x_{(1)}}.$$

To the best of our knowledge the properties of this statistic have not been studied and no percentage points are available, so no practical use can be made of it. This is probably no great loss, as there are reasons to believe that it has no particular advantages.

Consider now a different example. The table shows the lengths of stay (in days) of 92 patients in a hospital observation ward before they were transferred to a main ward (data by kind permission of J. Hoenig: Hoenig and Crotty, 1958, refers).

Length of stay in days	1	2	3	4	5	6	7	8	9	10	11	21 Total
Number of patients	11	18	28	8	12	5	5	1	1	1	1	92

Regarding  $x_{(92)} = 21$  as an outlier, what criterion might be used for identifying it? The assumption of a normal distribution for the lengths of stay is not plausible, but there is some evidence to support the use of a gamma distribution with origin at zero. With such a distribution, a test criterion is required to be invariant under changes of scale. As before, we look for test statistics of the form  $N/D$  where  $N$  measures the separation of the outlier from the rest of the sample, and  $D$ , in the same units as  $N$ , measures the spread of the sample.

If the underlying gamma distribution (denoted  $\Gamma(r, \lambda)$ ) has parameters  $r$  and  $\lambda$ , i.e. if it has probability density function

$$f(x) = \lambda(\lambda x)^{r-1} e^{-\lambda x}/\Gamma(r), \quad (3.0.3)$$

then its mean is  $r/\lambda$  and its variance is  $r/\lambda^2$ . If  $r$  is known, but  $\lambda$  is unknown, then the spread of the distribution can be measured not only by the sample standard deviation  $S$  but also by the sample mean  $\bar{x}$  or equivalently by the sample sum  $\sum x_i$ . This suggests that a useful statistic for identifying the upper outlier  $x_{(92)}$  would be

$$\frac{x_{(92)} - \bar{x}}{\bar{x}}$$

or equivalently

$$\frac{x_{(92)}}{\sum x_i}$$

where  $\sum x_i$  is the sum of all 92 observations.

As with  $T$  and  $T'$  in the normal case discussed above, the statistic is functionally related to, and hence equivalent to,

$$\frac{x_{(92)}}{\sum' x_i}$$

where  $\sum' x_i$  is the sum of the 91 observations omitting the outlier.

This statistic would of course have been inappropriate for judging an upper outlier in a *normal* sample. On the other hand, statistics of Dixon's

type such as

$$y(91, 92; 1, 91) = \frac{x_{(92)} - x_{(91)}}{x_{(91)} - x_{(1)}},$$

discussed above in the normal sample situation, are clearly applicable to gamma samples.

General considerations of this kind produce a wide choice of possible test statistics. We must now ask which test is 'best' for any particular situation, how can it be constructed, and how should its performance be assessed?

### 3.1 STATISTICAL BASES FOR CONSTRUCTION OF TESTS

Apart from intuitively based procedures, two widely applicable methods exist for setting up discordancy tests, as has been said in Section 2.5. These are the *maximum likelihood ratio principle* and the *principle of local optimality* perhaps restricted to the classes of unbiased, or invariant, tests. Naturally the construction of the tests depends in the first instance on the alternative hypothesis employed to account for the outliers (Section 2.3).

Consider for example the testing of a single upper outlier  $x_{(n)}$  in an exponential sample. Our working hypothesis is

$$H: F,$$

declaring that all the observations  $x_1, \dots, x_n$  belong to the distribution  $F$  with density  $\theta e^{-\theta x}$  ( $x > 0$ ),  $\theta$  being unknown. Suppose we have a slippage alternative  $\bar{H}$  stating that  $n-1$  of the observations belong to  $F$  and the remaining one,  $x_n$  say, to the exponential distribution  $G$  with density  $\lambda \theta e^{-\lambda \theta x}$  ( $x > 0$ ;  $\lambda < 1$ ). We may write

$$H: \lambda = 1$$

$$\bar{H}: \lambda < 1.$$

The log likelihood of the observations on hypothesis  $H$  is

$$L_H(\theta) = n \ln \theta - n\theta \bar{x} \quad (3.1.1)$$

where  $\bar{x}$  is the mean of  $x_1, \dots, x_n$ .  $L_H(\theta)$  is maximized by  $\theta = 1/\bar{x}$ , and its maximized value is

$$\hat{L}_H = -n \ln \bar{x} - n. \quad (3.1.2)$$

On hypothesis  $\bar{H}$ , the log likelihood of the observations is

$$L_{\bar{H}}(\theta, \lambda) = n \ln \theta + \ln \lambda - (n-1)\theta \bar{x}' - \lambda \theta x_n \quad (3.1.3)$$

where  $\bar{x}'$  is the mean of  $x_1, \dots, x_{n-1}$ .  $L_{\bar{H}}(\theta, \lambda)$  is maximized when

$$n/\theta - (n-1)\bar{x}' - \lambda x_n = 0$$

and

$$1/\lambda - \theta x_n = 0,$$

provided  $\lambda \leq 1$ , i.e. when  $\theta = 1/\bar{x}'$  and  $\lambda = \bar{x}'/x_n$ , provided  $x_n \geq \bar{x}'$ . Its maximized value is accordingly (3.1.2) if  $x_n < \bar{x}'$ , otherwise it is

$$\hat{L}_{\bar{H}} = -(n-1)\ln \bar{x}' - \ln x_n - n. \quad (3.1.4)$$

The test statistic based on the maximum likelihood ratio is  $\{\hat{L}_{\bar{H}} - \hat{L}_H\}$ . This is equal to zero if  $x_n < \bar{x}'$ , while if  $x_n \geq \bar{x}'$  it is

$$\begin{aligned} & -\{(n-1)\ln \bar{x}' + \ln x_n - n \ln \bar{x}\} \\ & = -(n-1)\ln \frac{n-T}{n-1} - \ln T, \quad \text{where } T = x_n/\bar{x}. \end{aligned} \quad (3.1.5)$$

It follows that the maximum likelihood ratio test is equivalent to rejecting  $H$  when  $T$  is large.

Strictly speaking, we are not in a position to use this test, because we do not know which of the observations is the discordant one belonging to  $G$  if  $\bar{H}$  is true. In practice this observation is assumed to be  $x_{(n)}$ , the outlier, and  $T_{(n)} = x_{(n)}/\bar{x}$  is used as test statistic. If it were not for the presence of the outlier we would not be moved to query  $H$  or to test the strength of the evidence for  $\bar{H}$ . This is an intuitive justification for the use of  $T_{(n)}$ , but clearly  $T_{(n)}$  is not the maximum likelihood ratio test statistic for the  $\bar{H}$  we have specified.

There are two ways in which we could legitimately establish  $T_{(n)}$  as the appropriate test statistic. The first faces up squarely to the fact that our desire for a test of discordancy stems from our reaction to one specific observation; namely, the greatest observation,  $x_{(n)}$ . An alternative hypothesis can be set up which reflects this, in the form

$$\begin{aligned} \bar{H}: & x_{(1)}, x_{(2)}, \dots, x_{(n-1)} \text{ belong to } F \\ & x_{(n)} \text{ belongs to } G. \end{aligned}$$

This hypothesis, which we may call the hypothesis of *labelled slippage*, identifies the *extreme* observation as the only possible discordant value.

If  $y_1, \dots, y_{n-1}$  is a random sample from  $F$  and  $y_n$  is a random observation from  $G$ , we can think of our ordered sample  $x_{(1)}, \dots, x_{(n)}$  as a particular realization  $y_1, y_2, \dots, y_n$  in which the observation  $y_n$  turns out to be the largest. Thus the likelihood under  $\bar{H}'$  is

$$P(x_{(1)}, \dots, x_{(n)})$$

where  $P(y_1, \dots, y_n)$  is the likelihood of  $y_1, \dots, y_n$  conditional on  $y_1 < \dots < y_{n-1} < y_n$ . Now each  $y_j$  ( $j = 1, \dots, n-1$ ) may be regarded as the time to the first event in a Poisson process of rate  $\theta$ , and  $y_n$  as the time to the first event in a Poisson process of rate  $\lambda\theta$ ; by superposing these  $n$  processes, assuming them independent, and considering which event occurs first, the probability that  $y_1$  is the smallest of the  $y$ 's is seen to be  $\theta/[(n-1)\theta + \lambda\theta] = 1/(n-1+\lambda)$ .

Continuing stepwise, we get

$$P(y_1 < \dots < y_n) = 1/[(n-1+\lambda)(n-2+\lambda)\dots(1+\lambda)]. \quad (3.1.6)$$

Hence the log likelihood of the observations on  $\bar{H}'$  is

$$L_{\bar{H}'}(\theta, \lambda) = n \ln \theta + \ln \lambda - \theta(n\bar{x} - x_{(n)}) - \lambda x_{(n)} + \sum_{j=1}^{n-1} \ln(j+\lambda). \quad (3.1.7)$$

This is maximized when

$$n/\theta - n\bar{x} + x_{(n)} - \lambda x_{(n)} = 0$$

and

$$1/\lambda - \theta x_{(n)} + \sum_{j=1}^{n-1} (j+\lambda)^{-1} = 0.$$

The maximizing value of  $\lambda$ ,  $\hat{\lambda}$  say, must therefore satisfy

$$\sum_{j=0}^{n-1} (j+\hat{\lambda})^{-1} = nT_{(n)} / [n - (1-\hat{\lambda})T_{(n)}]. \quad (3.1.8)$$

The maximized value of  $L_{\bar{H}'}(\theta, \lambda)$  comes out to be

$$\hat{L}_{\bar{H}'} = -n \ln \bar{x} - n - n \ln[n - (1-\hat{\lambda})T_{(n)}] + \sum_{j=0}^{n-1} \ln(j+\hat{\lambda}). \quad (3.1.9)$$

Under  $H$ , the log likelihood is

$$n \ln \theta - n\theta\bar{x} - \ln n!$$

with maximized value

$$\hat{L}_H = -n \ln \bar{x} - n - \ln n!$$

Hence  $\hat{L}_H - \hat{L}_{\bar{H}'}$  depends on the observations in terms of  $T_{(n)}$  and  $\hat{\lambda}$  only, and is therefore a function of  $T_{(n)}$  since in view of (3.1.8)  $\hat{\lambda}$  is a function of  $T_{(n)}$  only. The discordancy test statistic  $T_{(n)}$  is thus equivalent to the maximum likelihood ratio test statistic in the labelled slippage formulation.

The second way in which  $T_{(n)}$  is established directly as the appropriate test statistic is by a multiple decision procedure applied to a set of  $n$  alternative hypotheses

$$\begin{aligned} \bar{H}_i : x_i &\text{ comes from } G \text{ (some } i\text{)} \\ x_j &\text{ comes from } F \text{ (} j \neq i \text{)} \end{aligned} \quad (3.1.10)$$

for  $i = 1, 2, \dots, n$ . This formulation is similar to the *model B* type of slippage alternative hypothesis considered by Ferguson (1961a) (see Section 2.3 iv) specialized to the case of a single outlier and an exponential distribution. We noted in Section 2.5 Paulson's (1952b) use of a multiple decision approach to the corresponding set of alternative hypotheses for a location shifted outlier in a normal sample. The decision criterion is that of

maximizing the probability of adopting the correct  $\bar{H}_i$  when slippage has occurred, subject to a prescribed probability of correct adoption of the basic hypothesis  $H$  and to certain invariance conditions (index permutation, positive changes of scale, arbitrary changes of origin). In the present situation of an exponential basic model, changes in location are inappropriate and the procedure leads to adopting  $\bar{H}_i$  if  $T_i$  is maximized when  $j = i$  and is sufficiently large. Thus the appropriate test statistic is precisely  $T_{(n)}$ . Ferguson (1961a) applies the same type of multiple decision argument to the case of testing for a *model B* (variance-covariance slippage) type outlier in a multivariate normal sample.

Another way of handling (3.1.10) would be by means of a *two-stage maximum likelihood ratio* test: declaring as a discordant outlier that observation whose omission effects the greatest increase in maximized likelihood, provided that increase is significantly large. Thus we consider (see 3.1.5)

$$\hat{L}_{\bar{H}_i} - \hat{L}_H = \begin{cases} -(n-1)\ln\left(\frac{n-T_i}{n-1}\right) - \ln T_i & (T_i \geq 1), \\ 0 & (T_i \leq 1). \end{cases} \quad (3.1.11)$$

Choosing  $i$  to maximize (3.1.11) implies identifying the hypothesis  $\bar{H}_i$  for which  $x_i = x_{(n)}$ ; this  $\bar{H}_i$  is adopted and  $x_{(n)}$  declared a discordant outlier if  $T_{(n)}$  is sufficiently large.

Still considering the case of a single upper outlier in an exponential sample, let us now construct a discordancy test on the basis of *one-sided local (invariant) optimality*. Assuming first the slippage alternative  $\bar{H}$ , with log likelihood

$$L_{\bar{H}}(\theta, \lambda) = n \ln \theta + \ln \lambda - (n-1)\theta\bar{x}' - \lambda\theta x_n,$$

we have  $\partial L_{\bar{H}}(\theta, \lambda)/\partial \lambda = (1/\lambda) - \theta x_n$ .

Under the working hypothesis  $H$ ,  $\lambda = 1$ , and  $\partial L_{\bar{H}}/\partial \lambda$  is then equal to  $1 - \theta x_n$ . Replacing  $\theta$  by its maximizing value  $\hat{\theta} = 1/\bar{x}'$ , we obtain  $1 - (x_n/\bar{x}')$  as the locally optimal test statistic, which is equivalent to  $T = x_n/\bar{x}$ . If instead we take the labelled slippage alternative  $\bar{H}'$ , we get

$$\partial L_{\bar{H}'}(\theta, \lambda)/\partial \lambda = (1/\lambda) - \theta x_{(n)} + \sum_{j=1}^{n-1} (j+\lambda)^{-1}.$$

When  $\lambda = 1$  this is equal to  $\sum_{j=1}^n j^{-1} - \theta x_{(n)}$ ; substituting  $\theta = \hat{\theta} = 1/\bar{x}$  we get  $\sum_{j=1}^n j^{-1} - \frac{x_{(n)}}{\bar{x}}$  as the locally optimal test statistic, or in effect  $T_{(n)}$ .

Consider now the testing of a single upper outlier in a *normal* sample, for which the statistic  $T_{(n)} = \{x_{(n)} - \bar{x}\}/s$  is commonly used. As in the exponential case discussed above, it can be shown that  $T_{(n)}$  is effectively the maximum likelihood ratio test statistic for a labelled slippage alternative, also for the corresponding multiple decision formulation, and the unidentifiable equivalent  $T = \{x_n - \bar{x}\}/s$  is the corresponding statistic for the ordinary slippage

alternative. The alternative hypotheses must be appropriately chosen; that  $x_1, \dots, x_{n-1}$  (in the case of  $\bar{H}$ ) or  $x_{(1)}, \dots, x_{(n-1)}$  (in the cases of  $\bar{H}'$  or  $\bar{H}_n$ ) belong to  $F: N(\mu, \sigma^2)$  and that  $x_n$  or  $x_{(n)}$  belongs to a normal distribution  $G: N(\mu + a, \sigma^2)$  with a different mean  $\mu + a$  ( $a > 0$ ) but the same variance. If instead we make the alternative a normal distribution  $N(\mu, b\sigma^2)$  ( $b > 1$ ), i.e. with the *same* mean as  $F$  but a *larger* variance, a different and less tractable criterion emerges. For an account of the construction of such tests of discordancy for normal samples based on the local optimality principle, see Ferguson (1961a). His tests, based variously on sample skewness and kurtosis, have been referred to in Section 2.5 and are described in detail below (Section 3.4.3).

Test statistics obtained by either of the methods discussed above have the required invariance properties. They are, in the first place, invariant under permutation of the subscripts of  $x_1, \dots, x_n$  and hence can be expressed as functions of the ordered values  $x_{(1)}, \dots, x_{(n)}$ . (Any symmetric functions of the  $x_{(i)}$  which figure here can of course be written as the corresponding functions of the unordered  $x_i$ .) Ferguson sets out to ensure scale and location invariance in his test criteria and asserts (for the normal case with unknown mean and variance which he is examining) that this implies that the observations will appear in the expressions for the criteria purely in terms of ratios of intervals between ordered values, of the type  $\{x_{(a)} - x_{(b)}\}/\{x_{(c)} - x_{(d)}\}$ . For example, the familiar  $\{x_{(n)} - \bar{x}\}/s$  is a function of these ratios. (Here we have a justification for the use of Dixon statistics which are, of course, the ratios themselves.) Scale and location invariance will also hold in more specific circumstances, where certain parameter values are assumed known, although the condition that the test statistics are functions merely of ratios of differences between ordered values will not necessarily apply. For example, consider the testing of normal samples when the variance  $\sigma^2$  is known. In this situation  $\{x_{(n)} - \bar{x}\}/\sigma$ , for instance, is scale- and location-invariant, and is a valid discordancy statistic. Similarly if  $\mu$  but not  $\sigma^2$  is known,  $\{x_{(n)} - \mu\}/s$ , again not expressible in terms of Dixon statistics, is a valid statistic. In some contexts dual invariance will not be appropriate, as we have remarked above in relation to exponential samples.

Reverting to the maximum likelihood ratio principle, this has a further application of great importance in the *detection* of outliers, quite apart from the assessment of their discordancy. In data situations such as regression or the results of designed experiments, outliers may be present, essentially as outlying *residuals*, but such values may not in general be immediately recognized as outliers in the same way that extreme values are in univariate samples. One way of defining the ‘most outlying’ observation or point in a sample is as the one whose omission produces the greatest increase in the maximized likelihood; this amounts, of course, to identifying it as the one out of the  $n$  whose maximum likelihood ratio test statistic has the greatest value. The same principle can be used to detect the most outlying point in a

multivariate sample. Once detected, it can of course be *tested* for discordancy (see Chapters 6 and 7). This principle is embodied in the use of the multiple decision formulation of the alternative hypothesis.

### 3.1.1 Inclusive and exclusive measures, and a recursive algorithm for the null distribution of a test statistic

In Section 2.4 we have listed six basic types of discordancy test statistic. Most of these are ratios of the form  $N/D$ , as discussed in the introduction to this chapter: here  $N$  is a measure of the separation of the outlying value or values from the main mass of the sample and  $D$  is a measure of the spread of the sample. Measures used for  $N$  include the excess, the deviation and, in the gamma case, the extreme (as deviation from zero), all as defined in Section 2.4; measures used for  $D$  include the standard deviation, the range, the sum of squares of observations corrected to the sample mean, and, again in the gamma case, the sample mean or sum. One ambiguity immediately presents itself: should the means and measures of spread which enter into these statistics be calculated from the complete data set including the outlier (or outliers), or from the reduced data set excluding the outlier? We would appear to have a double set of statistics, based respectively on *inclusive measures* and *exclusive measures*, and to be faced with a decision as to which is preferable. It turns out, however, that a statistic  $N/D$  based on inclusive measures and its analogue based on exclusive measures are in many cases equivalent. The following examples will make this clear.

#### (i) Excess/Range statistic for testing single upper outlier

$$\text{Inclusive} \quad T = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \quad (3.1.12)$$

$$\text{Exclusive} \quad T' = \frac{x_{(n)} - x_{(n-1)}}{x_{(n-1)} - x_{(1)}} \quad (3.1.13)$$

Clearly

$$\frac{1}{T} - \frac{1}{T'} \equiv 1.$$

#### (ii) Deviation/Spread statistic for testing single lower outlier

$$\text{Inclusive} \quad T = \frac{\bar{x} - x_{(1)}}{s}$$

$$\text{Exclusive} \quad T' = \frac{\bar{x}' - x_{(1)}}{s'}$$

where

$$\begin{aligned} n\bar{x} &= \sum_{i=1}^n x_{(i)}, & (n-1)\bar{x}' &= \sum_{i=2}^n x_{(i)}, \\ (n-1)s^2 &= \sum_{i=1}^n (x_{(i)} - \bar{x})^2, & (n-2)s'^2 &= \sum_{i=2}^n (x_{(i)} - \bar{x}')^2. \end{aligned}$$

We have  $nsT = (n-1)s'T'$  and

$$\begin{aligned} (n-1)s^2 - (n-2)s'^2 &= (\bar{x} - x_{(1)})^2 + (n-1)(\bar{x}' - \bar{x})^2 \\ &= s^2 T^2 + \frac{1}{n-1} s^2 T'^2, \end{aligned}$$

whence

$$\frac{(n-1)^2}{nT^2} - \frac{n(n-2)}{(n-1)T'^2} \equiv 1. \quad (3.1.14)$$

When the inclusive and exclusive forms of a discordancy statistic  $N/D$  are functionally related, as above, the null distribution can be obtained by means of a very useful recursive argument. While it is not our intention in this book to give detailed proofs of all results used, we feel it is worthwhile setting out this recursive argument in some detail for a simple particular case, namely that of an upper outlier in an exponential sample. It will then be sufficient to state other results obtainable by the same type of argument as they arise, without giving the argument in detail.

#### *The distribution of the ratio of the greatest observation to the sum of the observations for an exponential sample*

In the introduction to this chapter we remarked on the usefulness of  $x_{(n)}/\sum x_i$  (or some simple function of it) for judging a single upper outlier in a gamma sample. For its detailed application we need to know its distribution in the null case, i.e. when there is *no discordant value* present. Let us consider this specifically for the exponential distribution,  $\Gamma(1, \lambda)$ .

Suppose  $X_1, \dots, X_n$  are  $n$  independent identically distributed (i.i.d.) random variables each exponentially distributed with probability density function

$$f(x) = \lambda e^{-\lambda x} \quad (x > 0). \quad (3.1.15)$$

Write

$$A_X = \sum_{j=1}^n X_j, \quad T_j = X_j/A_X \quad (j = 1, \dots, n).$$

Then  $\mathbf{T} = (T_1, \dots, T_n)'$  is a vector random variable of dimension  $n-1$ , since

$$\sum T_j = 1;$$

and  $A_X, \mathbf{T}$  are statistically independent.

When  $X_n$  is omitted,  $X_1, \dots, X_{n-1}$  are  $n-1$  i.i.d. random variables with the distribution (3.1.15) and we may rename these as  $Y_j$  ( $j = 1, \dots, n-1$ ), with

$$A_Y = \sum_{j=1}^{n-1} Y_j.$$

Also write  $T_j^* = Y_j/A_Y$  ( $j = 1, \dots, n-1$ ), and  $\mathbf{T}^* = (T_1^*, \dots, T_{n-1}^*)'$ , of dimension  $n-2$ . Clearly  $X_n, A_Y, \mathbf{T}^*$  are independent, hence  $X_n/A_Y$  and  $\mathbf{T}^*$  are independent.

Corresponding with  $X_n/A_X = T_n$ , write  $X_n/A_Y = T'_n$ . Clearly

$$\frac{1}{T_n} - \frac{1}{T'_n} = 1. \quad (3.1.16)$$

Hence when  $X_n/A_X = t$ , it follows that

$$X_n/A_Y = t/(1-t).$$

Suppose now that  $g_n(t)$  is the probability density function, and  $G_n(t)$  the distribution function, of the random variable  $X_{(n)}/A_X$ , where  $X_{(n)}$  is the greatest of the  $X_i$ . Then

$$\begin{aligned} g_n(t) \delta t &= P[X_{(n)}/A_X \in (t, t + \delta t)] \\ &= nP[X_n/A_X \in (t, t + \delta t), X_n = X_{(n)}] \\ &= nP[X_n/A_X \in (t, t + \delta t), X_1 < X_n, \dots, X_{n-1} < X_n] \\ &= nP[X_n/A_X \in (t, t + \delta t), X_j/A_Y < t/(1-t) \text{ for } j = 1, \dots, n-1] \\ &= nP[T_n \in (t, t + \delta t), T_j^* < t/(1-t) \text{ for } j = 1, \dots, n-1]. \end{aligned}$$

From (3.1.16),  $T_n$  and  $\mathbf{T}^*$  are independent. Hence

$$g_n(t) \delta t = nP[T_n \in (t, t + \delta t)]P[T_j^* < t/(1-t) \text{ for } j = 1, 2, \dots, n-1].$$

The second of these probabilities is

$$P\left(\max_j T_j^* < \frac{1}{1-t}\right) = G_{n-1}\left(\frac{1}{1-t}\right).$$

The first probability is equal to

$$P\left[T'_n \in (t', t' + \frac{dt'}{dt} \delta t)\right]$$

where  $t' = t/(1-t)$ , so that we conclude

$$g_n(t) \delta t = nP\left[T'_n \in \left(t', t' + \frac{dt'}{dt} \delta t\right)\right]G_{n-1}\left(\frac{t}{1-t}\right). \quad (3.1.17)$$

Since, from (3.1.15),  $(n-1)T'_n$  has the  $F$ -distribution on 2 and  $2(n-1)$  degrees of freedom, we have in the exponential case the recurrence relationship

$$g_n(t) = n(n-1)(1-t)^{n-2} G_{n-1}\left(\frac{t}{1-t}\right) \quad (3.1.18)$$

which gives the null distribution of the outlier statistic  $X_{(n)}/A_X$  for a sample of size  $n$ , in terms of the corresponding distribution for a sample of size  $n-1$ .

The range of possible values of  $X_{(n)}/A_X$  is from  $1/n$  to 1, hence  $G_n(t)=1$  for  $t \geq 1$ . The following recursive calculation arises from (3.1.18).

Range for $(1-t)^{-1}$	$G_{n-1}\left(\frac{t}{1-t}\right)$	Range for $t$	$g_n(t)$	$G_n(t)$
$[1, \infty]$	1	$[\frac{1}{2}, 1]$	$n(n-1)(1-t)^{n-2}$	$1 - n(1-t)^{n-1}$
$[\frac{1}{2}, 1]$	$1 - (n-1)\left(\frac{1-2t}{1-t}\right)^{n-2}$	$[\frac{1}{3}, \frac{1}{2}]$	$n(n-1)(1-t)^{n-2}$	$1 - n(1-t)^{n-1}$
			$- n(n-1)^2(1-2t)^{n-2}$	$+ \frac{n(n-1)}{2!}(1-2t)^{n-1}$
$[\frac{1}{3}, \frac{1}{2}]$	$1 - (n-1)\left(\frac{1-2t}{1-t}\right)^{n-2}$ $+ \frac{(n-1)(n-2)}{2!}$ $\times \left(\frac{1-3t}{1-t}\right)^{n-2}$	$[\frac{1}{4}, \frac{1}{3}]$	$n(n-1)(1-t)^{n-2}$ $- n(n-1)^2(1-2t)^{n-2}$ $+ \frac{n(n-1)^2(n-2)}{2}(1-3t)^{n-2}$	$1 - n(1-t)^{n-1}$ $+ \frac{n(n-1)}{2!}(1-2t)^{n-1}$ $- \frac{n(n-1)(n-2)}{3!}(1-3t)^{n-1}$

and so on. The density function consists of a succession of smoothly connected arcs in the intervals  $[\frac{1}{2}, 1], [\frac{1}{3}, \frac{1}{2}], [\frac{1}{4}, \frac{1}{3}], \dots, [\frac{1}{n-1}, \frac{1}{n}]$ . This well known result was first given by Fisher (1929).

There is, of course, nothing in the method of derivation of (3.1.18) which is specific to the exponential distribution and it can be applied in other circumstances. Additionally, it is easily modified for handling, say,  $X_{(1)}/A_X$ , which will also be of interest.

### 3.2 PERFORMANCE CRITERIA OF TESTS

We raised in Section 2.5 the question of what constitutes an appropriate performance criterion for a test of discordancy. A key measure is the significance level, although its interpretation is to some extent problematical (see Collett and Lewis, 1976). Comparison of tests of the same significance level must of course depend on the alternative hypothesis we have in mind for explaining the outliers (Section 2.3).

Consider first the *slippage alternative*. To fix ideas, suppose we are testing an upper outlier  $x_{(n)}$  in a univariate sample  $x_1, \dots, x_n$ . The null hypothesis is

$$H: F,$$

i.e. all the observations arise from a distribution  $F$  which is, say,  $N(\mu, \sigma^2)$  with  $\mu, \sigma^2$  unknown. We envisage a slippage alternative  $\bar{H}$  which states that  $n-1$  of the observations belong to  $F$  and the  $n$ th observation  $x_n$ , which we will now rename  $x_c$  and call the *contaminant*, belongs to a different distribution  $G$ . If  $F$  is  $N(\mu, \sigma^2)$ ,  $G$  may be  $N(\mu + \sigma\Delta, \sigma^2)$  (for slippage in location) or  $N(\mu, \sigma^2 \exp \Delta)$  (for slippage in dispersion); the hypotheses can then be written

$$H: \Delta = 0$$

versus

$$\bar{H}: \Delta > 0.$$

We wish to test  $x_{(n)}$  for discordancy. For the moment, let us distinguish between two kinds of test statistic, ‘general’ and ‘specific’ say. (We will not need to maintain this distinction for long.) To construct a *general* statistic  $Z_{(n)}$ , we start with a measure  $Z_i$  of the positioning of any observation  $x_i$  in relation to the rest of the sample; for example,  $Z_i$  could be

$$(x_i - \bar{x})/s \quad \text{or} \quad (x_i - \bar{x})/(x_{(n)} - x_{(1)}).$$

By particularizing to  $x_{(n)}$ , we get the corresponding discordancy statistic  $Z_{(n)}$ , e.g.

$$Z_{(n)} = (x_{(n)} - \bar{x})/s \quad \text{or} \quad (x_{(n)} - \bar{x})/(x_{(n)} - x_{(1)}).$$

A *specific* statistic, on the other hand, is sensibly defined only in relation to the outlier  $x_{(n)}$ , and cannot be meaningfully embedded in some set of statistics of like form ranging over the  $n$  sample members; e.g.

$$Z = (x_{(n)} - x_{(n-1)})/s.$$

Suppose our discordancy statistic is of the general type. The test takes the form: ‘Adjudge  $x_{(n)}$  discordant if  $Z_{(n)} > z_\alpha$ , where  $z_\alpha$  is the critical value for preassigned significance level  $\alpha$  defined by

$$P(Z_{(n)} > z_\alpha \mid H) = \alpha. \quad (3.2.1)$$

Since we assume a slippage alternative, one of the sample observations under  $\bar{H}$  will be the contaminant  $x_c$ ; and since  $Z_{(n)}$  is a general statistic, a corresponding measure  $Z_c$  exists for the contaminant. In the context of this particular set of assumptions, David (1970) suggested the following five probabilities as ‘reasonable measures’ of the performance of  $Z_{(n)}$ .

$$(i) \quad P_1 = P(Z_{(n)} > z_\alpha \mid \bar{H}). \quad (3.2.2)$$

This is the probability under  $\bar{H}$  that the outlier is identified as discordant, in other words the power function.

$$(ii) \quad P_2 = P(Z_c > z_\alpha \mid \bar{H}). \quad (3.2.3)$$

$$(iii) \quad P_3 = P(Z_c = Z_{(n)}, Z_{(n)} > z_\alpha \mid \bar{H}). \quad (3.2.4)$$

This is the probability that the contaminant is the outlier and is identified as discordant.

$$(iv) \quad P_4 = P(Z_c = Z_{(n)} > z_\alpha, Z_{(n-1)} < z_\alpha \mid \bar{H}). \quad (3.2.5)$$

$$(v) \quad P_5 = P(Z_c > z_\alpha \mid Z_c = Z_{(n)}; \bar{H}). \quad (3.2.6)$$

This is the probability that, when the contaminant is the outlier, it is identified as discordant.

David observes that:

$P_1$  measures the probability of significance [i.e. adjudged discordancy] for any reason whatever and is thus especially suitable for sounding a general alarm . . . .  $P_2$ ,  $P_3$  and  $P_4$  focus with increasing severity on the correct detection of the outlier . . . ; only  $P_4$  specifically excludes the possibility that good observations might be significant [adjudged discordant] in addition to . . . [the contaminant]. We see that

$$P_1 \geq P_2 \geq P_3 \geq P_4.$$

In point of fact  $P_1$ ,  $P_3$ , and  $P_5$  are useful measures, but the information conveyed by  $P_2$  and  $P_4$  would seem to be rather limited. Suppose for example that the contaminant  $x_c$  is the second greatest observation,  $x_{(n-1)}$ . For a ‘good’ test we want a high probability of identifying  $x_c$  as discordant but this should appropriately be done by reference to the null distribution of  $Z_{(n-1)}$  and not of  $Z_{(n)}$ , so that  $z_\alpha$  is not the appropriate critical value and  $P_2$  not the appropriate measure. Similar considerations apply to  $P_4$ , based as it is on the inequality  $Z_{(n-1)} < z_\alpha$ . In fact, David (1970) labels  $P_2$ ,  $P_4$  respectively as the ‘probability that  $X_1$  [i.e.  $x_c$  in our notation] is significantly large’ and the ‘probability that only  $X_1$  [i.e.  $x_c$ ] is significant’. While these definitions appear attractive, the concept of ‘significance’ on which they rest is ill-defined.

We can therefore discard  $P_2$  and  $P_4$  as performance criteria. Once we do this, the need to distinguish between general and specific tests for discordancy disappears, since  $P_1$ ,  $P_3$ , and  $P_5$  do not depend for their definition on the test being of one or the other type. (For this generalization the event ‘ $Z_c = Z_{(n)}$ ’ in the definitions of  $P_3$  and  $P_5$  needs to be rewritten as ‘ $x_c = x_{(n)}$ ’.)

$P_1$ ,  $P_3$ , and  $P_5$  may be considered to contain between them all the relevant information about the performance of a discordancy test against a slippage alternative.  $P_1$  is a convenient general measure, for the reason indicated by David (1970) (see above). There are strong arguments, however, for preferring  $P_3$  as a measure. Dixon (1950), discussing the assessment of a number of

discordancy criteria for outliers in normal samples, says:

The performance of the . . . criteria is measured by computing the proportion of the time the contaminating distribution provides an extreme value and the test discovers the value [i.e.  $P_3$ ]. Of course, performance could be measured by the proportion of the time the test gives a significant value when a member of the contaminating population is present in the sample, even though not at an extreme [i.e.  $P_1$ ]. However, since it is assumed that discovery of an outlier will frequently be followed by the rejection of an extreme we shall consider discovery a success only when the extreme value is from the contaminating distribution.

For a 'good' test, we require  $P_3$  and  $P_5$  to be high. We also want  $P_1 - P_3$  to be low; this is the probability that the test wrongly identifies a good observation as discordant.  $P_3/P_5$  is the probability that the contaminant shows up as the outlier, and it might appear that one would like this ratio to be as large as possible, in conflict with the requirement for a high value of  $P_5$ . Consider, however, two hypothetical tests with performance measures as follows:

$$\text{test } A: \quad P_3 = \frac{1}{2}, \quad P_5 = 1, \quad P_3/P_5 = \frac{1}{2}.$$

$$\text{test } B: \quad P_3 = \frac{1}{2}, \quad P_5 = \frac{1}{2}, \quad P_3/P_5 = 1.$$

In test *A*, the contaminant has only 50 per cent chance of showing up as the extreme value, indicating that the degree of contamination is not severe on average (the value of  $\Delta$  in  $\bar{H}$  is not large); however, when the contaminant does appear as the outlier, it is certain to be detected. In test *B*, on the other hand, contamination is more severe and the contaminant is always the extreme value. But it is only detected as discordant 50 per cent of times. What we require of a test is that it should identify contamination when this is sufficiently manifest, so *test A* is preferable to *test B*, and a high value of  $P_5$  is desirable rather than a high value of  $P_3/P_5$ . The same line of argument indicates that  $P_5$  takes precedence over  $P_3$  as a measure of test performance. To sum up, a good test may be characterized by high  $P_5$ , high  $P_1$ , and low  $P_1 - P_3$ .

So far we have been talking in terms of a slippage alternative. Consider now the assessment of performance of a discordancy test against a *mixture* alternative. As before, take the particular situation of an upper outlier  $x_{(n)}$  under test in a univariate sample. Under  $\bar{H}$  the number of contaminants in the sample is no longer fixed as in the slippage case, but is a binomially-distributed random number which may be 0, 1 or more. The following events are clearly relevant to the assessment of performance:

*D*: that the test identifies  $x_{(n)}$  as discordant.

*E*: that  $\bar{H}$  holds and the sample contains one or more contaminants.

*F*: that  $\bar{H}$  holds and  $x_{(n)}$  is a contaminant.

The direct analogue of the performance measure  $P_1$  defined in (3.2.2), i.e. the *power*, is  $P(D | \bar{H})$ . However  $P(D | E)$ , which is a power function

conditional on the actual presence of contamination, is a more useful measure than  $P(D | \bar{H})$ .

Analogously to the measures  $P_3, P_5$  of (3.2.4) and (3.2.6) we can define measures  $P(F \cap D | E)$  and  $P(D | F)$  (which is of course the same as  $P(D | F \cap E)$ ).

Characteristics of a good test are, by the previous line of argument, a high value of  $P(D | F)$ , a high value of  $P(D | E)$ , and a low value of  $P(D | \bar{H}) - P(D \cap F | \bar{H})$ , the probability that the test identifies as discordant an observation actually generated by the basic model.

In the case of a discordancy test against an *inherent* alternative, the situation simplifies. There is now no specifiable contaminant observation, and the probabilities  $P_3$  and  $P_5$  (and the events  $E$  and  $F$ ) are undefined. The appropriate measure of performance of the test is the power  $P(D | \bar{H})$ .

### 3.3 THE MULTIPLE OUTLIER PROBLEM

We have discussed the testing of a single outlier for discordancy. New problems of procedure arise when the number of observations which appear aberrant in relation to the main data mass is more than one. We may have, for instance, a normal sample of size  $n$ , with two upper outliers  $x_{(n-1)}$  and  $x_{(n)}$ , both of which are unusually far to the right of the other  $n-2$ , or a normal sample with two lower outliers  $x_{(2)}$  and  $x_{(1)}$  unusually far to the left, or again a sample with a lower and an upper outlier-pair,  $x_{(1)}$  and  $x_{(n)}$ , widely bracketing the main data mass. Again, a normal sample may contain three extreme values which appear to be outlying in relation to the main  $n-3$ , perhaps all three upper, perhaps two upper and one lower, and so on.

Similar situations may arise with gamma samples, though in the particular case of the exponential distribution its *J*-shape makes it likely that all outliers presenting themselves will be upper ones, not lower. Two or more outliers may in fact be encountered in samples from most univariate distributions or from distributions of higher dimensionality; and it is possible to find two or more outlying points in a regression, two or more outlying residuals underlying the observations from a designed experiment, or two or more outlying values in a time-series. In all these *multiple outlier* situations there are  $k (> 1)$  outliers or outlying points in a data set of size  $n$ , and the analyst envisages the possibility of up to  $k$  discordant values. Appropriate tests of discordancy will therefore be required.

For given sample size  $n$  there is an effective upper limit to the number of outliers  $k$ . To take an extreme case, one cannot consider  $k = n - 1$  observations as outliers in relation to the remaining 1! Indeed the concept of the 'main data mass' is hardly meaningful if  $k \geq n/2$ , say. Intuitively, an upper limit of the form  $k_{\max} = Cn^{\alpha}$  suggests itself for  $k$ , where  $C$  is a positive constant (perhaps 1) and  $\alpha$  a constant between 0 and 1. Taking  $\alpha = \frac{1}{2}$ ,  $C = 1$ , for example, we would get  $k_{\max} = \sqrt{n}$ , so that one would not deal with

more than 10 outliers in a sample of 100. Any attempt at quantification of  $k_{\max}$  would depend on a sensible choice of the working and alternative hypotheses in relation to each other, and information regarding the best achievable performance (Section 3.2) by discordancy tests (Section 3.1). Such studies have not so far been undertaken. In published work to date, discordancy tests and other procedures for multiple outliers have been mainly confined to  $k = 2$  and 3 irrespective of  $n$ .

Faced with a multiple outlier situation, there is a basic choice between two types of procedure, which we may call *block procedures* and *consecutive procedures* (sometimes referred to as ‘sequential’ procedures—see e.g. Dixon, 1950, and Section 3.3.2).

Suppose, to fix ideas, that we wish to test for discordancy two upper outliers  $x_{(n-1)}, x_{(n)}$  in an exponential sample of size  $n$ . Take  $x_{(1)}, \dots, x_{(n-2)}$  as belonging to the distribution  $F$  with density  $\theta e^{-\theta x}$  ( $x > 0$ ), and  $x_{(n-1)}$  and  $x_{(n)}$  as belonging to exponential distributions  $G_1, G_2$  with respective densities  $\lambda \theta e^{-\lambda \theta x}, \mu \theta e^{-\mu \theta x}$  ( $x > 0$ ). The working hypothesis is

$$H: \lambda = \mu = 1.$$

If we take as the single alternative hypothesis

$$\bar{H}: \lambda = \mu < 1,$$

we are led to a single discordancy test, as the result of which we either accept both outliers as consistent with the rest of the sample, or adjudge them both discordant. A possible test criterion in this context would be

$$[x_{(n-1)} + x_{(n)}] / \sum_{i=1}^n x_i.$$

This exemplifies a block procedure—we would be testing the multiple outliers *en bloc*.

On the other hand, we could go for a pair of consecutive alternatives to  $H$ , which might typically be as follows:

$$\begin{aligned}\bar{H}' &: \lambda = 1, \mu < 1 \\ \bar{H}'' &: \lambda < 1.\end{aligned}$$

The procedure would be first to test  $H$  against  $\bar{H}'$  using a test for a *single* upper outlier. If  $H$  is accepted, both outliers are declared consistent with the remainder of the sample, and the discordancy test terminates. If  $H$  is rejected,  $\bar{H}''$  is tested against a revised working hypothesis confined to  $x_{(1)}, \dots, x_{(n-1)}$ ,

$$H'': \lambda = 1.$$

Again we would use a test for a single upper outlier. We thus have a *consecutive* procedure, with three possible paths:

Accept  $H \rightarrow$  adjudge neither  $x_{(n)}$  nor  $x_{(n-1)}$  discordant.

Reject  $H$ , accept  $H'' \rightarrow$  adjudge  $x_{(n)}$  discordant, but not  $x_{(n-1)}$ .

Reject  $H$ , reject  $H'' \rightarrow$  adjudge both  $x_{(n)}$  and  $x_{(n-1)}$  discordant.

The above discussion illustrates the choice between block and consecutive procedures. An apparently similar kind of choice is, of course, familiar in the testing of the relevance of a subset of the regressor variables  $x_1, x_2, \dots$  in a regression analysis. In the regression situation the data context sometimes gives a guide as to when a block procedure is appropriate, in preference to the consecutive testing of variables one by one which is the norm. For example, we might be studying the effect of nine factors  $x_1, \dots, x_9$  on the efficiency,  $y$ , of a domestic heating device tested *in situ*. Suppose  $x_1, x_2$  are properties of the fuel used,  $x_3, \dots, x_7$  are different features of the internal construction of the house, and  $x_8, x_9$  are measurements of ambient temperature at ground and roof level outside the house. Then it is reasonable to test the significance of  $x_8$  and  $x_9$  jointly—a block procedure—on the basis that if  $y$  is affected at all by the outside temperature we should clearly take both of the outside temperature measurements into account.

In the usual multiple outlier situation, all the observations have the same status on the working hypothesis, and discordancy tests are only invoked when particular values show up as outliers; thus no guidance is available from the data context as to whether a block procedure should be used. The exception is when we have prior information which leads us to focus on some particular subset of the data—for example, if  $x_{(1)}, x_{(n)}$  were observations made by experimenter A and the other  $n-2$  by experimenter B.

In principle, the choice between a block procedure and a consecutive procedure in a multiple outlier situation depends on the relative performances of the test procedures in relation to an alternative hypothesis  $\bar{H}$ . We have to say ‘in principle’ rather than ‘in practice’, since in most cases no performance criteria have so far been evaluated. Dixon (1950) gives values of performance measure  $P_3$  (Section 3.2) for block tests of two upper (or lower) outliers in a normal sample using the test statistics

$$(i) \frac{\text{sum of squares about mean for } n-2 \text{ observations omitting outliers}}{\text{sum of squares about mean for } n \text{ observations including outliers}}$$

and

$$(ii) (x_{(n)} - x_{(n-2)}) / (x_{(n)} - x_{(1)}).$$

Ferguson (1961a) gives values of the power function  $P_1$  for general multiple outlier discordancy tests based on the sample skewness and the sample kurtosis for an unspecified number of outliers; see Worksheets below.

The first direct quantitative comparisons of block and consecutive procedures appear to be those of McMillan and David (1971) and McMillan (1971). McMillan and David consider a normal sample with known variance, unity say, containing two contaminants from a normal distribution also having unit variance but with mean slipped to the right. They evaluate  $P_3$  (Section 3.2) for a block discordancy test based on the sum of the two largest deviations from the sample mean; and they also consider a consecutive procedure based at each stage on the largest deviation from the mean, evaluating the probabilities that at least one contaminant is identified as discordant and that both contaminants are so identified; see Section 3.3.2 below. McMillan (1971) gives corresponding results in terms of studentized deviations for the case when the underlying variance is unknown. Hawkins (1973) extends McMillan's results and gives values of the power function for the consecutive test in various cases. As Hawkins says in his conclusion:

we [have discussed] the problem of repeated use of a single outlier statistic. The null hypothesis distributions are solved, but much research remains to be done on the alternative hypothesis distributions.

Consecutive procedures have an obvious appeal, but, as has long been recognized, they suffer in the form described above from one inherent limitation. This is the possible effect of *masking*. Suppose, to fix ideas, that two upper outliers  $x_{(n-1)}, x_{(n)}$  are to be tested for discordancy by up to two consecutive applications of a test using a statistic of the form  $N/D$ , where  $N$  is a measure of the separation of the greatest value from the rest of the sample and  $D$  is a measure of the spread of the sample. At the first stage of the test, where we consider  $x_{(n)}$  alone, the  $D$ -value will be large since it involves the outlier  $x_{(n-1)}$ , so that  $x_{(n)}$  may not be adjudged discordant; the procedure terminates without a second test and both  $x_{(n)}$  and  $x_{(n-1)}$  are declared consistent with the remainder of the sample. On the other hand a block test of  $x_{(n)}$  and  $x_{(n-1)}$  as a pair might identify them as highly discordant. This is what would happen for example, with a sample such as 3, 4, 7, 8, 10, 949, 951. In a phrase due to Murphy (1951),  $x_{(n-1)}$  has had a *masking effect* on the identification of  $x_{(n)}$ ; the masking phenomenon was discussed as early as 1936 by Pearson and Chandra Sekar. An interesting alternative danger has recently been described by Fieller (1976), that false conclusions may be drawn owing to an effect which he terms *swamping*. For example, consider the sample 3, 4, 7, 8, 10, 13, 951. A block procedure applied to the upper two values 13 and 951 may well declare them discordant as a pair; the extreme outlier 951 has 'carried' the otherwise unexceptionable value 13.

### 3.3.1 Block procedures for multiple outliers in univariate samples

The considerations governing discordancy tests for single outliers, discussed earlier in this chapter, extend to block-type discordancy tests for multiple

outliers—in the construction of tests, the existence in some cases of inclusive and exclusive measures and of recursive relations for null distributions, and in the setting up of performance criteria.

As in the single outlier case (Section 3.1), appealing test statistics can be set up on an intuitive basis; and ‘best’ tests can be constructed on the maximum likelihood ratio principle or again on the principle of local optimality. Consider again the testing of a pair of upper outliers  $x_{(n-1)}, x_{(n)}$  in an exponential sample. On intuitive grounds—maybe by generalization from the case of a single upper outlier—one could propose

$$T_{(n-1,n)} = \{x_{(n-1)} + x_{(n)}\}/2\bar{x} \quad (3.3.1)$$

as a sensible statistic. Again, Dixon-type statistics such as  $\{x_{(n)} - x_{(n-2)}\}/\{x_{(n-2)} - x_{(1)}\}$  have a natural appeal. On the other hand, let us see where the maximum likelihood ratio principle leads. Our working hypothesis is

$$H: F$$

declaring that all the observations  $x_1, \dots, x_n$  belong to the distribution  $F$  with density  $\theta e^{-\theta x}$  ( $x > 0$ ),  $\theta$  unknown.

Suppose first that we have a slippage alternative  $\bar{H}$  stating that  $n-2$  of the observations belong to  $F$  and the remaining two,  $x_{n-1}$  and  $x_n$  say, come from the exponential distribution  $G_\lambda$  with density  $\lambda\theta e^{-\lambda\theta x}$  ( $x > 0$ ;  $\lambda < 1$ ). Calculations similar to those of equations (3.1.1) to (3.1.5) lead to

$$T = (x_{n-1} + x_n)/2\bar{x} \quad (3.3.2)$$

as test statistic, providing that  $(x_{n-1} + x_n)/2 \geq \bar{x}''$ , the mean of  $x_1, \dots, x_{n-2}$ . If instead of  $\bar{H}$  we adopt the corresponding *labelled* slippage alternative

$$\bar{H}': x_{(1)}, \dots, x_{(n-2)} \text{ belong to } F$$

$$x_{(n-1)}, x_{(n)} \text{ belong to } G_\lambda,$$

calculations similar to those of equations (3.1.6) to (3.1.9) lead to  $T_{(n-1,n)}$  as test statistic. Alternatively,  $T_{(n-1,n)}$  can be set up on the basis of a multiple decision argument as in (3.1.10).

But  $\bar{H}, \bar{H}'$  are not the only pair of slippage alternatives. Consider instead the slippage alternative  $\bar{H}''$  stating that  $x_1, \dots, x_{n-2}$  belong to  $F$ ,  $x_{n-1}$  belongs to  $G_\lambda$ , and  $x_n$  belongs to the exponential distribution  $G_\mu$  with density  $\mu\theta e^{-\mu\theta x}$  ( $x > 0$ ;  $\mu < \lambda$ ). The maximum likelihood ratio statistic is now not  $T$  but

$$T'' = \left( \frac{x_{n-1}x_n}{\bar{x}^2} \right)^{1/n} \left( n - \frac{x_{n-1} + x_n}{\bar{x}} \right)^{(n-2)/n}. \quad (3.3.3)$$

The multiple decision argument leads to the statistic

$$T''_{(n-1,n)} = \left( \frac{x_{(n-1)}x_{(n)}}{\bar{x}^2} \right)^{1/n} \left( n - \frac{x_{(n-1)} + x_{(n)}}{\bar{x}} \right)^{(n-2)/n}. \quad (3.3.4)$$

On the other hand, the labelled slippage alternative corresponding to  $\bar{H}''$ , viz.

$$\bar{H}''' : x_{(1)}, \dots, x_{(n-2)} \text{ belong to } F$$

$$x_{(n-1)} \text{ belongs to } G_\lambda$$

$$x_{(n)} \text{ belongs to } G_\mu$$

does not now lead to  $T_{(n-1,n)}''$ , but gives a maximum likelihood ratio test statistic which cannot be expressed in closed form.

As regards performance criteria for tests of multiple outliers, the discussion of Section 3.2 carries over to the block test situation, the 'contaminant' now being a contaminant subset of two or more observations.

Published work on block procedures includes Grubbs (1950) and Tietjen and Moore (1972, but see the cautionary remark on Worksheet N4) for normal samples, and Likeš (1966) and Lewis and Fieller (1978) for gamma samples.

### 3.3.2 Consecutive procedures for multiple outliers in univariate samples

The possibility of testing multiple outliers consecutively for discordancy has been mentioned by a number of authors, mainly with reference to the masking effect; see Pearson and Chandra Sekar (1936), Dixon (1953), Ferguson (1961b), David (1970, p. 185), and Tietjen and Moore (1972). However, as regards quantitative discussion of actual tests and their properties, nearly all the references in the literature relate to block procedures. Notable exceptions are the papers by McMillan and David (1971), McMillan (1971), and Hawkins (1973) mentioned above.

Up to now authors have mostly used the word 'sequential' when referring to the successive testing of multiple outliers one at a time. We prefer the word *consecutive*. Sequential testing, in common statistical parlance, implies that the sample size is not fixed but is determined in each realization in relation to the values of the earlier observations. In successive testing of multiple outliers this sequential property applies to the number of times the test is used, not to the sample size, which is *fixed*.

Consecutive procedures present no separate problem of test construction, since they merely involve repeated use of single outlier tests from the available repertoire. As regards performance criteria, however, we are in a fresh situation. The measures described in Section 3.2 will need generalizing since there is, by definition, not one contaminant but several. Suppose for example that the alternative hypothesis  $\bar{H}$  envisages two discordant values, liable to appear as upper outliers, in a sample of  $n$ . The following events defined under  $\bar{H}$  would seem to be relevant:

$E_1$ : that  $x_{(n)}$  is one of the two contaminants

$E_2$ : that  $x_{(n-1)}$  is one of the two contaminants

$E = E_1 \cap E_2$ : that the two contaminants are the two outliers

$D_0$ : that  $x_{(n)}$  is not adjudged discordant (on the first test)

$D_1$ : that  $x_{(n)}$  is adjudged discordant (on the first test) but  $x_{(n-1)}$  is not adjudged discordant (on the second test)

$D_2$ : that  $x_{(n)}, x_{(n-1)}$  are both adjudged discordant (requiring two tests).

'Total' measures corresponding to  $P_1, P_3, P_5$  in Section 3.2 will be respectively  $P(D_2), P(D_2 \cap E), P(D_2 | E)$ . But 'partial' measures are also of interest, such as  $P(D_1 | E)$  and  $P(D_0 | E)$  ( $= 1 - P(D_1 | E) - P(D_2 | E)$ ).

The measures of performance used by McMillan and David (1971) and McMillan (1971) in their pioneering papers on consecutive testing are not in fact any of those we have listed above, but are

$$P(C_1), P(C_2), P(C_3)$$

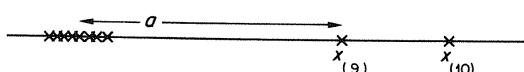
defined as follows. Denoting by  $R$  the discordancy region for the first-stage test based on all  $n$  observations,  $C_1$  is the event that at least one of the two contaminants is in  $R$  (and so adjudged discordant), and  $C_2$  the event that both contaminants are in  $R$ .  $C_3$  is the event that at least one of the two contaminants is in  $R$  and the other is in the discordancy region for a second test based on the reduced sample omitting the first contaminant. The values of  $P(C_1)$  and  $P(C_3)$  should correspond reasonably closely with  $P(D_1 | E) + P(D_2 | E)$  and  $P(D_2 | E)$  respectively, at any rate when the degree of contamination is marked enough to make it virtually certain that the contaminants appear as the outliers. However, in general the use of  $P(C_1)$ ,  $P(C_2)$ , and  $P(C_3)$  seems rather arbitrary.  $P(C_2)$  in particular seems difficult to interpret.

An event which can usefully be defined is

$D_3$ : that  $x_{(n)}$  is not adjudged discordant (on the first test), but  $x_{(n-1)}$  would be adjudged discordant on a hypothetical second test omitting  $x_{(n)}$  from the sample.

$P(D_3)$  is clearly a measure of the masking effect. It is relatively easy to calculate; we give a numerical illustration.

*Example.* Suppose we have a sample of size 10 from a normal distribution with unknown mean and variance, with eight of the values  $x_{(1)}, \dots, x_{(8)}$  grouped together near 0 and  $x_{(9)}, x_{(10)}$  far to the right at a say. Using consecutive testing, with a discordancy statistic of the form  $\frac{x_{(n)} - \bar{x}}{s}$  at the 5 per cent level, it will be impossible to identify  $x_{(10)}$  as discordant, owing to the masking effect of  $x_{(9)}$ , unless it exceeds 1.47a. For otherwise  $1 - P(D_3)$  will exceed 0.05.



### 3.4 DISCORDANCY TESTS FOR PRACTICAL USE

We now present detailed information on a wide range of useful tests. 'Useful' means two things here, first that the test performs reasonably well, even if not optimally, in relation to some meaningful alternative hypothesis; secondly that some information on percentage points is available—at least an inequality, if not an extensive tabulation. The main types of distribution for which useful tests are available are gamma and normal, and these are dealt with in Section 3.4.2, Section 3.4.3 respectively. Some tests for samples from other distributions, including uniform, log-normal, and Poisson, are described in Section 3.4.4.

#### 3.4.1 Guide to use of the tests

In each case (gamma, normal, uniform, etc.) we commence with some general discussion of the types of outlier situation where the distribution might be appropriate. We then present for each distribution a *contents list* of tests (pp. 77–79, 90–93, 116); we have labelled the tests G1, G2, ... for gamma samples, P1, P2, ... for Poisson samples, and so on, as is explained in detail later. For each individual test we then give a *worksheet*, which presents systematically the following information where available:

*Label and purpose of test*

*Test statistic*, denoted by  $T$  with the test label as subscript. For example, we denote the statistic for test N1 by  $T_{N1}$ .

*Test distribution*, i.e. the distribution of the statistic on the working hypothesis that the outlying value or set of values is consistent with the rest of the sample. The probability density function and the distribution function for this distribution are denoted respectively by  $f_n(t)$ ,  $F_n(t)$ , where  $n$  is the sample size.

*Recurrence relationship* for the test distribution (where appropriate).

*Simple inequality for the significance probability* (where appropriate). The significance probability attaching to an observed value  $t$  of a discordancy statistic  $T$  is denoted here by  $SP(t)$ . That is to say,  $SP(t)$  is the probability that, on the working hypothesis,  $T$  takes values more discordant than  $t$  (for most tests this means  $T > t$ ).

*Tabulated significance levels* in the form of references to where these will be found in the set of tables in the Appendix at the back of the book, together with source attribution.

*Further tables*: reference to books and journals extending our tabulated significance levels, or tabulating other quantities of interest, for example, power.

*References* to other published material on the test such as derivation of test distribution, optimality properties, power considerations, etc.

*Properties of test*: advantages and disadvantages, including statement of any secondary features of the data against which the test provides a particular

safeguard, such as a suspicious *least* value when testing for an *upper* outlier; whether the test has a theoretical validation, such as being a maximum likelihood ratio test for some alternative; information on power or other performance measures, if available.

Illustrative *examples* are also given for some of the tests.

The following notation is used for standard distributions, random variables and functions:

### *Notation for distributions or random variables*

$N(\mu, \sigma^2)$	normal with mean $\mu$ and variance $\sigma^2$
$t_\nu$	Student's $t$ with $\nu$ degrees of freedom
$F_{\nu_1, \nu_2}$	variance-ratio (or $F$ ) with $\nu_1$ and $\nu_2$ degrees of freedom
$\Gamma(r, \lambda)$	gamma with scale parameter $\lambda$ and shape parameter $r$ , i.e. with density $f(x) = [\lambda^r \Gamma(r)]^{-1} (x^{r-1}) \exp(-x/\lambda)$ ( $x > 0$ )
$E(\lambda)$	exponential with mean $\lambda$ , i.e. with density $f(x) = \lambda^{-1} \exp(-x/\lambda)$ ( $x > 0$ ), 0 ( $x < 0$ )—same as $\Gamma(1, \lambda)$
$E(\lambda; a)$	exponential with scale parameter $\lambda$ and origin at $a$ , i.e. with density $f(x) = \lambda^{-1} \exp[-(x-a)/\lambda]$ ( $x > a$ ), 0 ( $x < a$ )
$P(\mu)$	Poisson with mean $\mu$
$B(n, p)$	binomial with parameters $n, p$
$H(N; n, r)$	hypergeometric with parameters $N: n, r$

### *Notation for functions*

$\phi(t)$	probability density function of $N(0, 1)$ , i.e. $(2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}t^2)$
$\Phi(t)$	distribution function of $N(0, 1)$ , i.e. $\int_{-\infty}^t \phi(u) du$
$B(r, s)$	beta function with parameters $r$ and $s$ , i.e. $\Gamma(r)\Gamma(s)/\Gamma(r+s)$
$b_{r,s}(t)$	beta density with parameters $r$ and $s$ , i.e. $[B(r, s)]^{-1} t^{r-1} (1-t)^{s-1}$ $(0 \leq t \leq 1)$

### **3.4.2 Discordancy tests for gamma (including exponential) samples**

Until fairly recently, most of the published work on outliers in univariate samples has been in the context of normal distributions. However, problems of outliers in samples from gamma distributions, and in particular from exponential distributions, are of considerable practical importance. Outlier situations in exponential samples arise naturally in such contexts as life testing; outliers in  $\chi^2$  samples arise in analysis of variance; outliers in gamma samples of arbitrary shape parameter arise with skew-distributed data, for which a gamma distribution is often a useful pragmatic model; and outliers in both gamma and specifically exponential samples arise in any contexts where Poisson processes are appropriate basic models, e.g. in studying traffic flow, failures of electronic equipment, biological aggregation,

or even deaths from horse kicks. Attention to such problems in the literature has developed in recent years (Epstein, 1960a, 1960b; Laurent, 1963; Basu, 1965; Likeš, 1966; Kabe, 1970; Kale and Sinha, 1971; Joshi, 1972b; Sinha, 1972, 1973a, 1973b, 1973c; Veale and Kale, 1972; Mount and Kale, 1973; Kale, 1974a, 1975c; Lewis and Fieller, 1978).

Further applications arise through transformation. Procedures for outliers in exponential samples can sometimes be applied to outliers in samples from other distributions, such as the extreme-value distribution and the Weibull, by transforming the observations. For example, if the  $n$  values  $x_1, \dots, x_n$  are (on the working hypothesis) a sample from the extreme-value distribution with distribution function  $P(X \leq x) = \exp\{-\exp[-(x-a)/b]\}$ , then the  $n$  transformed values  $\exp(-x_1/b), \dots, \exp(-x_n/b)$  are a sample from the exponential distribution with mean  $\exp(-a/b)$ . Thus if  $b$  is known but  $a$  unknown, an outlier in the extreme-value sample can be tested by applying to the transformed values a suitable discordancy test for an exponential sample. See Section 3.4.4.

Outlier situations can also arise in the context of *shifted* exponential or gamma distributions. If the origin of the exponential distribution  $\mathbf{E}(\lambda)$  with density  $\lambda^{-1} \exp(-x/\lambda)$  ( $x > 0$ ), 0 ( $x < 0$ ) is shifted to  $x = a$ , say, we get the distribution  $\mathbf{E}(\lambda; a)$  with density  $\lambda^{-1} \exp[-(x-a)/\lambda]$  ( $x > a$ ), 0 ( $x < a$ ). (Similar remarks apply of course to the gamma distribution  $\Gamma(r, \lambda)$ .) Now some discordancy tests for exponential or gamma samples do require the assumption that the origin of the distribution is at zero, or at any rate is known; for example, the test based on the statistic  $x_{(n)} / \sum x_i$  assumes that the origin is zero, and it can obviously be adapted to *known* non-zero origin  $a$  by using  $(x_{(n)} - a) / (\sum x_i - na)$  as statistic. In contrast there are other tests which do not depend on knowledge of the origin, for example the Dixon-type test based on the statistic  $(x_{(n)} - x_{(n-1)}) / (x_{(n)} - x_{(1)})$ ; such tests are useful for two reasons. First, they are needed in the data contexts, sometimes encountered, where a shifted gamma or exponential distribution is the appropriate model. For instance, the development times of diapausing pupae of the cotton bollworm under conditions of constant temperature may be regarded as exponentially distributed with non-zero origin  $t_{\min}$ , this parameter being a minimum development time. Secondly, they are useful for testing outliers in samples from Pareto distributions, since the above-described transformation technique can be applied. Specifically, if the  $n$  values  $x_1, \dots, x_n$  are a sample from a Pareto distribution with origin  $a (> 0)$  and shape parameter  $r$ , then the  $n$  transformed values  $\ln x_1, \dots, \ln x_n$  are a sample from an exponential distribution with origin  $\ln a$  and scale parameter  $1/r$ , i.e. from  $\mathbf{E}(1/r; \ln a)$  in our notation. See Section 3.4.4.

### Contents List: Gamma Samples

The gamma distribution with scale parameter  $\lambda$  and shape parameter  $r$ , i.e. the distribution with density  $f(x) = [\lambda^r \Gamma(r)]^{-1} x^{r-1} \exp(-x/\lambda)$  ( $x > 0$ ), is

denoted by  $\Gamma(r, \lambda)$ . If the origin is shifted to  $a$ , the density is  $[\lambda^r \Gamma(r)]^{-1} (x-a)^{r-1} \exp[-(x-a)/\lambda]$  ( $x > a$ ).  $\Gamma(\nu/2, 2)$  is the  $\chi^2$ -distribution with  $\nu$  degrees of freedom,  $\chi_{\nu}^2$ .  $\Gamma(1, \lambda)$  is the exponential distribution with mean  $\lambda$ , denoted here by  $E(\lambda)$ . The corresponding distribution with origin shifted to  $x = a$  is denoted by  $E(\lambda; a)$ .

In all the tests given here,  $\lambda$  is assumed unknown. Except in test Ga13, the shape parameter  $r$  is assumed *known*. The tests are classified as follows:

Code	Distribution under the working hypothesis
G	gamma with unknown origin
E	exponential with unknown origin
Ga	gamma with known origin 0 (or more generally $a$ )
Ea	exponential with known origin 0 (or more generally $a$ )

Needless to say, any G- or Ga-test can be applied to an exponential sample as a special case ( $r = 1$ ); E- and Ea-tests are specific to the exponential case (generally because tables are only available for this case).

Label	Worksheet page no.	Description of test	Statistic
Ga1(Ea1)	79	Test for a single upper outlier $x_{(n)}$ in a gamma sample	$x_{(n)} / \sum x_i$
Ea2	80	Test for a single upper outlier $x_{(n)}$ in an exponential sample	$x_{(n)} - x_{(n-1)}$
E2	81	Test for a single upper outlier $x_{(n)}$ in an exponential sample irrespective of origin	$\frac{x_{(n)}}{x_{(n)} - x_{(n-1)}}$
Ga3(Ea3)	81	Test for a single lower outlier $x_{(1)}$ in a gamma sample	$x_{(1)} / \sum x_i$
E4	82	Test for a single lower outlier $x_{(1)}$ in an exponential sample with origin unknown	$\frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}$
Ga5(Ea5)	83	Test for $k$ ( $\geq 2$ ) upper outliers $x_{(n-k+1)}, \dots, x_{(n)}$ in a gamma sample	$\frac{x_{(n)} + \dots + x_{(n-k+1)}}{\sum x_i}$
Ea6	83	Test for an upper outlier-pair $x_{(n-1)}, x_{(n)}$ in an exponential sample	$\frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}}$
E6	83	Test for an upper outlier-pair $x_{(n-1)}, x_{(n)}$ in an exponential sample irrespective of origin	$\frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}}$
Ga7(Ea7)	84	Test for a lower and upper outlier-pair $x_{(1)}, x_{(n)}$ in a gamma sample	$\frac{x_{(n)}}{x_{(1)}}$
E8	85	Test for a lower and upper outlier-pair $x_{(1)}, x_{(n)}$ in an exponential sample with origin unknown	$\frac{x_{(n-1)} - x_{(2)}}{x_{(n)} - x_{(1)}}$
Ga9(Ea9)	85	Test for $k$ ( $\geq 2$ ) lower outliers $x_{(1)}, \dots, x_{(k)}$ in a gamma sample	$\frac{x_{(1)} + \dots + x_{(k)}}{\sum x_i}$

Label	Worksheet page no.	Description of test	Statistic
E10	86	Test for a lower outlier-pair $x_{(1)}, x_{(2)}$ in an exponential sample with origin unknown	$\frac{x_{(3)} - x_{(1)}}{x_{(n)} - x_{(1)}}$
Ea11	86	General Dixon-type test for an exponential sample, using knowledge of origin $a$	$\frac{x_{(s)} - x_{(r)}}{x_{(q)} - a}$
E11	87	General Dixon-type test for an exponential sample, irrespective of origin	$\frac{x_{(s)} - x_{(r)}}{x_{(q)} - x_{(p)}}$
E12	88	Test for presence of an undefined number of discordant values in an exponential sample	Shapiro and Wilk's W-statistic. See worksheet
Ga13	88	Testing for discordancy in a gamma sample of unknown shape parameter $r$ by transformation of the variables	See worksheet

Ga1(Ea1) *Discordancy test for a single upper outlier  $x_{(n)}$  in a gamma (or exponential) sample*

*Test statistic:*

$$T_{\text{Ga1}} = \frac{\text{outlier}}{\text{sum of observations}} = \frac{x_{(n)}}{\sum x_j}.$$

*Test distribution:*

For  $\Gamma(r, \lambda)$ ,  $f_n(t) = nb_{r, (n-1)r}(t)$  if  $t \geq \frac{1}{2}$ .

For  $\mathbb{E}(\lambda)$ ,  $F_n(t) = \sum_{j=0}^{\lfloor 1/t \rfloor} (-)^j \binom{n}{j} (1-jt)^{n-1}, \quad 0 \leq t \leq 1,$

where  $\lfloor 1/t \rfloor$  denotes the integer part of  $1/t$ .

*Recurrence relationship:*

*Inequality:*  $f_n(t) = nb_{r, (n-1)r}(t)F_{n-1}\{t/(1-t)\}.$

$$SP(t) \leq nP[\mathbf{F}_{2r, 2(n-1)r} > (n-1)t/(1-t)]; \quad \text{equality for } t \geq \frac{1}{2}.$$

*Tabulated significance levels:* Table I, pp. 290–291; reproduced (with appropriate change of notation) from Eisenhart, Hastay, and Wallis (1947), Tables 15.1 and 15.2, pages 390–391.

*References:* Fisher (1929), Cochran (1941).

*Properties of test:* No special features. All purpose, maximum likelihood ratio test for labelled slippage alternative.

*Example:* Table 3.1 shows a sample of 131 excess cycle times in steel manufacture.

Table 3.1

Excess cycle time $X$	Frequency	$X$	Frequency
1	18	11	6
2	12	12	7
3	18	13	2
4	16	14	1
5	10	15	3
6	4	21	3
7	9	32	2
8	9	35	1
9	2	92	1
10	7		131

The sample of size 130 obtained by omitting the outlier  $x_{(131)} = 92$  has mean  $\bar{x} = 6.44$ , variance  $s^2 = 38.14$ , standard deviation  $s = 6.18$ , and third and fourth moments about the mean  $m_3 = 493.4$ ,  $m_4 = 13444$ . Hence  $\bar{x}/s = 1.04$ ,  $m_3/s^3 = 2.09$ ,  $m_4/s^4 = 9.24$ , suggesting that the distribution may reasonably be assumed exponential ( $\mu/\sigma = 1$ ,  $\mu_3/\sigma^3 = 2$ ,  $\mu_4/\sigma^4 = 9$  for an exponential distribution). On this assumption we can test the outlier 92 for consistency with the other 130 values using test Ea1. The value of  $T_{\text{Ea1}}$  is  $t = 92/929 = 0.0990$ , so

$$\begin{aligned} SP(t) &\leq 131P\left(F_{2,260} > \frac{130 \times 0.0990}{0.9010}\right) \\ &= 131P(F_{2,260} > 14.28) \\ &\doteq 131\left(1 + \frac{14.28}{130}\right)^{-130} = 0.00008, \end{aligned}$$

i.e. the evidence for regarding the value 92 as being too large to have arisen from the same distribution as the other 130 values is very strong.

### Ea2 *Discordancy test for a single upper outlier $x_{(n)}$ in an exponential sample*

*Test statistic:*

$$T_{\text{Ea2}} = \frac{\text{excess}}{\text{outlier}} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)}}.$$

*Test distribution:*

$$F_n(t) = 1 - n(n-1)B\left(\frac{2-t}{1-t}, n-1\right) \quad (0 \leq t \leq 1).$$

*Tabulated significance levels:* Table III, page 293; abridged from Likeš (1966), Table 1, page 49, where 10 per cent, 5 per cent, and 1 per cent points are given for  $n = 2(1)20$ .

*Reference:* Likeš (1966).

*Properties of test:* Vulnerable to masking effect from  $x_{(n-1)}$ .

**Example:** Applying test Ea2 to the example discussed in Worksheet Ga1(Ea1), the value of  $T_{Ea2}$  with  $n = 131$  is  $t = \frac{92 - 35}{92} = 0.6196$ .

Hence

$$\begin{aligned} SP(t) &= 1 - F_n(t) = 131 \times 130B(3.629, 130) \\ &= 131 \times 130\Gamma(3.629)\Gamma(130)/\Gamma(133.629) \\ &= 131!(2.629)(1.629)(0.897)/[(132.63^{133.13})e^{-132.63}\sqrt{(2\pi)}] \\ &= 0.0013 \end{aligned}$$

Compare  $SP(t) \leq 0.00008$  for test Ea1.

**E2 Discordancy test for a single upper outlier  $x_{(n)}$  in an exponential sample with unknown origin**

*Test statistic:*

$$T_{E2} = \frac{\text{excess}}{\text{range}} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}.$$

*Test distribution:*

$$F_n(t) = 1 - (n-1)(n-2)B\left(\frac{2-t}{1-t}, n-2\right) \quad (0 \leq t \leq 1).$$

$T_{E2}$  for a sample of size  $n$  has the same test distribution as  $T_{Ea2}$  for a sample of size  $n-1$ .

*Tabulated significance levels:* Table III, page 293; see Worksheet Ea2.

*References:* Likeš (1966), Kabe (1970).

*Properties of test:* Dixon-type test. Vulnerable to masking effect from  $x_{(n-1)}$ .

**Ga3(Ea3)** *Discordancy test for a single lower outlier  $x_{(1)}$  in a gamma (or exponential) sample*

*Test statistic:*

$$T_{Ga3} = \frac{\text{outlier}}{\text{sum of observations}} = \frac{x_{(1)}}{\sum x_i}.$$

*Test distribution:*

$$\text{For } E(\lambda), \quad f_n(t) = n(n-1)(1-nt)^{n-2} \quad \left(0 \leq t \leq \frac{1}{n}\right),$$

$$0 \quad \left(t > \frac{1}{n}\right).$$

For  $\Gamma(2, \lambda)$  and  $\Gamma(3, \lambda)$ , the following expressions are available for small  $n$ :

$$\begin{aligned}\Gamma(2, \lambda): f_2(t) &= 12t(1-t) \quad (0 \leq t \leq \frac{1}{2}) \\ f_3(t) &= 60t(1-3t)(1-3t^2) \quad (0 \leq t \leq \frac{1}{3}) \\ f_4(t) &= 168t(1-4t)^2(1+3t-12t^2-4t^3) \quad (0 \leq t \leq \frac{1}{4}) \\ f_5(t) &= 360t(1-5t)^3(1+8t-18t^2-80t^3+64t^4) \quad (0 \leq t \leq \frac{1}{5}) \\ f_6(t) &= 660t(1-6t)^4(1+15t-360t^3+864t^5) \quad (0 \leq t \leq \frac{1}{6}) \\ \Gamma(3, \lambda): f_2(t) &= 60t^2(1-t)^2 \quad (0 \leq t \leq \frac{1}{2}) \\ f_3(t) &= 504t^2(1-3t)(1-2t+4t^2-18t^3+21t^4) \quad (0 \leq t \leq \frac{1}{3}) \\ f_4(t) &= 1980t^2(1-4t)^2(1-8t+28t^2-224t^3+1540t^4-5266t^5 \\ &\quad + 11032t^6-16832t^7+13696t^8) \quad (0 \leq t \leq \frac{1}{4})\end{aligned}$$

*Recurrence relationship:*

$$f_n(t) = nb_{r,(n-1)r}(t)(1 - F_{n-1}\{t/(1-t)\}).$$

*Inequality:*

$$SP(t) < nP(\mathbf{F}_{2r,2(n-1)r} < (n-1)t/(1-t)).$$

*Tabulated significance levels:* Table II, page 292; freshly compiled.

*Reference:* Lewis and Fieller (1978).

*Properties of test:* All purpose, maximum likelihood ratio test.

**E4 Discordancy test for a single lower outlier  $x_{(1)}$  in an exponential sample with unknown origin**

*Test statistic:*

$$T_{E4} = \frac{\text{excess}}{\text{range}} = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}.$$

*Test distribution:*

$$F_n(t) = (n-2)B\left(\frac{1+(n-2)t}{1-t}, n-2\right) \quad (0 \leq t \leq 1).$$

*Tabulated significance levels:* Table V, page 296; abridged from Likeš (1966), Table 2, page 51, where 10 per cent, 5 per cent, and 1 per cent points are given for  $n = 3(1)20$ .

*References:* Likeš (1966), Kabe (1970).

*Properties of test:* Dixon-type test. Note a practical difficulty in applying it: the smallest values  $x_{(1)}, x_{(2)}$  need to be given to a sufficient degree of accuracy, which frequently will not be the case in practice (e.g. excess cycle times data, Table 3.1).

**Ga5(Ea5) Discordancy test for  $k$  ( $\geq 2$ ) upper outliers in a gamma (or exponential) sample**

*Test statistic:*

$$T_{\text{Ga5}} = \frac{\text{sum of outliers}}{\text{sum of observations}} = \frac{x_{(n-k+1)} + \dots + x_{(n)}}{\sum x_j}.$$

*Inequality:*

$$SP(t) \leq \binom{n}{k} P\left(\mathbf{F}_{2kr, 2(n-k)r} > \frac{(n-k)t}{k(1-t)}\right).$$

*Reference:* Fieller (1976).

*Properties of test:* Maximum likelihood ratio test. Inequality unlikely to be useful unless  $k$  small.

**Ea6 Discordancy test for an upper outlier-pair  $x_{(n-1)}, x_{(n)}$  in an exponential sample**

*Test statistic:*

$$T_{\text{Ea6}} = \frac{\text{excess}}{\text{outlier}} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)}}.$$

*Test distribution:*

$$F_n(t) = 1 - n(n-1)(n-2) \left[ B\left(\frac{3-2t}{1-t}, n-2\right) - \frac{1}{2} B\left(\frac{3-t}{1-t}, n-2\right) \right].$$

*Reference:* Likeš (1966).

*Properties of test:* Can be used as a discordancy test for  $x_{(n)}$  if it is desired to insure against masking by  $x_{(n-1)}$ .

**E6 Discordancy test for an upper outlier-pair  $x_{(n-1)}, x_{(n)}$  in an exponential sample with unknown origin**

*Test statistic:*

$$T_{\text{E6}} = \frac{\text{excess}}{\text{range}} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}}.$$

*Test distribution:*

$$F_n(t) = 1 - (n-1)(n-2)(n-3) \left[ B\left(\frac{3-2t}{1-t}, n-3\right) - \frac{1}{2} B\left(\frac{3-t}{1-t}, n-3\right) \right].$$

$T_{\text{E6}}$  for a sample of size  $n$  has the same distribution as  $T_{\text{Ea6}}$  for a sample of size  $n-1$ .

*Reference:* Likeš (1966).

**Ga7(Ea7)** *Discordancy test for a lower and upper outlier-pair  $x_{(1)}, x_{(n)}$  in a gamma (or exponential) sample*

*Test statistic:*

$$T_{\text{Ga7}} = \frac{\text{upper outlier}}{\text{lower outlier}} = \frac{x_{(n)}}{x_{(1)}}.$$

*Test distribution:*

$$F_n(t) = \frac{n}{2^{nr}[\Gamma(r)]^n} \int_0^\infty u^{r-1} e^{-u/2} I^{n-1} du \quad \text{where} \quad I = \int_u^{tu} v^{r-1} e^{-v/2} dv.$$

*Recurrence relationship:*

$$f_n(t) = \frac{n(n-1)rt^{r-1}}{(1+t)^{2r}} F_{n-2}(t).$$

*Inequality:*

$$SP(t) \leq n(n-1)P(F_{2r,2r} > t).$$

*Tabulated significance levels:* Table IV, pp. 294–295; reproduced (with appropriate change of notation) from Pearson and Hartley (1966), Table 31, page 202.

*References:* Hartley (1950), David (1952).

*Properties of test:* As with test E4, not suitable where rounding makes value of  $x_{(1)}$  imprecise.

Hartley (1950) gives some values for the power of test Ga7(Ea7) in comparison with Bartlett's global test for heterogeneity of variances, the alternative hypothesis being that the  $n$  population variances are a random sample from a log-normal distribution. The relative power of test Ga7 is 100 per cent when  $n = 2$ , and takes values in the range 90–100 per cent for larger sample sizes (up to twelve). Hartley's figures must be treated with caution, in view of inaccuracies in his tables of percentage points of  $T_{\text{Ga7}}$  (later corrected by David, 1952).

*Example:* The times at which every fourth vehicle travelling westward along a main road in Hull passed an observer were recorded as follows (min:sec):

$$\begin{aligned} 19:57, & \quad 20:14, \quad 20:20, \quad 20:38, \quad 20:50, \\ 21:30, & \quad 21:38, \quad 21:46, \quad 22:07. \end{aligned}$$

There are eight time intervals, viz. 17, 6, 18, 12, 40, 8, 8, and 21 seconds; if the traffic flow is assumed to be random (i.e. in accord with a Poisson process), these will be independent values from a gamma distribution with shape parameter  $r = 4$ . Taking the values 40 and 6 as upper and lower outliers, their ratio is  $40/6 = 6.7$ . The 5 per cent significance point for  $T_{\text{Ga7}}$  with  $n = 8$ ,  $r = 4$  is 10.5, so on the basis of this test there is no reason to believe that the traffic flow was not random. (This conclusion is unaffected if

we make maximal allowance for rounding error and use values 40.5, 5.5 instead of 40, 6, giving a ratio 7.4.)

**E8** Discordancy test for a lower and upper outlier-pair  $x_{(1)}, x_{(n)}$  in an exponential sample with unknown origin

Test statistic:

$$T_{E8} = \frac{\text{reduced range}}{\text{range}} = \frac{x_{(n-1)} - x_{(2)}}{x_{(n)} - x_{(1)}}.$$

Test distribution:

$$F_n(t) = 1 - (n-1)!(1-t)^2 \sum_{j=1}^{n-3} \frac{(-)^{j+1} j}{(j+1)!(n-3-j)!(1+jt)\{n-1-(n-j-2)t\}}.$$

Reference: Kabe (1970).

**Ga9(Ea9)** Discordancy test for  $k$  ( $\geq 2$ ) lower outliers in a gamma (or exponential) sample

Test statistic:

$$T_{Ga9} = \frac{\text{sum of outliers}}{\text{sum of observations}} = \frac{x_{(1)} + \dots + x_{(k)}}{\sum x_i}$$

Test distribution:

For  $E(\lambda)$  and  $k = 2$ ,

$$f_n(t) = \frac{n(n-1)^2}{n-2} [(1 - \frac{1}{2}nt)^{n-2} - (1 - (n-1)t)^{n-2}]$$

$$\text{for } 0 < t < \frac{1}{n-1},$$

$$\frac{n(n-1)^2}{n-2} (1 - \frac{1}{2}nt)^{n-2} \quad \text{for } \frac{1}{n-1} < t < \frac{2}{n},$$

$$0 \quad \text{otherwise.}$$

Inequality:

$$SP(t) < \binom{n}{k} P(F_{2kr, 2(n-k)r} < \frac{(n-k)t}{k(1-t)}).$$

Reference: Fieller (1976), Lewis and Fieller (1978).

Properties of test: Maximum likelihood ratio test.

Example: Epstein (1960b, p. 171) considers a life test in which the failure times of ten items are observed, totalling  $\sum x_i = 600$  units. The failure times of the first two items to fail are the shortest of the ten, and total 24 units, so we can write

$x_{(1)} + x_{(2)} = 24$ . It is assumed that failure times under given conditions are exponentially distributed. Epstein tests whether the first two items to fail can be regarded as having failed abnormally early; making use of a straightforward  $F$ -test, he concludes in favour of this hypothesis. Suppose however that the ten items were placed on test at different starting times and that the two shortest failure times occurred, not necessarily first, but randomly in chronological sequence, so that there is no *a priori* reason to consider these two items as different from the rest. How strong is the evidence for regarding them as inconsistent with the rest in view of their failure times? The value of  $T_{Ea9}$  is

$$t = \frac{24}{600} = 0.04, \text{ which is in the range } 0 < t < \frac{1}{9}.$$

Hence

$$\begin{aligned} SP(t) &= \int_0^{0.04} \frac{810}{8} \left\{ (1-5u)^8 - (1-9u)^8 \right\} du \\ &= \frac{90}{8} \left\{ \frac{(0.64)^9}{9} - \frac{(0.80)^9}{5} - \frac{1}{9} + \frac{1}{5} \right\} = 0.721. \end{aligned}$$

This is the significance probability attaching to the observed ratio  $t = 0.04$ , i.e. there is no real evidence for regarding it as abnormally low—a contrary conclusion to Epstein's, on our modified premise.

#### E10 *Discordancy test for a lower outlier-pair $x_{(1)}, x_{(2)}$ in an exponential sample with unknown origin*

*Test statistic:*

$$T_{E10} = \frac{\text{excess}}{\text{range}} = \frac{x_{(3)} - x_{(1)}}{x_{(n)} - x_{(1)}}.$$

*Test distribution:*

$$F_n(t) = 1 - (n-3) \left[ (n-1)B\left(\frac{1+(n-3)t}{1-t}, n-3\right) - (n-2)B\left(\frac{1+(n-2)t}{1-t}, n-3\right) \right].$$

*References:* Likeš (1966), Kabe (1970).

*Properties of test:* As for test E4, page 82.

#### Ea11 *General Dixon-type discordancy test for an exponential sample, using knowledge of the origin $a$*

*Test statistic:*

$$T_{Ea11} = \frac{x_{(s)} - x_{(r)}}{x_{(q)} - a}, \quad 1 \leq r < s \leq q \leq n.$$

**Test distribution:**

$$1 - F_n(t) = \frac{n!}{(n-q)!} (1-t)$$

$$\times \left\{ \sum_{i=1}^{q-s} \sum_{k=1}^{s-r} \frac{(-)^{i+k} (q-r-i)! [(n-s+k)t + (n-q+i)(1-t)]^{-1}}{(i-1)!(k-1)!(q-s-i)!(s-r-k)!(q-i)!(n-s+k)} \right.$$

$$\left. + \sum_{j=1}^r \sum_{k=1}^{s-r} \frac{(-)^{q-s+j+k} (s-r+j-1)! [(n-s+k)t + (n-r+j)(1-t)]^{-1}}{(j-1)!(k-1)!(r-j)!(s-r-k)!(q-r+j-1)!(n-s+k)} \right\}$$

where the first of the double sums is omitted if  $q = s$ .  $T_{Ea11}$  for sample size  $n$  and observations  $x_{(q)}, x_{(r)}, x_{(s)}$  has the same distribution as  $T_{E11}$  for sample size  $n+p$  and observations  $x_{(p)}, x_{(q+p)}, x_{(r+p)}, x_{(s+p)}$ .

**References:** Likeš (1966), Kabe (1970).

**Properties of test:** Applicable to any combination of lower and/or upper outliers. For example,  $\frac{x_{(n-2)} - x_{(4)}}{x_{(n)} - a}$  would be a suitable statistic for a block test of discordancy of three lower and two upper outliers.

### E11 General Dixon-type discordancy test for an exponential sample irrespective of origin

**Test statistic:**

$$T_{E11} = \frac{x_{(s)} - x_{(r)}}{x_{(q)} - x_{(p)}}, \quad 1 \leq p \leq r < s \leq q \leq n, \quad q-p > s-r.$$

**Test distribution:**

$$1 - F_n(t) = \frac{(n-p)!}{(n-q)!} (1-t)$$

$$\times \left\{ \sum_{i=1}^{q-s} \sum_{k=1}^{s-r} \frac{(-)^{i+k} (q-r-i)! [(n-s+k)t + (n-q+i)(1-t)]^{-1}}{(i-1)!(k-1)!(q-s-i)!(s-r-k)!(q-p-i)!(n-s+k)} \right.$$

$$\left. + \sum_{j=1}^{r-p} \sum_{k=1}^{s-r} \frac{(-)^{q-s+j+k} (s-r+j-1)! [(n-s+k)t + (n-r+j)(1-t)]^{-1}}{(j-1)!(k-1)!(r-p-j)!(s-r-k)!(q-r+j-1)!(n-s+k)} \right\}$$

where the first of the double sums is omitted if  $q = s$  and the second if  $p = r$ .  $T_{E11}$  for sample size  $n$  and observations  $x_{(p)}, x_{(q)}, x_{(r)}, x_{(s)}$  has the same distribution as  $T_{Ea11}$  for sample size  $n-p$  and observations  $x_{(q-p)}, x_{(r-p)}, x_{(s-p)}$ .

**References:** Dixon (1950, 1951), Likeš (1966), Kabe (1970). Note that a factor  $(q-p-i)!$  needs inserting in the denominator of the first double sum in Kabe's equation (13), page 17.

*Properties of test:* Applicable to any combination of lower and/or upper outliers. For example,  $\frac{x_{(n-2)} - x_{(4)}}{x_{(n)} - x_{(1)}}$  would be a suitable statistic for a block test of discordancy for three lower and two upper outliers. E2, E4, E6 are important particular cases.

E12 *Two-sided test for the presence of an undefined number of discordant values in an exponential sample irrespective of origin*

*Test statistic:*

$$T_{E12} = \text{Shapiro and Wilk's 'W-Exponential' statistic}$$

$$= \frac{n(\bar{x} - x_{(1)})^2}{(n-1)\sum_{j=1}^n(x_j - \bar{x})^2}.$$

*Tabulated significance levels:* Table VI, page 297; abridged from Shapiro and Wilk (1972), Table 1, pages 361–362, where lower and upper 0.5, 1, 2.5, 5, and 10 per cent points and the 50 per cent point are given for  $n = 3(1)100$ .

*Further tables:* Shapiro and Wilk (1972) give values for the power of the test against 15 different inherent alternatives (see Chapter 2, page 31).

*Reference:* Shapiro and Wilk (1972).

*Properties of test:* A useful omnibus test against inherent alternatives. In the outlier context, significantly *high* values of  $T_{E12}$  indicate the presence of one or more high discordant values  $x_{(n)}, x_{(n-1)}, \dots$  and/or one *low* discordant value  $x_{(1)}$ ; significantly *low* values of  $T_{E12}$  indicate the presence of a number of low discordant values  $x_{(1)}, x_{(2)}, \dots$ .

Ga13 *Procedure for testing one or more outliers for discordancy in a gamma sample of unknown shape parameter r*

Transform the values  $x_1, \dots, x_n$  in the gamma sample to  $y_1 = \sqrt{x_1}, \dots, y_n = \sqrt{x_n}$ , and apply to the values  $y_1, \dots, y_n$  (all taken positively) a discordancy test for a sample from a normal distribution with unknown mean and variance (tests N1–N17, Section 3.4.3). (For if  $X$  is distributed as  $\Gamma(r, \lambda)$ , then  $\sqrt{X}$  is distributed approximately as  $N(\sqrt{[\lambda(r - \frac{1}{4})]}, \frac{1}{4}\lambda)$ ; and when  $r$  and  $\lambda$  are both unknown, the mean and variance of this approximating normal distribution are both unknown.)

For example, if three upper outliers  $x_{(n)}, x_{(n-1)}, x_{(n-2)}$  in the gamma sample are to be tested for discordancy,  $\sqrt{x_{(n)}}, \sqrt{x_{(n-1)}},$  and  $\sqrt{x_{(n-2)}}$  will be the three greatest values in the  $y$ -sample, and test N3 could appropriately be used, with  $[\sqrt{x_{(n)}} + \sqrt{x_{(n-1)}} + \sqrt{x_{(n-2)}} - 3\bar{y}]/s_y$  as test statistic, where  $\bar{y}$  and  $s_y$  are the mean and standard deviation of the  $y$ -values.

### 3.4.3 Discordancy tests for normal samples

Historically, the motivation for a statistical treatment of outliers came first from the problems of combining astronomical observations, and repeated measurements or determinations must always be one of the main contexts in which discordancy problems arise. In very many cases errors of measurement may plausibly be assumed to follow a normal distribution, whether through the operation of the central limit theorem on contributory error components, or purely as an empirical fact. It is not surprising, therefore, that the vast body of published methodology on outliers from the eighteenth century to the present day rests on the working hypothesis of a normal distribution. Indeed, it is only in the last fifteen years or so that outliers in exponential and other non-normal models have been specifically considered.

When the normal distribution is being used in this way as a kind of all-purpose probability model, the mean  $\mu$  and variance  $\sigma^2$  will both in general be unknown, and any discordancy test for outliers will reflect this. However, discordancy tests also arise in situations when information is available concerning  $\mu$  or  $\sigma^2$  or both. The value of  $\mu$  may be known. The variance  $\sigma^2$  may be known exactly, or again some information on its value may be available in the form of an estimate independent of the particular sample of observations under study for discordancy. This estimate may perhaps be an item of background information 'from the files'. A quite different context giving rise to such an estimate is in analysis of variance, when we may find a surprising value among a set of treatment means, and have available the residual mean square to assist in judging its discordancy. Outlier situations with  $\mu$  unknown but  $\sigma^2$  known may arise in quality control, where past experience provides reasonably accurate knowledge of the process variance. Outlier situations with  $\sigma^2$  unknown but  $\mu$  known may arise, for example, in paired comparison situations where the sample values we are considering for discordancy are differences between corresponding responses and so have mean  $\mu = 0$  on the working hypothesis. The case  $\mu, \sigma^2$  both known is of limited methodological interest, any discordancy test being based simply on the appropriate extreme-value distribution. However, we have included this case as it has practical interest. It could arise, for example, in the validation of tables of random normal deviates; or, less esoterically, in reaching decisions on classification for taxonomic, anthropological, or even legal purposes. For example, in a well known British legal case dating from shortly after the First World War, the husband's basis for bringing divorce proceedings was a 331-day period between his departure for military service abroad and the birth of his wife's child. Extensive data on the duration of pregnancies (whilst not strictly normal) allow one to assume reasonably precise values for mean and variance. A 331-day gestation period is surprising; is it credible or must it be assumed discordant?

Finally, the technique of transformation of the observations leads to a further range of applications of normal-sample discordancy tests, as in the

case of exponential samples (page 77). Tests designed for normal samples with  $\mu$ ,  $\sigma^2$  both unknown can be applied to outliers in samples from gamma distributions with unknown shape parameter. Tests designed for normal samples with known variance  $\sigma^2$  are particularly useful, since they can be applied to outliers in Poisson samples and binomial samples. For example, if the  $n$  values  $x_1, \dots, x_n$  are on the working hypothesis a sample from a Poisson distribution  $P(\mu)$ , then the  $n$  transformed values  $\sqrt{x_1 + \frac{1}{4}}, \dots, \sqrt{x_n + \frac{1}{4}}$  are (provided the mean  $\mu$  is not too small) a sample from a distribution approximately  $N(\sqrt{\mu}, \frac{1}{4})$ . For details of discordancy testing in the Poisson and binomial cases, see Section 3.4.4; for the gamma case  $\Gamma(r, \lambda)$  with unknown  $r$ , see Section 3.4.2 (test Ga13).

### Contents List: Normal Samples

The tests are classified as follows, according to the information available regarding the mean and variance of the normal distribution  $N(\mu, \sigma^2)$  assumed in the working hypothesis.

Code	Information
N	$\mu$ and $\sigma^2$ both unknown
Nv	$\mu$ unknown. Information available on $\sigma^2$ independent of the sample in the form of an estimate $v = s_v^2$ such that $vs_v^2$ is distributed as $\chi_v^2$
N $\mu$	$\mu$ known, $\sigma^2$ unknown
N $\sigma$	$\sigma^2$ known, $\mu$ unknown
N $\mu\sigma$	$\mu$ and $\sigma^2$ both known

In the case N,  $\mu$  is estimated on the working hypothesis by  $\bar{x} = \sum x_i/n$ , and  $\sigma^2$  by  $s^2 = \sum (x_i - \bar{x})^2/(n-1)$ . In the case Nv,  $\sigma^2$  can be estimated by the independent estimators  $s^2$ ,  $s_v^2$ , or by the pooled estimator

$$\tilde{s}^2 = [\sum (x_i - \bar{x})^2 + vs_v^2]/(n-1+\nu).$$

In the case N $\mu$ ,  $\sigma^2$  is estimated by  $s^2(\mu) = \sum (x_i - \mu)^2/n$ .

The sum of squares of the deviations from  $\bar{x}$  of the  $n$  observations  $x_1, x_2, \dots, x_n$ ,  $\sum_{i=1}^n (x_i - \bar{x})^2$ , is denoted by  $S^2$ . Thus  $s^2 = S^2/(n-1)$ . If  $x_{(n)}$  is omitted, the sum of squares of the deviations of the remaining  $n-1$  observations from their own mean is denoted by  $S_{n-1}^2$ .  $S_{n-1,n}^2$  is the corresponding sum of squares when  $x_{(n-1)}, x_{(n)}$  are both omitted, and so on. The quantity  $\sum_{i=1}^n (x_i - \mu)^2$ , which is of relevance to case N $\mu$ , is denoted by  $S^2(\mu)$ , with  $s^2(\mu) = S^2(\mu)/n$ , and with similar definitions for  $S_n^2(\mu)$ ,  $S_{n-1,n}^2(\mu)$  etc. In the case Nv, the sum of squares  $\sum (x_i - \bar{x})^2 + vs_v^2$  is denoted by  $\tilde{S}^2$ .

Tests not involving the sample mean or population mean are coded both as N and N $\mu$  (or, where  $\sigma^2$  is known, both as N $\sigma$  and N $\mu\sigma$ ). Examples: N7(N $\mu$ 7), test statistic  $(x_{(n)} - x_{(n-1)})/(x_{(n)} - x_{(1)})$ ; N $\sigma$ 6(N $\mu\sigma$ 6), test statistic  $(x_{(n)} - x_{(1)})/\sigma$ .

In a situation where  $\mu$  is known but no appropriate test is listed under code  $N\mu$  because no significance levels are available, or similarly where  $\sigma^2$  is known but no appropriate test is listed under code  $N\sigma$ , there may be an appropriate N-test which can be used though with some loss of efficiency. For example, to test for a lower and upper outlier-pair  $x_{(1)}, x_{(n)}$  with  $\mu$  known,  $\sigma^2$  unknown, N6 could be used, since significance levels for a test based on  $(x_{(n)} - x_{(1)})/s(\mu)$  are not available. Where both  $\mu$  and  $\sigma^2$  are known an  $N\sigma$ -test (or possibly an  $N\mu$ -test) can be used if necessary.

In view of the symmetry of the normal distribution, any test for an upper outlier, upper outlier-pair etc., can be used for a lower outlier, lower outlier-pair etc., with the obvious modifications. For example, two lower outliers  $x_{(1)}, x_{(2)}$  can be tested for discordancy by test N3 using the statistic  $(2\bar{x} - x_{(2)} - x_{(1)})/s$ . To save space, such tests are only given here in terms of the upper outlier situation.

Label	Worksheet page no.	Description of test	Statistic
N1	93	Test for upper outlier $x_{(n)}$	$\frac{x_{(n)} - \bar{x}}{s}$ or equivalently $S_n^2/S^2$
N2	94	Test for extreme outlier (two-sided form of N1)	$\max\left(\frac{x_{(n)} - \bar{x}}{s}, \frac{\bar{x} - x_{(1)}}{s}\right)$ or equivalently $\min(S_n^2/S^2, S_1^2/S^2)$
N3	95	Test for $k (\geq 2)$ upper outliers $x_{(n-k+1)}, \dots, x_n$	$\frac{x_{(n-k+1)} + \dots + x_{(n)} - k\bar{x}}{s}$
N4	96	Test for $k (\geq 2)$ upper outliers $x_{(n-k+1)}, \dots, x_{(n)}$	$S_{n-k+1, \dots, n-1, n}^2/S^2$
N5	96	Test for lower and upper outlier-pair $x_{(1)}, x_{(n)}$	$S_{1,n}^2/S^2$
N6	97	Test for lower and upper outlier-pair $x_{(1)}, x_{(n)}$	$\frac{x_{(n)} - x_{(1)}}{s}$
N7( $N\mu$ 7)	97	Dixon-type test for upper outlier $x_{(n)}$	$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$
N8( $N\mu$ 8)	98	Dixon-type test for extreme outlier (two-sided form of N7( $N\mu$ 7))	$\max\left[\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}, \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}\right]$
N9( $N\mu$ 9)	98	Dixon-type test for upper outlier $x_{(n)}$	$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}$
N10( $N\mu$ 10)	99	Dixon-type test for upper outlier $x_{(n)}$	$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(3)}}$
N11( $N\mu$ 11)	99	Dixon-type test for two upper outliers $x_{(n-1)}, x_{(n)}$	$\frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}}$
N12( $N\mu$ 12)	100	Dixon-type test for two upper outliers $x_{(n-1)}, x_{(n)}$	$\frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}}$
N13( $N\mu$ 13)	100	Dixon-type test for two upper outliers $x_{(n-1)}, x_{(n)}$	$\frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}}$
N14	100	Test for one or more upper outliers	Sample skewness $g_1 = \sqrt{b_1}$

Label	Worksheet page no.	Description of test	Statistic
N15	10	Two-sided test for one or more outliers, irrespective of their directions	Sample kurtosis $b_2$
N16	102	Block test for $k$ outliers irrespective of directions (i.e. of how many upper and how many lower)	Tietjen and Moore's $E_k$ -statistic. See worksheet
N17	102	Test for presence of an undefined number of discordant values	Shapiro and Wilk's W-statistic. See worksheet
Nv1	103	Test for upper outlier $x_{(n)}$	$\frac{x_{(n)} - \bar{x}}{s_v}$
Nv2	104	Test for upper outlier $x_{(n)}$	$\frac{x_{(n)} - \bar{x}}{\tilde{s}}$
Nv3		Test for extreme outlier (two-sided form of Nv1)	$\max\left(\frac{x_{(n)} - \bar{x}}{s_v}, \frac{\bar{x} - x_{(1)}}{s_v}\right)$
Nv4	105	Test for extreme outlier (two-sided form of Nv2)	$\max\left(\frac{x_{(n)} - \bar{x}}{\tilde{s}}, \frac{\bar{x} - x_{(1)}}{\tilde{s}}\right)$
Nv5	106	Test for $k (\geq 2)$ upper outliers $x_{(n-k+1)}, \dots, x_{(n)}$	$\frac{x_{(n-k+1)} + \dots + x_{(n)} - k\bar{x}}{\tilde{s}}$
Nv6	106	Test for lower and upper outlier-pair $x_{(1)}, x_{(n)}$	$\frac{x_{(n)} - x_{(1)}}{s_v}$
Nv7	107	Test for lower and upper outlier-pair $x_{(1)}, x_{(n)}$	$\frac{x_{(n)} - x_{(1)}}{\tilde{s}}$
N $\mu$ 1	107	Test for upper outlier $x_{(n)}$	$\frac{x_{(n)} - \mu}{s(\mu)}$
N $\mu$ 2	108	Test for extreme outlier (two-sided form of N $\mu$ 1)	$\max\left(\frac{ x_{(n)} - \mu }{s(\mu)}, \frac{ \mu - x_{(1)} }{s(\mu)}\right)$ or equivalently $\min(S_n^2(\mu)/S^2(\mu), S_1^2(\mu)/S^2(\mu))$
N $\mu$ 3	108	Test for $k (\geq 2)$ upper outliers $x_{(n-k+1)}, \dots, x_{(n)}$	$\frac{x_{(n-k+1)} + \dots + x_{(n)} - k\mu}{s(\mu)}$
N $\mu$ 4	108	Test for two upper outliers $x_{(n-1)}, x_{(n)}$	$S_{n-1,n}^2(\mu)/S^2(\mu)$
N $\mu$ 5	109	Test for lower and upper outlier-pair $x_{(1)}, x_{(n)}$	$S_{1,n}^2(\mu)/S^2(\mu)$
N $\mu$ 6	109	Two-sided test for one or more extreme outliers	$\frac{\sum_{j=1}^n (x_j - \mu)^4}{ns^4(\mu)}$
N $\mu$ 14	109	Block test for $k$ outliers irrespective of directions	See worksheet
N $\sigma$ 1	110	Test for upper outlier $x_{(n)}$	$\frac{x_{(n)} - \bar{x}}{\sigma}$
N $\mu\sigma$ 1	111	Test for upper outlier $x_{(n)}$	$\frac{x_{(n)} - \mu}{\sigma}$
N $\sigma$ 2	112	Test for extreme outlier (two-sided form of N $\sigma$ 1)	$\max\left(\frac{x_{(n)} - \bar{x}}{\sigma}, \frac{\bar{x} - x_{(1)}}{\sigma}\right)$

Label	Worksheet page no.	Description of test	Statistic
N $\sigma$ 3	112	Test for $k (\geq 2)$ upper outliers $x_{(n-k+1)}, \dots, x_{(n)}$	$\frac{x_{(n-k+1)} + \dots + x_{(n)} - k\bar{x}}{\sigma}$
N $\mu\sigma$ 3	112	Test for $k (\geq 2)$ upper outliers $x_{(n-k+1)}, \dots, x_{(n)}$	$\frac{x_{(n-k+1)} + \dots + x_{(n)} - k\mu}{\sigma}$
N $\sigma$ 4	113	Test for two upper outliers $x_{(n-1)}, x_{(n)}$	$S_{n-1,n}^2/\sigma^2$
N $\mu\sigma$ 4	113	Test for two upper outliers $x_{(n-1)}, x_{(n)}$	$S_{n-1,n}^2(\mu)/\sigma^2$
N $\sigma$ 5	113	Test for lower and upper outlier-pair $x_{(1)}, x_{(n)}$	$S_{1,n}^2/\sigma^2$
N $\mu\sigma$ 5	113	Test for lower and upper outlier-pair $x_{(1)}, x_{(n)}$	$S_{1,n}^2(\mu)/\sigma^2$
N $\sigma$ 6(N $\mu\sigma$ 6)	113	Test for lower and upper outlier-pair $x_{(1)}, x_{(n)}$	$\frac{x_{(n)} - x_{(1)}}{\sigma}$
N $\sigma$ 7(N $\mu\sigma$ 7)	114	Test for upper outlier $x_{(n)}$	$\frac{x_{(n)} - x_{(n-1)}}{\sigma}$
N $\sigma$ 8(N $\mu\sigma$ 8)	114	Test for two upper outliers $x_{(n-1)}, x_{(n)}$	$\frac{x_{(n-1)} - x_{(n-2)}}{\sigma}$
N $\sigma$ 9(N $\mu\sigma$ 9)	115	Test for $k$ lower and $k$ upper outliers	$\frac{x_{(n-k+1)} - x_{(k)}}{\sigma}$

N1 *Discordancy test for a single upper outlier  $x_{(n)}$  in a normal sample with  $\mu$  and  $\sigma^2$  unknown*

*Test statistic:*

$$T_{N1} = \text{internally studentized extreme deviation from mean} = \frac{x_{(n)} - \bar{x}}{s}.$$

An equivalent statistic is:

$$\frac{\text{reduced sum of squares}}{\text{total sum of squares}} = \frac{S_n^2}{S^2} = 1 - \frac{n}{(n-1)^2} T_{N1}^2.$$

*Recurrence relationship:*

$$f_n(t) = \frac{n}{n-1} \left( \frac{n}{\pi} \right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)} \left( 1 - \frac{nt^2}{(n-1)^2} \right)^{(n-4)/2} F_{n-1} \left[ \left( \frac{n^2(n-2)t^2}{((n-1)((n-1)^2 - nt^2)} \right)^{\frac{1}{2}} \right]$$

$$\left( \frac{1}{\sqrt{n}} \leq t \leq \frac{n-1}{\sqrt{n}} \right)$$

with

$$F_2(t) = 0 \left( t < \frac{1}{\sqrt{2}} \right), 1 \left( t > \frac{1}{\sqrt{2}} \right).$$

*Inequality:*

$$SP(t) \leq nP\left(t_{n-2} > \left[\frac{n(n-2)t^2}{(n-1)^2 - nt^2}\right]^{\frac{1}{2}}\right).$$

This is an equality when  $t \geq [(n-1)(n-2)/2n]^{\frac{1}{2}}$ .

*Tabulated significance levels:* Table VIIa, page 298; abridged from Grubbs and Beck (1972), Table I, pages 848–850, where 0.1, 0.5, 1, 2.5, 5, and 10 per cent points are given for  $n = 3(1)147$ .

*Further tables:* Dixon (1950) gives graphs of the performance measure  $P_3$  (see page 66) for  $n = 5, 15$  and for alternatives of slippage in location and slippage in dispersion by one and two observations; the figures are derived from sampling experiments of size 200 at most. Ferguson (1961a) gives tables of power  $P_1$  (pages 65–66) for the alternative of slippage in location by a single observation. David and Paulson (1965) give graphs of performance measure  $P_2$  (page 66) in relation to the same alternative, for  $n = 4(2)10$ . McMillan (1971) gives graphs of the performance measures  $P(C_1), P(C_2), P(C_3)$  (page 74) when N1 is used consecutively for the testing of two upper outliers; some corrections to these results are given by Moran and McMillan (1973).

*References:* Pearson and Chandra Sekar (1936), Dixon (1950, 1962), Grubbs (1950, 1969), Kudo (1956a), Ferguson (1961a, 1961b), Quesenberry and David (1961), David and Paulson (1965), Stefansky (1971), McMillan (1971), Moran and McMillan (1973).

*Properties of test:* N1 is the maximum likelihood ratio test for a location-slippage alternative in which one observation arises from a normal distribution  $\mathbf{N}(\mu + a, \sigma^2)$ ,  $a > 0$ . For this alternative, it has the optimal property of being the scale- and location-invariant test of given size which maximizes the probability  $P_3$  of identifying the contaminant as discordant. Vulnerable to masking effect when there is more than one contaminant, but less so than N7(N $\mu$ 7). Not very suitable for consecutive use when testing several outliers; preferable in this case to use a block procedure or to use N15 consecutively.

## N2 Two-sided discordancy test for an extreme outlier in a normal sample with $\mu$ and $\sigma^2$ unknown

*Test statistic:*

$$T_{N2} = \max\left(\frac{x_{(n)} - \bar{x}}{s}, \frac{\bar{x} - x_{(1)}}{s}\right).$$

An equivalent statistic is:

$$\min\left(\frac{S_n^2}{S^2}, \frac{S_1^2}{S^2}\right) = 1 - \frac{n}{(n-1)^2} T_{N2}^2.$$

**Inequality:**

$SP(t) \leq 2P(T_{N_1} > t)$ . Equality holds for  $t > \sqrt{[(n-1)(n-2)/2n]}$ .

**Tabulated significance levels:** Table VIIb, page 298; derived from Pearson and Hartley (1966), Table 26b, page 188.

**Further tables:** Ferguson (1961a) gives tables of power  $P_1$  for alternatives of slippage in location by a single observation and by two observations.

**References:** Kudo (1956), Ferguson (1961a, 1961b), Quesenberry and David (1961), Tietjen and Moore (1972). Tietjen and Moore, working in terms of the equivalent statistic  $\min(S_n^2/S^2, S_1^2/S^2)$ , present the inequality for  $SP(t)$  as an equality.

**Properties of test:** Maximum likelihood ratio test for a location-slippage alternative in which one observation arises from a normal distribution  $N(\mu + a, \sigma^2)$ ,  $a \neq 0$ . For this alternative, has the optimal property of being the scale- and location-invariant test of given size which maximizes the probability ( $P_3$ , page 66) of identifying the contaminant as discordant. Vulnerable to masking effect in small samples when there are two outliers in the same direction.

**N3 Discordancy test for  $k$  upper outliers  $x_{(n-k+1)}, \dots, x_{(n-1)}, x_{(n)}$  in a normal sample with  $\mu$  and  $\sigma^2$  unknown**

**Test statistic:**

$T_{N_3}$  = sum of internally studentized deviations from the mean

$$= \frac{x_{(n-k+1)} + \dots + x_{(n-1)} + x_{(n)} - k\bar{x}}{s}.$$

**Inequality:**

$$SP(t) \leq \binom{n}{k} P\left(t_{n-2} > \left[ \frac{n(n-2)t^2}{k(n-k)(n-1)-nt^2} \right]^{\frac{1}{2}} \right).$$

This is an equality when  $t \geq [k^2(n-1)(n-k-1)/(nk+n)]^{\frac{1}{2}}$ .

**Tabulated significance levels:** Table IXa, page 304; freshly compiled on the basis of simulations of sizes 10 000.

**References:** Murphy (1951), Kudo (1956a), Ferguson (1961b), McMillan (1971), Fieller (1976). McMillan gives results for the comparative performance of tests N3 and N4 as applied to two upper outliers ( $k = 2$ ); see N4 Worksheet.

**Properties of test:** N3 is the maximum likelihood ratio test for a location-slippage alternative in which  $k$  observations arise from a common normal distribution  $N(\mu + a, \sigma^2)$ ,  $a > 0$ . For this alternative, it has the optimal

property of being the scale- and location-invariant test of given size which maximizes the probability of identifying the  $k$  contaminants as discordant.

**N4 Discordancy test for  $k(\geq 2)$  upper outliers  $x_{(n-k+1)}, \dots, x_{(n-1)}, x_{(n)}$  in a normal sample with  $\mu$  and  $\sigma^2$  unknown**

*Test statistic:*

$$T_{N4} = \frac{\text{reduced sum of squares}}{\text{total sum of squares}} = \frac{S_{n-k+1, \dots, n-1, n}^2}{S^2}.$$

*Tabulated significance levels:* Table IXb, page 304; values for  $k = 2$  abridged from Grubbs and Beck (1972), Table II, pages 851–853, where 0.1, 0.5, 1, 2.5, 5, and 10 per cent points are given for  $n = 4(1)149$ ; values for  $k = 3$  and  $k = 4$  abridged from Tietjen and Moore (1972), Table I, pages 587–590, where 1, 2.5, 5, and 10 per cent points are given for  $k = 1(1)10$  and  $n = \max[3, 2k](1)20(5)50$ . Note that values of  $T_{N4}$  smaller than the tabulated level are significant.

*References:* Grubbs (1950, 1969), Dixon (1950), McMillan (1971), Tietjen and Moore (1972), Fieller (1976).

*Properties of test:* N4 is the maximum likelihood ratio test for a location-slippage alternative in which  $k$  observations arise from separate normal distributions each with variance  $\sigma^2$  but with distinct means all exceeding  $\mu$ .

A study by McMillan (1971) of the performance measure  $P_2$  (see page 66) for tests N3 and N4 in the case  $k = 2$  indicates that N4 is more robust than N3 against departures from the relevant alternative.

Commonly the number,  $k$ , of outliers to be tested will have been chosen, either as being the number of manifest outliers, or as a parameter in a data-processing procedure. As an alternative Tietjen and Moore (1972) suggest proceeding in the following way: find the ‘largest gap’, i.e. the largest of the intervals  $x_{(n)} - x_{(n-1)}$ ,  $x_{(n-1)} - x_{(n-2)}$ , …, to the right of the mean  $\bar{x}$ , and fix upon the observations to the right of this gap ( $k$  in number, say) for testing as upper outliers. Tietjen and Moore show that N4 applied as a block test to these  $k$  outliers has rather better performance, in terms of proportion of contaminants correctly identified as discordant, than consecutive test procedures using either N14 or N17, and much better performance than consecutive procedures using either N1 or (an unspecified) one of the Dixon-type tests.

**N5 Discordancy test for a lower and upper outlier-pair  $x_{(1)}, x_{(n)}$  in a normal sample with  $\mu$  and  $\sigma^2$  unknown**

*Test statistic:*

$$T_{N5} = \frac{\text{reduced sum of squares}}{\text{total sum of squares}} = \frac{S_{1,n}^2}{S^2}.$$

**Tabulated significance levels:** Table Xa, page 306; freshly compiled on the basis of simulations of sizes 10 000. Values of  $T_{N5}$  smaller than the tabulated level are significant.

**References:** Grubbs (1950), Ferguson (1961b), Fieller (1976).

**Properties of test:** Maximum likelihood ratio test for a location-slippage alternative in which two observations arise from separate normal distributions  $N(\mu + a_1, \sigma^2)$ ,  $N(\mu + a_2, \sigma^2)$ ,  $a_1 < 0 < a_2$ .

**N6 Discordancy test for a lower and upper outlier-pair  $x_{(1)}, x_{(n)}$  in a normal sample with  $\mu$  and  $\sigma^2$  unknown**

**Test statistic:**

$$T_{N6} = \text{internally studentized range} = \frac{x_{(n)} - x_{(1)}}{s}.$$

**Inequality:**

$$SP(t) \leq n(n-1)P\left(t_{n-2} > \left[\frac{(n-2)t^2}{2n-2-t^2}\right]^{\frac{1}{2}}\right).$$

This is an equality when  $t \geq [\frac{3}{2}(n-1)]^{\frac{1}{2}}$ .

**Tabulated significance levels:** Table XIa, page 307; abridged from Pearson and Hartley (1966), Table 29c, page 200, where lower and upper limits and lower and upper 0.5, 1, 2.5, 5, and 10 per cent points are given for  $n = 3(1)20(5)100, 150, 200, 500, 1000$ .

**Further tables:** Shapiro, Wilk, and Chen (1968) give values for the power of the test against 45 different inherent alternatives (see Chapter 2, page 31).

**References:** David, Hartley, and Pearson (1954), Pearson and Stephens (1964), Shapiro, Wilk and Chen (1968).

**Properties:** As a test against an *inherent* alternative, N6 has good power properties against a variety of symmetric distributions alternative to the normal, but performs poorly with respect to asymmetric alternatives.

**N7(Nμ7) Discordancy test for a single upper outlier  $x_{(n)}$  in a normal sample with  $\sigma^2$  unknown**

**Test statistic:**

$$T_{N7} = \frac{\text{excess}}{\text{range}} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \quad (\text{Dixon's } r_{10} \text{ statistic}).$$

**Test distribution:** For small  $n$ , we have:

$$f_3(t) = \frac{3\sqrt{3}}{2\pi} (t^2 - t + 1)^{-1}$$

$$f_4(t) = \frac{2}{\sqrt{3}} f_3(t) [(1 - 2t)(4t^2 - 4t + 3)^{-\frac{1}{2}} - (t - 2)(3t^2 - 4t + 4)^{-\frac{1}{2}}].$$

*Tabulated significance levels:* Table XIIIa, page 311; abridged from Dixon (1951), Table I, page 73, where upper 0.5, 1, 2 per cent points, upper and lower 5, 10, 20, 30, 40 per cent points, and the 50 per cent point are given for  $n = 3(1)30$ .

*Further tables:* Dixon (1950) gives graphs of the performance measure  $P_3$ , based on sampling experiments of comparatively small size (66–200 replications). Ferguson (1961a) gives tables of power  $P_1$  for the alternative of slippage in location by one observation.

*References:* Dixon (1950, 1951), Ferguson (1961a, 1961b).

*Properties of test:* Mainly effective when there is at most one discordant value, otherwise vulnerable to possible masking effect of  $x_{(n-1)}$  and/or  $x_{(1)}$ ; see properties of test N1. The performances of tests N7(N $\mu$ 7) as measured both by  $P_3$  and  $P_1$ , against the alternative of slippage in location by a single observation, are effectively the same for sample sizes up to 15.

N8(N $\mu$ 8) *Two-sided discordancy test for an extreme outlier in a normal sample with  $\sigma^2$  unknown*

*Test statistic:*

$$T_{N8} = \max \left[ \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}, \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}} \right].$$

*Inequality:*

$$SP(t) \leq 2P(T_{N7} > t).$$

This is an equality when  $t \geq \frac{1}{2}$ .

*Tabulated significance levels:* Table XIIIb, page 311; freshly compiled on the basis of simulations of sizes 10 000.

*Reference:* King (1953).

*Properties of test:* Two-sided form of N7(N $\mu$ 7).

N9(N $\mu$ 9) *Discordancy test for a single upper outlier  $x_{(n)}$  in a normal sample with  $\sigma^2$  unknown*

*Test statistic:*

$$T_{N9} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}} \quad (\text{Dixon's } r_{11} \text{ statistic}).$$

*Test distribution:* For  $n = 4$  we have:

$$f_4(t) = \frac{3\sqrt{3}}{\pi} (t^2 - t + 1)^{-\frac{1}{2}} [1 + (t-2)[3(4 - 4t + 3t^2)]^{-\frac{1}{2}}].$$

**Tabulated significance levels:** Table XIIIc, page 311; abridged from Dixon (1951), Table II, page 74, where more extensive values are given, as detailed in Worksheet N7(N $\mu$ 7).

**Further tables:** Dixon (1950) gives graphs of  $P_3$ ; see Worksheet N7(N $\mu$ 7).

**References:** Dixon (1950, 1951).

**Properties of test:** Advantage: avoids any possible masking effect of lowest sample value  $x_{(1)}$  (by inflation of the denominator). Disadvantage: vulnerable to masking effect of  $x_{(n-1)}$ . See also Worksheet N12(N $\mu$ 12) in relation to performance.

**N10(N $\mu$ 10) Discordancy test for a single upper outlier  $x_{(n)}$  in a normal sample with  $\sigma^2$  unknown**

**Test statistic:**

$$T_{N10} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(3)}} \quad (\text{Dixon's } r_{12} \text{ statistic}).$$

**Tabulated significance levels:** Table XIIIId, page 311; abridged from Dixon (1951), Table III, page 75, where more extensive values are given, as detailed in Worksheet N7(N $\mu$ 7).

**Further tables and references:** As for N9(N $\mu$ 9).

**Properties of test:** Avoids any possible masking effect of the two lowest observations  $x_{(1)}, x_{(2)}$  on the testing of  $x_{(n)}$ , but is vulnerable to any masking effect of  $x_{(n-1)}$ . See also Worksheet N13(N $\mu$ 13) in relation to performance.

**N11(N $\mu$ 11) Discordancy test for an upper outlier-pair  $x_{(n-1)}, x_{(n)}$  in a normal sample with  $\sigma^2$  unknown**

**Test statistic:**

$$T_{N11} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}} \quad (\text{Dixon's } r_{20} \text{ statistic}).$$

**Tabulated significance levels:** Table XIIIe, page 311; abridged from Dixon (1951), Table IV, page 76, where more extensive values are given, as detailed in Worksheet N7(N $\mu$ 7).

**Further tables and references:** As for N9(N $\mu$ 9).

**Properties of test:** N11(N $\mu$ 11) can also be used as a discordancy test for a single upper outlier  $x_{(n)}$  which avoids the risk of masking by  $x_{(n-1)}$ .

**N12(N $\mu$ 12)** *Discordancy test for an upper outlier-pair  $x_{(n-1)}, x_{(n)}$  in a normal sample with  $\sigma^2$  unknown*

*Test statistic:*

$$T_{N12} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}} \quad (\text{Dixon's } r_{21} \text{ statistic}).$$

*Tabulated significance levels:* Table XIIIIf, page 311; abridged from Dixon (1951), Table V, page 77, where more extensive values are given, as detailed in Worksheet N7(N $\mu$ 7).

*Further tables and references:* As for N9(N $\mu$ 9).

*Properties of test:* Avoids any possible masking effect from  $x_{(1)}$ . Can be used as a discordancy test for a single upper outlier  $x_{(n)}$ , and for this purpose is to be preferred to test N9(N $\mu$ 9), since its performance is similar against a single contaminant and it avoids the risk of masking from  $x_{(n-1)}$  if there is more than one contaminant.

**N13(N $\mu$ 13)** *Discordancy test for an upper outlier-pair  $x_{(n-1)}, x_{(n)}$  in a normal sample with  $\sigma^2$  unknown*

*Test statistic:*

$$T_{N13} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}} \quad (\text{Dixon's } r_{22} \text{ statistic}).$$

*Tabulated significance levels:* Table XIIIg, page 311; abridged from Dixon (1951), Table VI, page 78, where more extensive values are given, as detailed in Worksheet N7(N $\mu$ 7).

*Further tables:* As for N9(N $\mu$ 9).

*References:* Dixon (1950, 1951), Ferguson (1961b).

*Properties of test:* Avoids any possible masking effect from the two lowest observations  $x_{(1)}, x_{(2)}$ . Can be used as a discordancy test for a single upper outlier  $x_{(n)}$ , and for this purpose is superior to test N10(N $\mu$ 10), having a similar performance against a single contaminant but being more robust against the presence of a second upper outlier at  $x_{(n-1)}$ .

**N14** *Discordancy test for one or more upper (or lower) outliers in a normal sample with  $\mu$  and  $\sigma^2$  unknown*

*Test statistic:*

$$T_{N14} = \text{sample skewness} = \left[ \frac{\sum_{j=1}^n (x_j - \bar{x})^3}{ns^3} \right]^{\frac{1}{2}}.$$

The value tested for discordancy is  $x_{(n)}$  or  $x_{(1)}$  according as the sign of  $T_{N14}$  is + or -. For more than one outlier, apply test consecutively.

**Tabulated significance levels:** Table XIVa, page 312; abridged from Pearson and Hartley (1966), Table 34B, page 207, where 5 per cent and 1 per cent points are given for  $n = 25(5)50(10)100(25)200(50)1000(200)2000(500)5000$ , and from Ferguson (1961a), Table I, page 281, where estimated 10, 5, and 1 per cent points are given for  $n = 5(5)25$ .

**Further tables:** Ferguson (1961a) gives tables of power  $P_1$  against the alternative of slippage in location by a single observation. Shapiro, Wilk, and Chen (1968) give values for power against 45 different *inherent* alternatives (see Chapter 2, page 31).

**References:** Ferguson (1961a, 1961b), Shapiro, Wilk, and Chen (1968).

**Properties of test:** N14 is the locally best invariant test of given size against a location-slippage alternative in which  $k$  of the  $n$  observations arise from separate normal distributions  $N(\mu + a_1, \sigma^2), \dots, N(\mu + a_k, \sigma^2)$ ,  $a_1 > 0, a_2 > 0, \dots, a_k > 0$ , whatever the values of the  $a$ 's, and whatever the value of  $k$  provided only that the contamination proportion  $k/n$  under the alternative hypothesis is less than  $\frac{1}{2}$ .

Its power is nearly as good as that of N1 against slippage in location for a single observation by medium or large amounts. It also has good power against inherent Cauchy and log-normal alternatives.

#### N15 Discordancy test for one or more outliers (irrespective of their directions) in a normal sample with $\mu$ and $\sigma^2$ unknown

**Test statistic:**

$$T_{N15} = \text{sample kurtosis} = \frac{\sum_{j=1}^n (x_j - \bar{x})^4}{ns^4}.$$

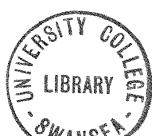
The value tested for discordancy is whichever of  $x_{(n)}$  or  $x_{(1)}$  is further from  $\bar{x}$ . Discordancy is indicated by *high* values of the statistic. For more than one outlier, apply test consecutively.

**Tabulated significance levels:** Table XIVb, page 312; abridged from Pearson and Hartley (1966), Table 34C, page 208, where lower and upper 1 per cent and 5 per cent points are given for  $n = 50(25)150(50)700(100)1000(200)2000(500)5000$ , and from Ferguson (1961a), Table II, page 282, where estimated 1, 5, and 10 per cent points are given for  $n = 5(5)25$ .

**Further tables:** Ferguson (1961a) gives tables of power against alternatives of slippage in location by a single observation and by two observations. Shapiro, Wilk, and Chen (1968) give values for power against 45 different inherent alternatives (see Chapter 2, page 31).

**References:** Ferguson (1961a, 1961b), Shapiro, Wilk, and Chen (1968).

**Properties of test:** N15 is the locally best unbiased invariant test of given size against a location-slippage alternative in which  $k$  of the  $n$  observations arise



from separate normal distributions  $\mathbf{N}(\mu + a_1, \sigma^2), \dots, \mathbf{N}(\mu + a_k, \sigma^2)$ , where  $a_1, \dots, a_k$  differ from zero but are otherwise arbitrary, provided that the contamination proportion  $k/n$  under the alternative hypothesis is less than 0.21. N15 is also the locally best invariant test of given size against a dispersion-slippage alternative in which  $k$  of the observations arise from separate normal distributions  $\mathbf{N}(\mu, b_1\sigma^2), \dots, \mathbf{N}(\mu, b_k\sigma^2)$ ,  $b_1 > 1, \dots, b_k > 1$ , irrespective of the proportion  $k/n$ .

Its power is nearly as good as that of N2 against slippage in location for a single observation by medium or large amounts. Against slippage in location by two observations it is superior to N2 in power, greatly so when the sample size is less than, say, 20.

N15 has the advantage of being robust against possible masking effect. It is suitable for consecutive use in the possible presence of more than one outlier.

**N16 Two-sided discordancy test for  $k$  outliers (irrespective of directions) in a normal sample with  $\mu$  and  $\sigma^2$  unknown**

*Test statistic:*

$$T_{N16} = \text{Tietjen and Moore's } E_k\text{-statistic} = \frac{\sum_{j=1}^{n-k} (r_{(j)} - \bar{r}_{n-k})^2}{\sum_{j=1}^n (r_{(j)} - \bar{r})^2}$$

where  $r_j = |x_j - \bar{x}|$ , the absolute deviation of  $x_j$  from the sample mean;  $\{r_{(j)}\}$  are the values of the  $r_j$  in ascending order,  $r_{(1)} < r_{(2)} < \dots < r_{(n)}$ ;  $\bar{r}$  is the mean of all the  $r$ 's; and  $\bar{r}_{n-k}$  is the mean of the  $(n-k)$  lowest  $r$ 's, i.e.

$$\bar{r}_{n-k} = (r_{(1)} + \dots + r_{(n-k)})/(n - k).$$

*Tabulated significance levels:* Table XV, page 313, extracted from Tietjen and Moore (1972), Table II, pages 591–593, where 1, 5, and 10 per cent points are given for  $k = 1(1)10$  and  $n = [\max(3, 2k)](1)20(5)50$ . Two erroneous entries in Tietjen and Moore's Table IIa have been amended ( $n = 10$ ,  $k = 3$ ;  $n = 10$ ,  $k = 4$ ).

*Reference:* Tietjen and Moore (1972).

*Properties of test:* A pragmatic test procedure.

**N17 Two-sided test for the presence of an undefined number of discordant values in a normal sample with  $\mu$  and  $\sigma^2$  unknown**

*Test statistic:*

$$T_{N17} = \text{Shapiro and Wilk's } W\text{-statistic} \\ = \left( \sum_{i=1}^{[n/2]} a_{n,n-i+1} [x_{(n-i+1)} - x_{(i)}] \right)^2 / S^2$$

where  $[n/2]$  denotes the integer part of  $n/2$ , and the  $a_{n,j}$  are tabulated constants (Table XVIb, page 315; extracted from Shapiro and Wilk (1965), Table 5, pages 603–604, where values of these constants are given for  $n = 2(1)50$ .

*Test distribution:* For  $n = 3$ , we have:

$$f_3(t) = \frac{3}{\pi} (t - t^2)^{-\frac{1}{2}} \quad (\frac{3}{4} \leq t \leq 1).$$

*Tabulated significance levels:* Table XVIa, page 314; abridged from Shapiro and Wilk (1965), Table 6, page 605, where lower and upper 1, 2, 5, and 10 per cent points, and the 50 per cent point, are given for  $n = 3(1)50$ . Values of  $T_{N17}$  smaller than the tabulated level are significant.

*Further tables:* Shapiro, Wilk, and Chen (1968) give values for the power of the test against 45 different inherent alternatives (see Chapter 2, page 31). Chen (1971) gives values for the power against contamination by a given number of observations (either 1 or 2) in small samples ( $n \leq 10$ ), or with a given contamination probability per observation (0.05, 0.10 or 0.20) in larger samples (up to  $n = 50$ ); he deals with shifts both in location and in dispersion.

*References:* Shapiro and Wilk (1965), Shapiro, Wilk, and Chen (1968), Chen (1971).

*Properties:* A useful omnibus test, both against inherent alternatives and against slippage alternatives.

Nv1 *Discordancy test for a single upper outlier  $x_{(n)}$  in a normal sample with  $\mu$  unknown and an independent estimate of  $\sigma^2$  known*

*Test statistic:*

$$T_{Nv1} = \text{externally studentized extreme deviation from the mean} = \frac{x_{(n)} - \bar{x}}{s_\nu}.$$

*Recurrence relationship:*

$$f_n(t) = \left( \frac{n^3}{(n-1)\pi\nu} \right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left( 1 + \frac{nt^2}{(n-1)\nu} \right)^{-(\nu+1)/2} F_{n-1}\left(\frac{nt}{n-1}\right)$$

with  $F_1(t) = 0$  ( $t < 0$ ),  $1(t > 0)$ .

*Inequality:*

$$SP(t) < nP(t_\nu > [n/(n-1)]^{\frac{1}{2}}t).$$

**Tabulated significance levels:** Table VIIa, page 300; abridged from Pearson and Hartley (1966), Table 26, pages 185–186, where 10, 5, 2.5, 1, 0.5, and 0.1 per cent points are given for  $n = 3(1)10, 12$  and  $\nu = 10(1)20, 24, 30, 40, 60, 120, \infty$ , and further 5 and 1 per cent points for the additional  $\nu$ -values 5(1)9.

**Further tables:** David and Paulson (1965) give graphs of performance measure  $P_2$  (see page 66) for an alternative model of slippage in location by one observation. McMillan (1971), subsequently amended by Moran and McMillan (1973), gives graphs of performance measures  $P(C_1)$ ,  $P(C_2)$ ,  $P(C_3)$  (see page 74) for the consecutive testing of two upper outliers using Nv1.

**References:** Nair (1948, 1952), David (1956a, 1956b), David and Paulson (1965), McMillan (1971), Moran and McMillan (1973).

**Properties of test:** Makes no use of the internal estimate of variance; if there is at most one contaminant this wastes information, and test Nv2 is preferable; on the other hand it offers a safeguard against the risk of masking if there is more than one contaminant.

**Nv2 Discordancy test for a single upper outlier  $x_{(n)}$  in a normal sample with  $\mu$  unknown and an independent estimate of  $\sigma^2$  known**

**Test statistic:**

$T_{Nv2}$  = externally and internally studentized extreme deviation  
from the mean =  $\frac{x_{(n)} - \bar{x}}{\tilde{s}}$ .

**Recurrence relationship:**

$$f_n(t) = \left( \frac{n^3}{\pi(n-1)(n-1+\nu)} \right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{n-1+\nu}{2}\right)}{\Gamma\left(\frac{n-2+\nu}{2}\right)} \left( 1 - \frac{nt^2}{(n-1)(n-1+\nu)} \right)^{(n-4+\nu)/2} \\ \times F_{n-1} \left[ \left( \frac{n^2(n-2+\nu)t^2}{(n-1)^2(n-1+\nu) - n(n-1)t^2} \right)^{\frac{1}{2}} \right],$$

$$t \leq [(n-1)(n-1+\nu)/n]^{\frac{1}{2}},$$

with  $F_1(t) = 0$  ( $t < 0$ ),  $1$  ( $t > 0$ ).

**Inequality:**

$$SP(t) < n P\left(t_{n-2+\nu} > \left[ \frac{n(n-2+\nu)t^2}{(n-1)(n-1+\nu) - nt^2} \right]^{\frac{1}{2}}\right).$$

**Tabulated significance levels:** Table VIIIC, page 302; derived from Quesenberry and David (1961), Tables 1 and 2, page 388, where 5 and 1 per cent points of  $(n-1+\nu)^{-\frac{1}{2}} T_{Nv2}$  are given for  $n = 3(1)10, 12, 15, 20$  and  $\nu = 0(1)10, 12, 15, 20, 24, 30, 40, 50$ .

**Further tables:** David and Paulson (1965) give graphs of performance measure  $P_2$  (see page 66) for an alternative model of slippage in location by one observation. McMillan (1971), with subsequent amendments by Moran and McMillan (1973), gives graphs of performance measures  $P(C_1)$ ,  $P(C_2)$ ,  $P(C_3)$  (see page 74) when Nv2 is used consecutively for testing two upper outliers.

**References:** Kudo (1956a), Quesenberry and David (1961), David and Paulson (1965), McMillan (1971), Moran and McMillan (1973).

**Properties of test:** For a location-slippage alternative in which one observation arises from a normal distribution  $N(\mu + a, \sigma^2)$ ,  $a > 0$ , Nv2 has the optimal property of being the scale- and location-invariant test of given size which maximizes the probability of identifying the contaminant as discordant.

**Nv3 Two-sided discordancy test for an extreme outlier in a normal sample with  $\mu$  unknown and an independent estimate of  $\sigma^2$  known**

**Test statistic:**

$$T_{Nv3} = \text{externally studentized extreme absolute deviation from the mean} \\ = \max\left(\frac{x_{(n)} - \bar{x}}{s_\nu}, \frac{\bar{x} - x_{(1)}}{s_\nu}\right).$$

**Inequality:**

$$SP(t) < 2P(T_{Nv1} > t) < 2nP(t_v > [n/(n-1)]^{1/2}t).$$

**Tabulated significance levels:** Table VIIIb, page 301; derived from Halperin, Greenhouse, Cornfield, and Zalokar (1955), Tables 1 and 2, pages 187–188, where bounds on the 5 and 1 per cent points are given for  $n = 3(1)10, 15, 20, 30, 40, 60$  and  $\nu = 3(1)10, 15, 20, 30, 40, 60, 120, \infty$  subject to  $\nu \geq n$ .

**Reference:** Halperin *et al.* (1955).

**Properties of test:** An appropriate test for comparing treatment means in analysis of variance.

**Nv4 Two-sided discordancy test for an extreme outlier in a normal sample with  $\mu$  unknown and an independent estimate of  $\sigma^2$  known**

**Test statistic:**

$$T_{Nv4} = \text{externally and internally studentized extreme absolute deviation from the mean} \\ = \max\left(\frac{x_{(n)} - \bar{x}}{\tilde{s}}, \frac{\bar{x} - x_{(1)}}{\tilde{s}}\right).$$

*Inequality:*

$$SP(t) < 2P(T_{Nv2} > t) < 2nP\left(t_{n-2+\nu} > \left[\frac{n(n-2+\nu)t^2}{(n-1)(n-1+\nu)-nt^2}\right]^{\frac{1}{2}}\right).$$

*Tabulated significance levels:* Table VIIId, page 303; derived from Quesenberry and David (1961), Tables 3 and 4, pages 389–390, where bounds on the 5 and 1 per cent points of  $(n-1+\nu)^{-\frac{1}{2}}T_{Nv4}$  are given for  $n = 3(1)10, 12, 15, 20$  and  $\nu = 0(1)10, 12, 15, 20, 24, 30, 40, 50$ . An error in Quesenberry and David's Table 3 (the entry for  $n = 7, \nu = 4$ ) has been corrected.

*References:* Kudo (1956a), Quesenberry and David (1961).

*Properties of test:* For a location-slippage alternative in which one observation arises from a normal distribution  $N(\mu + a, \sigma^2)$ ,  $a \neq 0$ , Nv4 has the optimal property of being the scale- and location-invariant test of given size which maximizes the probability of identifying the contaminant as discordant.

**Nv5** *Discordancy test for  $k (\geq 2)$  upper outliers  $x_{(n-k+1)}, \dots, x_{(n-1)}, x_{(n)}$  in a normal sample with  $\mu$  unknown and an independent estimate of  $\sigma^2$  known*

*Test statistic:*

$T_{Nv5}$  = sum of jointly (externally and internally) studentized deviations from the mean

$$= \frac{x_{(n-k+1)} + \dots + x_{(n-1)} + x_{(n)} - k\bar{x}}{\tilde{s}}.$$

*Inequality:*

$$SP(t) \leq \binom{n}{k} P\left(t_{n-2+\nu} > \left[\frac{n(n-2+\nu)t^2}{k(n-k)(n-1+\nu)-nt^2}\right]^{\frac{1}{2}}\right).$$

This is an equality when  $t \geq [k^2(n-1+\nu)(n-k-1)/(nk+n)]^{\frac{1}{2}}$ .

*References:* McMillan (1971), Fieller (1976).

*Properties of test:* As for N3.

**Nv6** *Discordancy test for a lower and upper outlier-pair  $x_{(1)}, x_{(n)}$  in a normal sample with  $\mu$  unknown and an independent estimate of  $\sigma^2$  known*

*Test statistic:*

$$T_{Nv6} = \text{externally studentized range} = \frac{x_{(n)} - x_{(1)}}{s_\nu}.$$

*Inequality:*

$$SP(t) < n(n-1)P(t_\nu > t\sqrt{2}).$$

**Tabulated significance levels:** Table XIc, pp. 308–309; abridged from Pearson and Hartley (1966), Table 29, pages 191–193, where 10, 5, and 1 per cent points are given for  $n = 2(1)20$  and  $\nu = 1(1)20, 24, 30, 40, 60, 120, \infty$ .

**Further tables:** Harter (1969a) gives upper and lower 0.1, 0.5, 1, 2.5, 5, 10(10)40 per cent points and the median, and extends the sample sizes to  $n = 22(2)40(10)100$ .

**References:** Dixon (1950), Thompson (1955), Moore (1957), David (1962), Harter (1969a), Fieller (1976).

**Nv7 Discordancy test for a lower and upper outlier-pair  $x_{(1)}, x_{(n)}$  in a normal sample with  $\mu$  unknown and an independent estimate of  $\sigma^2$  known**

**Test statistic:**

$$T_{Nv7} = \text{externally and internally studentized range} = \frac{x_{(n)} - x_{(1)}}{\bar{s}},$$

**Inequality:**

$$SP(t) \leq n(n-1)P(t_{n-2+\nu} > [(n-2+\nu)t^2/(2n-2+2\nu-t^2)]^{\frac{1}{2}}).$$

This is an equality when  $t \geq [\frac{3}{2}(n-1+\nu)]^{\frac{1}{2}}$ .

**Reference:** Fieller (1976).

**Nμ1 Discordancy test for a single upper outlier  $x_{(n)}$  in a normal sample with  $\mu$  known and  $\sigma^2$  unknown**

**Test statistic:**

$$T_{N\mu 1} = \frac{x_{(n)} - \mu}{s(\mu)}.$$

**Recurrence relationship:**

$$f_n(t) = \left(\frac{n}{\pi}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \left(1 - \frac{t^2}{n}\right)^{\frac{n-3}{2}} F_{n-1}\left(\left[\frac{(n-1)t^2}{n-t^2}\right]^{\frac{1}{2}}\right) \quad (-\sqrt{n} \leq t \leq \sqrt{n})$$

with  $F_1(t) = 0$  ( $t < -1$ ),  $\frac{1}{2}$  ( $-1 < t < 1$ ), 1 ( $t > 1$ ).

**Tabulated significance levels:** Table VIIc, page 298; freshly compiled on the basis of simulations of sizes 10 000.

**Properties of test:** Note that  $T_{N\mu 1}$  can take negative values. The statistic  $S_n^2(\mu)/S^2(\mu) = 1 - (1/n)T_{N\mu 1}^2$  is therefore not equivalent to  $T_{N\mu 1}$ , in contradistinction to the one-one relationship between  $S_n^2/S^2$  and  $T_{N1}$ . The occurrence of a negative value for  $T_{N\mu 1}$  has probability  $1/2^n$  on the working hypothesis and is therefore rare except for very small samples.

N $\mu$ 2 *Two-sided discordancy test for an extreme outlier in a normal sample with  $\mu$  known and  $\sigma^2$  unknown*

*Test statistic:*

$$T_{N\mu 2} = \max\left(\frac{|x_{(n)} - \mu|}{s(\mu)}, \frac{|\mu - x_{(1)}|}{s(\mu)}\right).$$

*Recurrence relationship:*

$$f_n(t) = nb_{\frac{1}{2},(n-1)/2}(t)F_{n-1}[t/(1-t)] \quad (0 \leq t \leq 1).$$

*Inequality:*

$$SP(t) \leq 2nP(t_{n-1} > [(n-1)t/(1-t)]^{\frac{1}{2}}).$$

This is an equality when  $t \geq \frac{1}{2}$ .

*Tabulated significance levels:* Table VIId, page 298; derived from Eisenhart, Hastay, and Wallis (1947), Tables 15.1 and 15.2, pages 390–391, where 5 and 1 per cent points of  $(T_{N\mu 2})^2/n$  are given (as part of a larger table) for  $n = 2(1)10, 12, 15, 20, 24, 30, 40, 60, 120$ .

*References:* Cochran (1941), Fieller (1976), Lewis and Fieller (1978).

*Properties of test:* Maximum likelihood ratio test when the alternative is that one observation arises from a normal distribution  $N(\mu, b\sigma^2)$ ,  $b > 1$ . Note that  $(T_{N\mu 2})^2/n$  has the same distribution as  $T_{Ga1}$  with  $r = \frac{1}{2}$ .

N $\mu$ 3 *Discordancy test for  $k (\geq 2)$  upper outliers  $x_{(n-k+1)}, \dots, x_{(n)}$  in a normal sample with  $\mu$  known and  $\sigma^2$  unknown*

*Test statistic:*

$$T_{N\mu 3} = \frac{x_{(n-k+1)} + \dots + x_{(n-1)} + x_{(n)} - k\mu}{s(\mu)}.$$

*Inequality:*

$$SP(t) \leq \binom{n}{k} P(t_{n-1} > [(n-1)t^2/(nk - t^2)]^{\frac{1}{2}}).$$

*Tabulated significance levels:* Table IXc, page 305; freshly compiled on the basis of simulations of sizes 10 000.

N $\mu$ 4 *Discordancy test for two upper outliers  $x_{(n-1)}, x_{(n)}$  in a normal sample with  $\mu$  known and  $\sigma^2$  unknown*

*Test statistic:*

$$T_{N\mu 4} = S_{n-1,n}^2(\mu) / S^2(\mu).$$

*Tabulated significance levels:* Table IXd, page 305; freshly compiled on the basis of simulations of sizes 10 000. Values of  $T_{N\mu 4}$  smaller than the tabulated level are significant.

**N $\mu$ 5** Discordancy test for a lower and upper outlier-pair  $x_{(1)}, x_{(n)}$  in a normal sample with  $\mu$  known and  $\sigma^2$  unknown

Test statistic:

$$T_{N\mu 5} = S_{1,n}^2(\mu) / S^2(\mu).$$

*Tabulated significance levels:* Table Xb, page 306; freshly compiled on the basis of simulations of sizes 10 000. Values of  $T_{N\mu 5}$  smaller than the tabulated level are significant.

**N $\mu$ 6** Discordancy test for one or more outliers (irrespective of their directions) in a normal sample with  $\mu$  known and  $\sigma^2$  unknown

Test statistic:

$$T_{N\mu 6} = \text{sample kurtosis based on deviations from } \mu$$

$$= \frac{\sum_{j=1}^n (x_j - \mu)^4}{ns^4(\mu)}.$$

The value tested for discordancy is  $x_{(n)}$  or  $x_{(1)}$ , according as  $|x_{(n)} - \mu|$  is greater or less than  $|\mu - x_{(1)}|$ . The presence of a discordant value is indicated by a *high* value of the test statistic. For more than one outlier, apply test consecutively.

*Tabulated significance levels:* Table XIVc, page 313; freshly compiled on the basis of simulations of sizes 10 000.

*Reference:* Ferguson (1961a).

*Properties of test:* N $\mu$ 6 is the locally best invariant test of given size against a location-slippage alternative in which  $k$  of the  $n$  observations arise from separate normal distributions  $N(\mu + a_1, \sigma^2), \dots, N(\mu + a_k, \sigma^2)$ ,  $a_1 \neq 0, \dots, a_k \neq 0$ , provided that the contamination proportion  $k/n$  under the alternative hypothesis is less than  $\frac{1}{3}$  (strictly, provided that  $k < \frac{1}{3}(n+2)$ ).

The test is suitable for consecutive use in the possible presence of more than one outlier.

**N $\mu$ 14** Discordancy test for  $k$  outliers (irrespective of directions) in a normal sample with  $\mu$  known and  $\sigma^2$  unknown

Test statistic:

$$T_{N\mu 14} = \frac{\sum_{j=n-k+1}^n d_{(j)}^2}{S^2(\mu)}$$

where  $d_j = |x_j - \mu|$ , the absolute deviation of  $x_j$  from the population mean; and  $\{d_{(j)}\}$  are the values of the  $d_j$  in ascending order,  $d_{(1)} < d_{(2)} < \dots < d_{(n)}$ .

*Inequality:*

$$SP(t) < \binom{n}{k} P\left(F_{k,n-k} > \frac{(n-k)t}{k(1-t)}\right).$$

*References:* Fieller (1976), Lewis and Fieller (1978).

*Properties of test:*  $N\mu 14$  is the maximum likelihood ratio test for an alternative in which  $k$  of the  $n$  observations arise from a common normal distribution  $\mathbf{N}(\mu, b\sigma^2)$ ,  $b > 1$ .

The test statistic  $T_{N\mu 14}$  has the same distribution as  $T_{G_{a5}}$  with  $r = \frac{1}{2}$ .

**No1** *Discordancy test for a single upper outlier  $x_{(n)}$  in a normal sample with  $\sigma^2$  known and  $\mu$  unknown*

*Test statistic:*

$$T_{N\sigma 1} = \text{standardized extreme deviation from the mean} = \frac{x_{(n)} - \bar{x}}{\sigma}.$$

*Test distribution:*

$$F_n(t) = \exp\left\{-\frac{1}{2n} \frac{d^2}{dt^2}\right\} [\Phi(t)]^n.$$

*Recurrence relationship:*

$$f_n(t) = \left(\frac{n^3}{2\pi(n-1)}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2} \frac{nt^2}{n-1}\right) F_{n-1}\left(\frac{nt}{n-1}\right),$$

with  $f_2(t) = \frac{2}{\sqrt{\pi}} \exp(-t^2)$ .

*Inequality:*

$$SP(t) < n\Phi[-n^{\frac{1}{2}}t/(n-1)^{\frac{1}{2}}].$$

*Tabulated significance levels:* Table VIIe, page 299; abridged from Grubbs (1950), Table III, page 45, where 10, 5, 1, and 0.5 per cent points are given for  $n = 2(1)25$ .

*Further tables:* A table in Nair (1948), also in Pearson and Hartley (1966), gives lower and upper 10, 5, 2.5, 1, 0.5, and 0.1 per cent points for  $n = 3(1)9$ . Dixon (1950) gives graphs of performance measure  $P_3$ . David (1956) gives a table of values of power  $P_1$  against the alternative that one observation arises from a normal distribution  $\mathbf{N}(\mu + a, \sigma^2)$ ,  $a > 0$ , for  $a/\sigma = 1(1)4$  and  $n = 3(1)10, 12, 15, 20, 25$ . David and Paulson (1965) correct some errors in this table. McMillan and David (1971) give graphs of performance measures  $P(C_1)$ ,  $P(C_2)$ ,  $P(C_3)$  (page 74) for the consecutive use of No1 in testing two upper outliers.

**References:** McKay (1935), Nair (1948), Grubbs (1950, 1969), Dixon (1950, 1962), David (1956), Kudo (1956a), Ferguson (1961b), McMillan and David (1971), Fieller (1976).

**Properties of test:**  $N\mu\sigma 1$  is the maximum likelihood ratio test for the above-stated alternative of slippage in location by  $a > 0$  for one observation. For this alternative, it has the optimal property of being the scale- and location-invariant test of given size which maximizes the probability  $P_3$  of identifying the contaminant as discordant. For the same alternative, some typical values of power  $P_1$  are as shown in Table 3.2 (David and Paulson, 1965).

Unlike test  $N1$ ,  $N\mu\sigma 1$  can be used effectively when there is more than one contaminant, being relatively unaffected by the risk of masking.

**$N\mu\sigma 1$  Discordancy test for a single upper outlier  $x_{(n)}$  in a normal sample with both  $\mu$  and  $\sigma^2$  known**

**Test statistic:**

$T_{N\mu\sigma 1}$  = standardized extreme deviation from the population mean

$$= \frac{x_{(n)} - \mu}{\sigma}.$$

**Test distribution:**

$$F_n(t) = [\Phi(t)]^n.$$

**Tabulated significance levels:** Table VIIg, page 299; the entries for sample sizes up to  $n = 30$  are extracted from Pearson and Hartley (1966), Table 24, page 184, where lower and upper 10, 5, 2.5, 1, 0.5, and 0.1 per cent points are given for  $n = 1(1)30$ ; the entries for  $n > 30$  have been freshly compiled.

**Reference:** Dixon (1962).

**Properties of test:** Maximum likelihood ratio test when the alternative is that one observation arises from a normal distribution  $N(\mu + a, \sigma^2)$ ,  $a > 0$ .

Table 3.2

	$a/\sigma$	2	3	4	5
$n \backslash$					
Test at 5% significance level	3	0.31	0.63	0.87	0.98
	10	0.27	0.62	0.89	0.99
	25	0.21	0.54	0.86	0.98
Test at 1% significance level	3	0.14	0.40	0.71	0.92
	10	0.12	0.41	0.76	0.95
	25	0.09	0.35	0.72	0.94

**N $\sigma$ 2** Two-sided discordancy test for an extreme outlier in a normal sample with  $\sigma^2$  known and  $\mu$  unknown

*Test statistic:*

$$T_{N\sigma 2} = \max \left( \frac{x_{(n)} - \bar{x}}{\sigma}, \frac{\bar{x} - x_{(1)}}{\sigma} \right).$$

*Inequality:*

$$SP(t) < 2P(T_{N\sigma 1} > t).$$

*Tabulated significance levels:* Table VIIIf, page 299; extracted from Halperin, Greenhouse, Cornfield and Zalokar (1955), Tables 1 and 2, pages 187–188 ( $\nu = \infty$ ; see Worksheet Nv3).

*Reference:* Kudo (1956a).

*Properties of test:* As for N2, except that N $\sigma$ 2 (unlike N2) is relatively unaffected by masking from other outliers.

**N $\sigma$ 3** Discordancy test for  $k (\geq 2)$  upper outliers  $x_{(n-k+1)}, \dots, x_{(n-1)}, x_{(n)}$  in a normal sample with  $\sigma^2$  known and  $\mu$  unknown

*Test statistic:*

$$T_{N\sigma 3} = \text{sum of standardized deviations from the mean} \\ = \frac{x_{(n-k+1)} + \dots + x_{(n-1)} + x_{(n)} - k\bar{x}}{\sigma}.$$

*Inequality:*

$$SP(t) < \binom{n}{k} \Phi[-n^{1/2}t/(kn - k^2)^{1/2}].$$

*Tabulated significance levels:* Table IXE, page 305; values for  $k = 2, n \leq 20$  extracted from McMillan and David (1971), Table 1, page 82, where 5 and 1 per cent points are given for  $n = 4(1)27$ ; values for  $k = 2, n \geq 30$  and for  $k = 3$  and  $k = 4$  freshly compiled on the basis of simulations of sizes 10 000.

*Further tables:* McMillan and David (1971) give graphs of the performance measure  $P_2$  for the case  $k = 2$ .

*References:* Kudo (1956a), McMillan and David (1971), Fieller (1976).

*Properties of test:* As for N3.

**N $\mu\sigma$ 3** Discordancy test for  $k (\geq 2)$  upper outliers  $x_{(n-k+1)}, \dots, x_{(n-1)}, x_{(n)}$  in a normal sample with both  $\mu$  and  $\sigma^2$  known

*Test statistic:*

$$T_{N\mu\sigma 3} = \text{sum of standardized deviations from the population mean} \\ = \frac{x_{(n-k+1)} + \dots + x_{(n-1)} + x_{(n)} - k\mu}{\sigma}.$$

*Tabulated significance levels:* Table IXg, page 306; freshly compiled on the basis of simulations of sizes 10 000.

**N $\sigma$ 4** *Discordancy test for two upper outliers  $x_{(n-1)}, x_{(n)}$  in a normal sample with  $\sigma^2$  known and  $\mu$  unknown*

*Test statistic:*

$$T_{N\sigma 4} = S_{n-1,n}^2 / \sigma^2.$$

*Tabulated significance levels:* Table IXf, page 305; freshly compiled on the basis of simulations of sizes 10 000. Values of  $T_{N\sigma 4}$  smaller than the tabulated level are significant.

**N $\mu\sigma$ 4** *Discordancy test for two upper outliers  $x_{(n-1)}, x_{(n)}$  in a normal sample with both  $\mu$  and  $\sigma^2$  known*

*Test statistic:*

$$T_{N\mu\sigma 4} = S_{n-1,n}^2(\mu) / \sigma^2.$$

*Tabulated significance levels:* Table IXh, page 306; freshly compiled on the basis of simulations of sizes 10 000. Values of  $T_{N\mu\sigma 4}$  smaller than the tabulated level are significant.

**N $\sigma$ 5** *Discordancy test for a lower and upper outlier-pair  $x_{(1)}, x_{(n)}$  in a normal sample with  $\sigma^2$  known and  $\mu$  unknown.*

*Test statistic:*

$$T_{N\sigma 5} = S_{1,n}^2 / \sigma^2.$$

*Tabulated significance levels:* Table Xc, page 306; freshly compiled on the basis of simulations of sizes 10 000. Values of  $T_{N\sigma 5}$  smaller than the tabulated level are significant.

**N $\mu\sigma$ 5** *Discordancy test for a lower and upper outlier-pair  $x_{(1)}, x_{(n)}$  in a normal sample with both  $\mu$  and  $\sigma^2$  known*

*Test statistic:*

$$T_{N\mu\sigma 5} = S_{1,n}^2(\mu) / \sigma^2.$$

*Tabulated significance levels:* Table Xd, page 306; freshly compiled on the basis of simulations of sizes 10 000. Values of  $T_{N\mu\sigma 5}$  smaller than the tabulated level are significant.

**N $\sigma$ 6(N $\mu\sigma$ 6)** *Discordancy test for a lower and upper outlier-pair  $x_{(1)}, x_{(n)}$  in a normal sample with  $\sigma^2$  known*

*Test statistic:*

$$T_{N\sigma 6} = \text{standardized range} = \frac{x_{(n)} - x_{(1)}}{\sigma}.$$

**Test distribution:**

$$F_n(t) = n \int_{-\infty}^{\infty} \phi(x)[\Phi(x) - \Phi(x-t)]^{n-1} dx.$$

**Tabulated significance levels:** Table XIb, page 307; abridged from Harter (1969a), Table A7, pages 372–374, where lower and upper 0.01, 0.05, 0.1, 0.5, 1, 2.5, 5, 10(10)40 per cent points and the 50 per cent point are given for  $n = 2(1)20(2)40(10)100$ .

**Further tables:** Dixon (1950) gives graphs of performance measure  $P_3$ .

**References:** Tippett (1925), Pearson (1926, 1932), Pearson and Hartley (1942), Dixon (1950, 1962), Harter (1969a).

**N $\sigma$ 7(N $\mu\sigma$ 7) Discordancy test for a single upper outlier  $x_{(n)}$  in a normal sample with  $\sigma^2$  known**

**Test statistic:**

$$T_{N\sigma 7} = \frac{x_{(n)} - x_{(n-1)}}{\sigma}.$$

**Test distribution:**

$$F_n(t) = 1 - n \int_{-\infty}^{\infty} \phi(x+t)[\Phi(x)]^{n-1} dx.$$

**Tabulated significance levels:** Table XIIIh, page 312; derived from Irwin (1925), Table II, page 239, where values of  $F_n(t)$  are given for  $t = 0.1(0.1)5.0$  and  $n = 2, 3, 10(10)100(100)1000$ .

**Further tables:** Dixon (1950) gives graphs of performance measure  $P_3$ .

**References:** Irwin (1925), Dixon (1950).

**Properties of test:** Analogous in concept to a Dixon-type test. Performance  $P_3$  is comparable to that of test N $\sigma$ 1 if there is just one contaminant, but compares unfavourably with N $\sigma$ 1 if there is more than one contaminant, owing to incidence of masking by  $x_{(n-1)}$ . However, N $\sigma$ 7 could be a useful test for a *lower* outlier (with test statistic  $[x_{(2)} - x_{(1)}]/\sigma$ ) in a life-test data situation in which for practical reasons only the shortest lifetimes were actually observed.

**N $\sigma$ 8(N $\mu\sigma$ 8) Discordancy test for an upper outlier-pair  $x_{(n-1)}, x_{(n)}$  in a normal sample with  $\sigma^2$  known**

**Test statistic:**

$$T_{N\sigma 8} = \frac{x_{(n-1)} - x_{(n-2)}}{\sigma}.$$

*Test distribution:*

$$F_n(t) = 1 - n(n-1) \int_{-\infty}^{\infty} \phi(x+t)[1-\Phi(x+t)][\Phi(x)]^{n-2} dx.$$

*Tabulated significance levels:* Table XIIIi, page 312; derived from Irwin (1925), Table III, page 242, where values of  $F_n(t)$  are given for  $t = 0.1(0.1)2.0$  and  $n = 3, 10(10)100(100)1000$ , and also for  $t = 2.1(0.1)4.0$  in the case  $n = 3$ .

*Reference:* Irwin (1925).

*Properties of test:* Analogous to a Dixon-type test. As with  $N\sigma 7$ , could be an appropriate test in a life-testing context.  $N\sigma 7$  and  $N\sigma 8$  have a historical standing as being two of the earliest published tests in modern outlier methodology.

**$N\sigma 9(N\mu\sigma 9)$  Discordancy test for  $k$  lower and  $k$  upper outliers ( $k \geq 2$ ) in a normal sample with  $\sigma^2$  known**

*Test statistic:*

$$T_{N\sigma 9} = (k-1)\text{th standardized quasi-range} = \frac{x_{(n-k+1)} - x_{(k)}}{\sigma}.$$

*Tabulated significance levels:* Table XII, page 310; abridged from Harter (1969b), Table A7, pages 295–319, where lower and upper 0.01, 0.05, 0.1, 0.5, 1, 2.5, 5, 10(10)40 per cent points and the 50 per cent point are given for  $k = 1(1)9$  and  $n = 2k(1)20(2)40(10)100$ .

*Reference:* Harter (1969b).

*Properties of test:* Advantage is taken here of Harter's very extensive tables of quasi-ranges to provide a test for discordancy in the tails of a large sample whose main central mass can be assumed normal.

### 3.4.4 Discordancy tests for samples from other distributions

Many of the discordancy tests given in Section 3.4.2 for exponential samples can be used for samples from Pareto distributions and from distributions of asymptotic extreme-value type (i.e. Gumbel, Fréchet, and Weibull distributions), by simple transformation of the data. Likewise, various discordancy tests given in Section 3.4.3 for normal samples can be used for samples from log-normal distributions. Since a square root transformation converts a Poisson random variable into a variable distributed approximately normally with variance  $\frac{1}{4}$ , whatever the value of the Poisson mean provided it is not too small, the  $N\sigma$  tests in Section 3.4.3 can be used for samples from Poisson distributions. Similarly, the  $N\sigma$  tests can be used for samples from binomial distributions.

Apart from the use of transformations, a few specific discordancy tests are available for Poisson and binomial samples and for samples from other distributions.

Details for the various distributions are given below as follows:

	page
Discordancy tests for Pareto samples	116
Discordancy tests for Gumbel, Fréchet, and Weibull samples	117
Discordancy tests for log-normal samples	118
Discordancy tests for uniform samples	118
Discordancy tests for Poisson samples	120
Discordancy tests for binomial samples	122
Discordancy tests for truncated exponential samples	124

### *Discordancy tests for Pareto samples*

In addition to its well known role in economics as a model for the distribution of incomes, the Pareto distribution can be used as a pragmatic model for other skew-distributed data characterized by a main mass of low values at one end and a gradation to a long tail of infrequently occurring high values at the other. In data of this nature, high-valued outliers requiring test may well arise.

A Pareto random variable  $X$  is characterized by two parameters, the minimum value  $a$  ( $a > 0$ ), and the shape parameter  $r$  ( $r > 0$ ). Its distribution function can be written

$$P(X < x) = 0 \quad (x \leq a), \quad 1 - (a/x)^r \quad (x \geq a).$$

If  $Y = \ln X$ , the distribution function of  $Y$  is

$$P(Y < y) = 0 \quad (y \leq \ln a), \quad 1 - \exp[-r(y - \ln a)] \quad (y \geq \ln a),$$

i.e.  $Y$  has an exponential distribution with origin  $\ln a$  and scale parameter  $1/r$ . Suppose then that we have, on the working hypothesis, a Pareto sample  $x_{(1)}, \dots, x_{(n)}$ , containing one or more outliers. The transformed quantities  $\ln x_{(1)}, \dots, \ln x_{(n)}$  will also be in ascending order, and so can be written  $y_{(1)}, \dots, y_{(n)}$ , an ordered sample from the exponential  $Y$ -distribution. If  $a$  and  $r$  are both unknown, the outlying value or values in the  $x$ -sample can be tested for discordancy by applying to the corresponding  $y$ -values the appropriate test for an exponential sample with unknown origin. The available tests given in Section 3.4.2 are the Dixon-type tests E2, E4, E6, E8, E10, E11.

If  $a$  is known, the transformation  $Z = \ln(X/a)$  should be used. On the working hypothesis the transformed quantities  $z_{(1)} = \ln[x_{(1)}/a], \dots, z_{(n)} = \ln[x_{(n)}/a]$  belong to an exponential distribution with origin at 0 (and density  $re^{-rz}$  ( $z > 0$ )). All the G and E tests listed in Section 3.4.2 are applicable to the transformed sample.

### Discordancy tests for Gumbel, Fréchet, and Weibull samples

The asymptotic extreme-value distributions of the first, second, and third types, in other words the Gumbel, Fréchet, and Weibull distributions, are well known as models for extreme observations such as annual maximum wind speeds, floods (as greatest-value phenomena), endurance limits in fatigue testing (as smallest-value phenomena), annual minimum temperatures, oldest ages of individuals in a population, and shortest lives of manufactured items. In analysing extreme-value data it is obviously important to remove where possible the biasing effect of any contaminant values which may be present, and the testing of outliers for discordancy is of particular relevance.

The Gumbel distribution depends on a location parameter  $a$  and a positive scale parameter  $b$ ; the Fréchet and Weibull distributions depend on these two parameters and also on a positive shape parameter  $r$ . In terms of these parameters their distribution functions  $P(X < x)$  are given in Table 3.3. Note that each distribution has two forms, according as it relates to greatest-value or smallest-value extremes. It can be shown that if the shape parameter is reparametrized as  $\lambda = 1/r$  (Weibull),  $\lambda = -1/r$  (Fréchet), the Gumbel distribution corresponds to the limiting case  $\lambda \rightarrow 0$ .

If  $X$  has a Gumbel greatest-value distribution, the transformed random variable  $Y = \exp(-X/b)$  has an exponential distribution with origin 0 and scale parameter  $\exp(-a/b)$ . If we know the value of  $b$  we can test the discordancy of an outlier or a set of outliers in a sample from the  $X$ -distribution by transforming each observed value  $x_i$  to  $y_i = \exp(-x_i/b)$  and using on the  $y$ 's a discordancy test for an exponential sample with origin 0. Note that with this particular transformation an upper outlier  $x_{(n)}$  in the  $x$ -sample converts to a lower outlier  $y_{(1)}$  in the  $y$ -sample, so the test on the  $y$ -values must be chosen accordingly.

Corresponding transformations to the one just described are available for the Gumbel smallest-value distribution and the various Fréchet and Weibull distributions. In each case, the  $Y$ -distribution will have a density of the form  $(1/\lambda)\exp(-y/\lambda)$  ( $y > 0$ ) on the working hypothesis. All the tests listed in Section 3.4.2 are applicable to a sample from such a distribution.

Table 3.3

	Distribution of greatest values	Distribution of smallest values
Gumbel	$\exp[-e^{-(x-a)/b}], -\infty < x < \infty$	$1 - \exp[-e^{-(a-x)/b}], -\infty < x < \infty$
Fréchet	$\exp\left[-\left(\frac{x-a}{b}\right)^r\right], a < x$	$1 - \exp\left[-\left(\frac{a-x}{b}\right)^r\right], x < a$
Weibull	$\exp\left[-\left(\frac{a-x}{b}\right)^r\right], x < a$	$1 - \exp\left[-\left(\frac{x-a}{b}\right)^r\right], a < x$

Table 3.4

X-distribution	Required knowledge of parameters	Transformation to be applied	Scale parameter $\lambda$ of Y-distribution	$x_{(n)}$ transforms to	$x_{(1)}$ transforms to
Gumbel greatest-value	$b$ known	$Y = \exp(-X/b)$	$\exp(-a/b)$	$y_{(1)}$	$y_{(n)}$
Gumbel smallest-value	$b$ known	$Y = \exp(X/b)$	$\exp(a/b)$	$y_{(n)}$	$y_{(1)}$
Fréchet greatest-value	$a, r$ known	$Y = (X - a)^{-r}$	$b^{-r}$	$y_{(1)}$	$y_{(n)}$
Fréchet smallest-value	$a, r$ known	$Y = (a - X)^{-r}$	$b^{-r}$	$y_{(n)}$	$y_{(1)}$
Weibull greatest-value	$a, r$ known	$Y = (a - X)^r$	$b^r$	$y_{(1)}$	$y_{(n)}$
Weibull smallest-value	$a, r$ known	$Y = (X - a)^r$	$b^r$	$y_{(n)}$	$y_{(1)}$

Table 3.4 gives the various transformations in a form for working use. Admittedly their utility is limited by the knowledge of parameter values required before they can be applied, unlike the Pareto case (page 116).

#### *Discordancy tests for log-normal samples*

If  $X$  is a log-normal random variable with parameters  $\mu$  and  $\sigma$ ,  $\ln X$  is  $N(\mu, \sigma^2)$  or equivalently  $\log_{10} X$  is  $N(0.434\mu, (0.434)^2\sigma^2)$ . Thus to test for discordancy any outlier or outliers in a log-normal sample, we need only take the logarithms of the observations (to base e or 10 as convenient) and apply an appropriate normal sample test from those listed in Section 3.4.3 to the transformed sample.

No such facility is available for the generalized three-parameter log-normal distribution, in which  $\ln(X - \xi)$  is  $N(\mu, \sigma^2)$ ,  $\xi$  being an unknown location parameter.

#### *Discordancy tests for uniform samples*

Denote the lower and upper bounds of the uniform distribution by  $a, b$  respectively, so that its density is  $1/(b - a)$  for  $a < x < b$ , 0 otherwise. Given the ordered sample  $x_{(1)} < x_{(2)} < \dots < x_{(n)}$ , it is well known that the  $n + 1$  intervals

$$x_{(1)} - a, x_{(2)} - x_{(1)}, \dots, x_{(n)} - x_{(n-1)}, b - x_{(n)}$$

are distributed as  $n + 1$  independent exponential random variables with a common (arbitrary) scale parameter, conditional upon their sum having the constant value  $b - a$ . For testing outliers, therefore, we can use the fact that the ratio of any two non-intersecting combinations of these intervals will have an  $F$ -distribution on the working hypothesis. This leads at once to a general Dixon-type discordancy test applicable to any combination of lower and upper outliers, the statistic being

$$T_U = \frac{x_{(s)} - x_{(r)}}{x_{(q)} - x_{(p)}}, \quad 1 \leq p \leq r < s \leq q \leq n, \quad s - r < q - p,$$

if  $a$  and  $b$  are unknown. For example, an appropriate statistic to test two upper outliers for discordancy would be

$$\frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}}.$$

If  $a$  is known, it is preferable to use the statistic

$$\frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - a};$$

this can be included as a particular case of  $T_U$  by defining  $x_{(0)} = a$  and extending the range of possible values of  $p$  down to zero. Likewise the value of  $b$ , if known, can be used in these tests by defining  $x_{(n+1)} = b$ ,  $q \leq n+1$ .

The test distribution for  $T_U$  is

$$f_n(t) = b_{s-r, q-p-s+r}(t) \quad (0 \leq t \leq 1).$$

No tabulated significance levels are required, other than those provided in standard  $F$ -tables, since

$$SP(t) = P\left(F_{2(s-r), 2(q-p-s+r)} > \frac{q-p-s+r}{s-r} \frac{t}{1-t}\right).$$

For the important particular case of a single outlier, say an upper one,  $x_{(n)}$ , the test statistic  $\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$  has distribution  $f_n(t) = (n-2)(1-t)^{n-3}$  ( $0 \leq t \leq 1$ ), and

$$SP(t) = (1-t)^{n-2}.$$

*Example:* Maguire, Pearson, and Wynn (1952) give the following table (reproduced by permission of the Biometrika Trustees) of time intervals in days between successive compensable accidents for one shift in a section of a mine. The early interval 23 days is an outlier.

3	1	0	2	4	5	1	2	2	1	3	0	4
23	0	1	3	2	0	1	0	0	8	2	3	8
0	1	3	0	3	4	8	2	1	3	0	2	Total 222
0	0	2	0	5	0	0	0	1	0	4	2	
2	0	1	0	0	0	0	0	8	2	3	0	
3	0	2	0	0	2	2	2	0	0	12	0	
0	0	0	0	2	1	0	8	0	1	10	14	
2	1	0	3	4	1	2	8	0	1	4	1	

There were 98 intervals and thus 99 accidents, occurring at times 0, 3, 26, 26, 26, 28, ..., 209, 210, 214, and 222 days. Assuming that the accidents occurred randomly at a constant average rate, these 99 times can be regarded as points in a Poisson process (or rather an equivalent lattice process, since their values are

rounded to the nearest integer—but this can be ignored in the analysis). On the working hypothesis, therefore, they constitute an ordered sample

$$x_{(1)} = 0, x_{(2)} = 3, x_{(3)} = 26, \dots, x_{(98)} = 214, x_{(99)} = 222,$$

from a uniform distribution with unknown bounds  $a < 0$ ,  $b > 222$ . Let us test  $x_{(1)}$  and  $x_{(2)}$  as a lower outlier-pair, using the test statistic  $T_U = [x_{(3)} - x_{(1)}]/[x_{(99)} - x_{(1)}]$ . The observed value of  $T_U$  is  $t = 26/222 = 0.1171$ , and the significance probability of this value is  $SP(t) = P(F_{4,192} > \frac{96}{2} \frac{0.1171}{0.8829}) = P(F_{4,192} > 6.37)$  which is much smaller than 0.001. We must therefore regard as discordant the outlier-pair  $x_{(1)}, x_{(2)}$ , i.e. the combined interval 3 + 23 days from the first to the third accident. This is a stronger result than the discordancy of the single interval 23 days which it *en passant* implies.

### *Discordancy tests for Poisson samples*

Outlier situations in Poisson samples may arise in any of the numerous practical contexts giving rise to Poisson-distributed data, in particular where the data are counts of events occurring randomly in a given time, or counts of individuals scattered randomly over a given length, area, or volume.

Suppose, to fix ideas, that we wish to test for discordancy an upper outlier  $x_{(n)}$  in a sample from a Poisson distribution whose mean  $\mu$  is unknown. The argument used earlier (page 55) in the case of a gamma sample might suggest  $x_{(n)}/\sum x_i$  as a possible statistic, being of the form  $N/D$  and invariant under change of scale. This statistic cannot be used as it stands, since its null distribution depends on  $\mu$ . However, the null distribution of  $x_{(n)}$  *conditional on the observed value of*  $\sum x_i$  does *not* depend on  $\mu$ , since the distribution of  $x_1, \dots, x_n$  conditional on the observed value of  $\sum x_i$  is multinomial with parameters  $(\sum x_j, \frac{1}{n}, \dots, \frac{1}{n})$ . Using this fact, Doornbos (1966) has shown how to set up a discordancy test for (say)  $x_{(n)}$ , based on its null distribution conditional on  $\sum x_i$ . The table of significance levels, say at 5 per cent, will thus show a critical value of  $x_{(n)}$  corresponding to each pair of entry values  $n, \sum x_i$ . (As always with discrete distributions, the significance probability attaching to each critical integer value will not be exactly 5 per cent.)

Discordancy tests for four outlier situations are given below, namely a single upper outlier, a single lower outlier, an upper outlier-pair, and a lower outlier-pair.

As an alternative to these specific tests we may, as already mentioned, use the fact that the transform  $\sqrt{(X + \frac{1}{4})}$  of a Poisson random variable  $X$  with mean  $\mu$  is approximately  $N(\sqrt{\mu}, \frac{1}{4})$ , providing that  $\mu$  is not too small, say at least 4 or 5. Thus to test for discordancy any outlier or outliers in a Poisson sample of unknown but not too small mean, we can take the transformed

values  $\sqrt{(x_j + \frac{1}{4})}$  ( $j = 1, \dots, n$ ) and apply an appropriate  $N\sigma$  test with  $\sigma = \frac{1}{2}$  to the transformed sample.

**P1** *Discordancy test for a single upper outlier  $x_{(n)}$  in a Poisson sample of unknown mean*

*Test statistic:*  $T_{P1}$  = outlier  $x_{(n)}$ , tested conditionally on the observed sum of observations  $\sum x_j$ .

*Inequality:*

$$np_1 - \frac{n(n-1)}{2} p_1^2 < SP(t) < np_1$$

where  $p_1 = P\left(B\left(\sum x_j, \frac{1}{n}\right) \geq t\right)$ .

*Tabulated significance levels:* Table XVIIa, page 316; freshly compiled.

*Further tables:* Doornbos (1966), Table I, gives nominal 5 per cent and 1 per cent critical values of  $x_{(n)}$  for  $n = 2(1)10$  and  $\sum x_j = 2(1)25$ , together with the actual levels of significance attaching to each (necessarily discrete) entry. See also Section 5.3.3 and Table XXV.

*Reference:* Doornbos (1966).

**P2** *Discordancy test for a single lower outlier  $x_{(1)}$  in a Poisson sample of unknown mean*

*Test statistic:*  $T_{P2}$  = outlier  $x_{(1)}$ , tested conditionally on the observed sum of observations  $\sum x_j$ .

*Inequality:*

$$np_2 - \frac{n(n-1)}{2} p_2^2 < SP(t) < np_2$$

where  $p_2 = P\left(B\left(\sum x_j, \frac{1}{n}\right) \leq t\right)$ .

*Tabulated significance levels:* Table XVIIb, page 317; freshly compiled.

*Reference:* Doornbos (1966).

**P3** *Discordancy test for an upper outlier-pair  $x_{(n-1)}, x_{(n)}$  in a Poisson sample of unknown mean*

*Test statistic:*  $T_{P3}$  = sum of outliers  $x_{(n-1)} + x_{(n)}$ , tested conditionally on the observed sum of observations  $\sum x_j$ .

*Inequality:*

$$SP(t) < \binom{n}{2} P\left(\mathbf{B}\left(\sum x_j, \frac{2}{n}\right) \geq t\right).$$

*Tabulated significance levels:* Table XVIIIa, page 318; freshly compiled.

*Further tables:* Doornbos (1966), Table III, gives nominal 5 per cent and 1 per cent critical values of  $x_{(n-1)} + x_{(n)}$  for  $n = 4(1)10$  and  $\sum x_j = 5(1)25$ .

*Reference:* Doornbos (1966).

#### P4 *Discordancy test for a lower outlier-pair $x_{(1)}, x_{(2)}$ in a Poisson sample of unknown mean*

*Test statistic:*  $T_{P4}$  = sum of outliers  $x_{(1)} + x_{(2)}$ , tested conditionally on the observed sum of observations  $\sum x_j$ .

*Inequality:*

$$SP(t) < \binom{n}{2} P\left(\mathbf{B}\left(\sum x_j, \frac{2}{n}\right) \leq t\right).$$

*Tabulated significance levels:* Table XVIIIb, page 319; freshly compiled.

*Further tables:* Doornbos (1966), Table IV, gives nominal 5 per cent and 1 per cent critical values of  $x_{(1)} + x_{(2)}$  for the following cases:  $n = 4$  and  $n = 5$ ,  $\sum x_j = 4(1)25$ ;  $n = 6$ ,  $\sum x_j = 14(1)25$ ;  $n = 7$  and  $n = 8$ ,  $\sum x_j = 17(1)25$ .

*Reference:* Doornbos (1966).

#### *Discordancy tests for binomial samples*

Suppose we have a sample  $x_1, \dots, x_n$  in which each observation  $x_j$  is a value from a binomial distribution  $\mathbf{B}(m, p_j)$  with  $m$  known,  $p_j$  unknown. On the working hypothesis the  $p_j$  are all equal, say to  $p$  (unknown). Data of this kind could arise, for instance, in sampling inspection of mass-produced items, where successive samples each of  $m$  items are inspected and  $x_j$  is the number of defectives found on the  $j$ th occasion; or again in experiments to compare, say, the germination rates of seeds under different conditions,  $m$  seeds being tested under each condition, with replication of the control.

The null distribution of a statistic such as  $x_{(n)}/\sum x_j$  depends on the unknown  $p$ , but the null distribution of  $x_1, \dots, x_n$  conditional on the observed value of  $\sum x_j$  does not depend on  $p$ , being multihypergeometric with parameters  $(nm; \sum x_j; m, \dots, m)$ . Therefore, as with Poisson samples discussed above (see page 120), discordancy tests for outliers are carried out conditionally on the observed value of  $\sum x_j$ . In principle, the three quantities  $n$ ,  $m$ , and  $\sum x_j$  are needed for entering any table of significance levels of (say)  $x_{(n)}$ .

Discordancy tests are given below for an upper outlier and a lower outlier; these are in effect the same test. For convenience, the table of significance levels that we give for this test is entered with  $n$ ,  $m$ , and  $m - x_{(n)}$ , the tabulated quantity being  $\sum x_j$ .

Once again, transformation is available as an alternative to these specific discordancy tests. For a binomial random variable  $X$  with parameters  $m$ ,  $p$  the transform  $\sin^{-1}[(X/m)^{\frac{1}{2}}]$  is distributed approximately  $N(\sin^{-1}(p^{\frac{1}{2}}), 1/(4m))$ . Thus to test for discordancy any outlier or outliers in a binomial sample with known  $m$  but unknown  $p$ , apply an appropriate  $N\sigma$  test with  $\sigma = 1/(2m^{\frac{1}{2}})$  to the sample of transformed values  $\sin^{-1}[(x_j, m)^{\frac{1}{2}}]$ .

### B1 Discordancy test for a single upper outlier $x_{(n)}$ in a binomial sample of unknown probability parameter

*Test statistic:*  $T_{B1}$  = outlier  $x_{(n)}$ , tested conditionally on the observed sum of observations  $\sum x_j$ .

*Inequality:*

$$np_1 - \frac{n(n-1)}{2} p_1^2 < SP(t) < np_1$$

where  $p_1 = P(H(nm; \sum x_j, m) \geq t)$ .

*Tabulated significance levels:* Table XIX, pp. 320–322; freshly compiled.

*Reference:* Doornbos (1966).

*Example:* See the example in Worksheet B2.

### B2 Discordancy test for a single lower outlier $x_{(1)}$ in a binomial sample of unknown probability parameter

*Test statistic:*  $T_{B2}$  = outlier  $x_{(1)}$ , tested conditionally on the observed sum of observations  $\sum x_j$ .

*Inequality:*

$$np_2 - \frac{n(n-1)}{2} p_2^2 < SP(t) < np_2$$

where  $p_2 = P(H(nm; \sum x_j, m) \leq t)$ .

*Tabulated significance levels:* Table XIX, pp. 320–322. This is the table for test B1 (upper outlier); to use it for test B2, replace the given binomial sample  $x_1, \dots, x_n$  with lower outlier  $x_{(1)}$  by the complementary binomial sample  $y_1 = m - x_1, \dots, y_n = m - x_n$ , and apply test B1 to the *upper* outlier  $y_{(n)} = m - x_{(1)}$ .

*Reference:* Doornbos (1966).

**Example:** Suppose that, in the inspection of quality of a manufactured item, five specimens are selected randomly from each of ten batches and tested to destruction, with the following results:

Batch	A	B	C	D	E	F	G	H	I	J
Number of good items out of five	5	4	4	5	4	1	4	5	3	4

Can one take it as obvious that batch F is out of line with the others?

On the basic model, we have a sample of ten from a distribution  $B(m, p)$  where  $m = 5$  and  $p$  is unknown. We wish to test the lower outlier  $x_6 = x_{(1)} = 1$  for discordancy. To convert into an upper outlier test, we consider instead the numbers of failures  $y_1, \dots, y_{10} = 0, 1, 1, 0, 1, 4, 1, 0, 2, 1$ . The upper outlier  $y_{(10)} = 4 = m - 1$ . Entering Table XIX with  $n = 10$  and  $m = 5$ , we see that this outlier would be significant at 1 per cent if  $\sum y_j$  were 7 or less, and would be significant at 5 per cent if  $\sum y_j$  were 10 or less. In fact  $\sum y_j = 11$ , so the evidence for regarding batch F as discordant is weak.

### *Discordancy tests for truncated exponential samples*

If the exponential distribution with density  $(1/\lambda)exp(-x/\lambda)$  ( $x > 0$ ) is truncated at  $x = a$  we get the distribution with density  $(1/\lambda) \times [1 - exp(-a/\lambda)]^{-1} exp(-x/\lambda)$  ( $0 < x < a$ ), 0 ( $x > a$ ). Such a distribution might arise, for example, in life testing data where for practical reasons the test is not allowed to continue for longer than some preassigned duration  $a$ . Upper outlier problems are perhaps not particularly likely to be found with such data, in view of the truncation. Two Dixon-type discordancy tests are available in the literature (Wani and Kabe, 1971), and we give details of these below.

#### **TE1** *Discordancy test for a single upper outlier $x_{(n)}$ in a truncated exponential sample*

*Test statistic:*

$$T_{TE1} = \frac{\text{excess}}{\text{range}} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}.$$

*Test distribution:*

$$f_n(t) = \frac{(n-1)!}{(1-e^{-a})^n} \sum_{j=2}^{n-1} \frac{(-)^{n+j}}{(j-2)!(n-j)!} \left\{ \frac{1}{u_j^2} - \frac{1}{(n-u_j)^2} e^{-na} \right. \\ \left. - \left[ \frac{1}{u_j^2} - \frac{1}{(n-u_j)^2} + \frac{na}{u_j(n-u_j)} \right] e^{-au_j} \right\}$$

where  $u_j = n - j + 1 - (n-j)t$ ,  $(0 \leq t \leq 1)$ .

**Reference:** Wani and Kabe (1971).

**TE2** *Discordancy test for a single lower outlier  $x_{(1)}$  in a truncated exponential sample*

**Test statistic:**

$$T_{\text{TE2}} = \frac{\text{excess}}{\text{range}} = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}.$$

**Test distribution:**

$$f_n(t) = \frac{(n-1)!}{(1-e^{-a})^n} \sum_{j=3}^n \frac{(-)^{n+j}}{(j-3)!(n-j)!} \left\{ \frac{1}{u_j^2} - \frac{1}{(n-u_j)^2} e^{-na} - \left[ \frac{1}{u_j^2} - \frac{1}{(n-u_j)^2} + \frac{na}{u_j(n-u_j)} \right] e^{-au_j} \right\}$$

where  $u_j = n - j + 1 + (j-2)t$ ,  $(0 \leq t \leq 1)$ .

**Reference:** Wani and Kabe (1971).

## CHAPTER 4

# *Accommodation of Outliers in Univariate Samples: Robust Estimation and Testing*

We have outlined in Chapter 2 the need for accommodation of outliers by means of estimation or testing procedures which are *robust* against the presence of outliers—in the phrase of Anscombe and Barron (1966), ‘desensitized to outliers’. ‘What is a robust procedure?’ Huber asks, in his 1972 Wald Lecture *Robust Statistics: A Review*, and goes on to say:

one never has a very accurate knowledge of the true underlying distribution; . . . the performance of some of the classical tests or estimates is very unstable under small changes of the underlying distribution; . . . some alternative tests or estimates . . . lose very little efficiency for an exactly normal law, but show a much better and more stable performance under deviations from it.

While for years one had been concerned mostly with what was later called ‘robustness of validity’ (that the actual confidence levels should be close to, or at least on the safe side of the nominal levels), one realized now that ‘robustness of performance’ (stability of power, or of the length of confidence intervals) was at least as important . . .

From the beginning, ‘robustness’ has been a rather vague concept; . . . if one wants to choose in a rational fashion between different robust competitors to a classical procedure, one has to make precise the goals one wants to achieve. (Huber, 1972)

From our point of view, this implies that the procedures need to perform satisfactorily under alternative models of the kinds which generate outliers. Our attention is accordingly focused on alternative models of the categories discussed in Chapter 2, and specifically the *mixture*, *slippage*, and *exchangeable* models rather than the *inherent* type of alternative. Inherent alternatives, such as a Cauchy distribution for data normally distributed on the basic model, are covered by more general (non-outlier-specific) robustness procedures. (What is meant by ‘satisfactory’ performance is discussed in some detail in Section 4.1.)

In this chapter, we restrict attention to a univariate sample of  $n$  observations  $x_1, \dots, x_n$ , all of which (on the basic model) belong to a distribution  $F$ .

Many of the detailed results published so far relate to normal or exponential  $F$ . In our general discussion (Section 4.1) of how to construct robust procedures and evaluate their performance, we shall frequently find it convenient to illustrate ideas on the assumption that  $F$  is normal; our basic results will apply, *mutatis mutandis*, to other distributions. In Section 4.2 we review robust procedures where there is no specific declaration of the distributional form of  $F$ . In Section 4.3 we will consider in some detail the particular case where  $F$  is normal.

Most of the existing work based on the *exchangeable* type of alternative model relates to exponential samples. This work is often further particularized by the assumption that there is just one discordant observation. Available results for exponential samples will be discussed in detail in Section 4.4.

In the context of *slippage* alternatives, robust procedures have been discussed for normal samples by a number of writers, including Anscombe (1960a), Tiao and Guttman (1967), Veale and Huntsberger (1969), Guttman and Smith (1969, 1971), Guttman (1973a), and also, in the particular case of samples of size 3, Anscombe and Barron (1966) and Willke (1966). In all these papers (not only the ones dealing with samples of size 3!) the number,  $k$ , of discordant observations in a sample is assumed to be either one or two.

Undoubtedly the type of alternative model most commonly envisaged, at any rate for normal samples, is a mixture model. Tukey (1960), in a seminal paper 'A survey of sampling from contaminated distributions', discusses robust estimation for samples where the basic model is normal,

$$H: F,$$

and the alternative is a mixture of two normals,

$$\bar{H}: (1-\lambda)F + \lambda G \quad (0 < \lambda < 1),$$

the normal distribution  $G$  having either the same mean as  $F$  but a larger variance, or the same variance but a shifted mean. Tukey calls the mixture  $(1-\lambda)F + \lambda G$  a *contaminated distribution*, the basic distribution  $F$  being contaminated by the distribution  $G$ ; the parameter  $\lambda$ , commonly a quite small fraction, is the amount of contamination, or contamination fraction, or just the *contamination*. Sample observations which come from  $G$  are *contaminants*; under  $\bar{H}$ , the number of contaminants in a sample of  $n$  observations will be a binomial random variable with parameters  $n, \lambda$ . We have already used the term 'contaminant' in Sections 3.2 and 3.3 in the context of a slippage alternative. Slippage models can of course be regarded as contamination models in which the number of contaminants in a sample is fixed. Accordingly, some writers refer to the mixture model as one of 'random contamination' and the slippage model as one with a fixed number of contaminants. However, we will use Tukey's terminology, which has become

fairly widely adopted; ‘contamination’ will be understood to be in the mixture-model sense, unless a slippage model is explicitly specified.

In a contamination model, the contaminating distribution  $G$  need not necessarily be normal even if  $F$  is normal. It may be symmetric about the mean of the basic normal distribution  $F$ , or it may not. We distinguish the cases of *symmetric contamination* and *asymmetric contamination* in the following manner for the case of a symmetric basic distribution  $F$ . Contamination is symmetric if the contaminating distribution is symmetric about the centre of the distribution  $F$ . Thus if  $F$  is normal,  $G$  could be normal with the same mean as  $F$  but with a greater variance; or, more generally, it could be of arbitrary symmetric form centred at the mean of  $F$ . With asymmetric contamination,  $G$  may be symmetric about some value different from the centre of  $F$ , or it may be asymmetric. For example, again when  $F$  is normal,  $G$  may be normal (or non-normal symmetric) with a different mean from that of  $F$ , or it may be of arbitrary asymmetric form. We shall be mainly concerned with symmetric contamination of a basic symmetric distribution  $F$  mixed with only *one* contaminating distribution  $G$ . (One could envisage  $F$  being mixed with a set of contaminating distributions, e.g.

$$\bar{H}: (1 - \lambda_1 - \lambda_2)F + \lambda_1 G_1 + \lambda_2 G_2$$

or, still more generally,

$$\bar{H}: \int G(\lambda, x) dK(\lambda)$$

where  $F$  is a particular distribution, say  $F(x) = G(\lambda_0, x)$ , from a one-parameter family  $G(\lambda, x)$ , and  $K$  is a mixing measure. For example,  $F$  might be  $N(\mu, \sigma^2)$  and  $G(\lambda, x)$ , might be  $N(\mu, \lambda\sigma^2)$  with  $\lambda$  exponentially distributed and  $\lambda_0 = 1$ ; the mixture specified by  $\bar{H}$  is in this case a double-exponential distribution, illustrating that with such a mixture situation we have really moved over to an *inherent alternative*.)

If the basic distribution  $F$  is  $N(\mu, \sigma^2)$ , inferences about  $\mu$  may assume  $\sigma^2$  known or unknown; likewise, inferences about  $\sigma^2$  may assume  $\mu$  known or unknown. It might be thought that the cases  $\sigma^2$  known or  $\mu$  known would be only of academic interest, but this is not so; examples of practical situations involving knowledge of  $\sigma^2$  or of  $\mu$  do arise (cf. Sections 3.4.3 and 3.4.4).

What is the effect of contamination? It may be considerable, even for very small values of  $\lambda$ . Suppose, for example, that  $F: N(\mu, \sigma^2)$  is contaminated in the ratio  $1 - \lambda : \lambda$  by another normal distribution  $G: N(\mu, b\sigma^2)$  ( $b > 1$ ) having  $b$  times its variance. In a sample of size  $n$  from the mixture, there will be  $R = 0, 1, \dots$  contaminants, the random variable  $R$  having a binomial distribution with parameters  $n, \lambda$ . The sample mean  $\bar{x}$  will be an unbiased estimator of  $\mu$ , with sampling variance

$$\text{var}(\bar{x}) = \frac{1}{n^2} E_R[(n - R)\sigma^2 + Rb\sigma^2] = \frac{\sigma^2}{n} [1 + (b - 1)\lambda]. \quad (4.0.1)$$

The contamination has caused the sampling variance to increase, relative to  $\sigma^2/n$ , by a factor  $1+(b-1)\lambda$ . For  $\lambda=0.05$  and  $b=9$ , i.e. for 5 per cent contamination by a distribution  $G$  with three times the standard deviation of  $F$ —a not untoward situation—this factor is 1.4; for 10 per cent contamination by the same  $G$  it is 1.8, a loss of efficiency of 44 per cent.

Now consider the effect of the contamination on the performance of the sample variance  $s^2$  as an estimator of  $\sigma^2$ . A straightforward calculation gives, conditional on  $R$  contaminants,

$$E(s^2 | R) = \frac{\sigma^2}{n} [(n-R) + Rb] \quad \text{as in (4.0.1),}$$

$$\begin{aligned} E(s^4 | R) &= \frac{3\sigma^4}{n^2} (n-R+Rb^2) \\ &\quad + \frac{(n^2-2n+3)\sigma^4}{n^2} [(n-R)(n-R-1) + 2(n-R)Rb + R(R-1)b^2]. \end{aligned}$$

Hence, using  $E(R)=n\lambda$ ,  $E(R^2)=n^2\lambda^2+n\lambda-n\lambda^2$ , we get

$$\text{var}(s^2) = \frac{2\sigma^4}{n-1} \left[ 1 + \frac{1}{2}\lambda(b-1)(3b+1) - \frac{1}{2}\lambda^2(b-1)^2 - \frac{3\lambda(1-\lambda)(b-1)^2}{2n} \right]. \quad (4.0.2)$$

The contamination has caused the sampling variance of  $s^2$  to increase, relative to the sampling variance  $2\sigma^4/(n-1)$  under the basic model, by the factor in square brackets in (4.0.2). For  $\lambda=0.05$  and  $b=9$ , the value of this factor is  $6.52-(4.56/n)$ . Even with  $\lambda$  as small as 0.01 and  $b=9$ , the factor is  $2.12-(0.95/n)$ ; with a sample size as small as 7, a mere 1 per cent contamination by  $N(\mu, 9\sigma^2)$  causes a loss of efficiency of 50 per cent in using the sample variance to estimate  $\sigma^2$ !

This striking effect must be due to the incidence, even if infrequent, of extreme values from the contaminating distribution. One might think that contaminants which have so pronounced an effect on the efficiency of estimation would show up unmistakably as outliers and could be rejected on the basis of some discordancy test. This is not so. To entertain such a hope with the sample of size 7 discussed above would be vain; that it would be equally so with large samples has been demonstrated cogently by Tukey (1960). He considers by way of example a sample of 1000 observations from a contaminated distribution  $(1-\lambda)F+\lambda G$ , where  $F$  is  $N(\mu, \sigma^2)$ ,  $G$  is  $N(\mu, 9\sigma^2)$ , and  $\lambda=0.01$  (the 1 per cent contamination of our earlier example). Some typical percentiles of the two distributions are as shown in Table 4.1. Thus the two cumulative distributions are indistinguishable for practical purposes for values of the variable between  $\mu-2\frac{1}{2}\sigma$  and  $\mu+2\frac{1}{2}\sigma$ . Of the ten expected observations from  $G$ , only about 40 per cent (corresponding to

Table 4.1

Cumulative probability	Percentile of $F$	Percentile of $(1-\lambda)F + \lambda G$
0.25	$\mu - 0.67\sigma$	$\mu - 0.68\sigma$
0.05	$\mu - 1.64\sigma$	$\mu - 1.67\sigma$
0.01	$\mu - 2.33\sigma$	$\mu - 2.41\sigma$
0.006	$\mu - 2.51\sigma$	$\mu - 2.64\sigma$

standardized deviates  $\pm 2.5/3 = \pm 0.833$ )

will fall outside these bounds, and we may thus expect only two observations from the upper tail of the broader, rarer constituent [i.e.  $G$ ], and another two from the lower tail. Beyond the same limits we will expect about six (in each tail) from the narrower constituent [i.e.  $F$ ]. Unless one or both of the two are very extreme, the indication of non-normality will be very slight. . . . A sample of one thousand is likely to be of little help. (Tukey, 1960)

## 4.1 PERFORMANCE CRITERIA

### 4.1.1 Efficiency measures for estimators

Suppose we are estimating the location parameter  $\mu$  of a symmetric population  $F$  from a sample of size  $n$ . Our choice of estimator might be unrestricted, or we might decide on the other hand to confine ourselves to some restricted class of estimators, such as linear combinations  $a_1x_{(1)} + \dots + a_nx_{(n)}$  of the sample order statistics. If the basic model holds good,

$$H: F,$$

an ‘optimal’ estimator of  $\mu$  can be defined among the estimators of the class we are considering; we will denote this by  $\check{\mu}$ . For example,  $\check{\mu}$  might be the maximum likelihood estimator  $\hat{\mu}$ , or the linear unbiased estimator of form  $\sum a_i x_{(i)}$  with minimum variance (the ‘best linear unbiased estimator’, or BLUE); while these estimators coincide for a normal distribution, they will be quite different for, say, a Cauchy distribution. However we define it,  $\check{\mu}$  serves as a yardstick for what an estimator can achieve in relation to the basic model. To fix ideas, suppose  $F$  is  $N(\mu, \sigma^2)$ ;  $\check{\mu}$  then would typically be the sample mean  $\bar{x}$ . We now propose some rival estimator,  $T$ , which shall be robust in relation to some outlier-generating model  $\bar{H}$ .

Starting with the simplest situation, we take  $\bar{H}$  to be a simple and symmetric alternative, providing merely for the observations to belong to a specified mixture distribution with known symmetric contamination. That is,

$$\bar{H}: (1-\lambda)F + \lambda G \quad (4.1.1)$$

where typically  $G$  could be  $N(\mu, b\sigma^2)$ ,  $b > 1$ , and  $\lambda, b$  are known. In view of the symmetry of the model we can reasonably assume  $E(T | \bar{H}) = \mu$ , and the question of bias in  $T$  does not arise.

If  $\text{var}(\check{\mu} | \bar{H})$  did not appreciably exceed  $\text{var}(\check{\mu} | H)$ ,  $\check{\mu}$  would itself be robust, and there would be no need to seek for a rival estimator  $T$ ; the ratio

$$\frac{\text{var}(\check{\mu} | \bar{H})}{\text{var}(\check{\mu} | H)}$$

provides a quantitative indication of the need for a robust estimator alternative to  $\check{\mu}$ . We can assume that this ratio exceeds unity substantially, as in the example discussed under equation (4.0.2).

For robustness of  $T$  we require that  $\text{var}(T | \bar{H})$  shall be substantially less than  $\text{var}(\check{\mu} | \bar{H})$ ; also, of course,  $T$  must be a reasonable alternative to the optimal  $\check{\mu}$  when the data obey the basic model  $F$ , i.e. we require that  $\text{var}(T | H)$  shall not be 'unduly' greater than  $\text{var}(\check{\mu} | H)$ . So  $\text{var}(T | \bar{H})/\text{var}(\check{\mu} | \bar{H})$  is to be 'small', and  $\text{var}(T | H)/\text{var}(\check{\mu} | H)$  'not much greater than unity'. Each of these ratios is, of course, a relative efficiency measure, and discussions of performance are often phrased in terms of efficiency. An alternative terminology is that of *protection* and *premium*; these concepts, equivalent to the above relative efficiency measures, were introduced by Anscombe (1960a) in a classical paper. We quote the passage in which he introduces these terms as part of his exposition of a basic philosophy in the matter of accommodating outliers:

Rejection rules are not significance tests. . . . when a chemist doing routine analyses, or a surveyor making a triangulation, makes routine use of a rejection rule, he is not studying whether spurious readings occur (he may already be convinced they do sometimes), but guarding himself from their adverse effect. . . .

A rejection rule is like a householder's fire insurance policy. Three questions to be considered in choosing a policy are

- (1) What is the premium?
- (2) How much protection does the policy give in the event of fire?
- (3) How much danger really is there of a fire?

Item (3) corresponds to the study of whether spurious readings occur in fact . . . The householder, satisfies that fires *do* occur, does not bother much about (3), provided the premium seems moderate and the protection good. (Anscombe, 1960a)

The 'fire' here is the occurrence of outliers. The discussion is couched in terms of rejection rules, but applies to accommodation procedures in general. Anscombe goes on to ask:

In what currency can we express the premium charged and the protection afforded by a rejection rule? . . . variance will be considered here, although in principle any other measure of expected loss could be used. The premium payable may then be taken to be the percentage increase in the variance of estimation errors due to using the rejection rule, when in fact all the observations come from a homogeneous normal source; the protection given is the reduction in variance (or mean squared error) when spurious readings are present. (Anscombe, 1960a)

Clearly, in the case we have considered so far,

$$\text{premium} = \frac{\text{var}(T | H) - \text{var}(\check{\mu} | H)}{\text{var}(\check{\mu} | H)} \quad (4.1.2)$$

and

$$\text{protection} = \frac{\text{var}(\check{\mu} | \bar{H}) - \text{var}(T | \bar{H})}{\text{var}(\check{\mu} | \bar{H})}. \quad (4.1.3)$$

Whether the measurement of robustness is discussed in terms of variance ratios or relative efficiencies on the one hand, or premium and protection on the other, is a matter of taste. Papers which make explicit use of premium and protection include Anscombe and Barron (1966), Tiao and Guttman (1967), Guttman and Smith (1969, 1971), Guttman (1973a), and Desu, Gehan, and Severo (1974).

From the simple situation just discussed, we may generalize in several directions.

### (i) Asymmetric contamination

If  $F$  is  $N(\mu, \sigma^2)$  and  $G$  is, say,  $N(\mu + a, \sigma^2)$ ,  $E(T | \bar{H})$  will not in general be equal to  $\mu$ , and we have to take account of bias in our estimators. In this situation,

a natural criterion for judging them is their mean squared error, a measure which takes into account both their inherent variability and their distance from the estimand. (Jaeckel, 1971a)

This leads us to replacing  $\text{var}(\check{\mu} | \bar{H})$ ,  $\text{var}(T | \bar{H})$  in our discussion by the mean squared error values

$$MSE(\check{\mu} | \bar{H}) = E[(\check{\mu} - \mu)^2 | \bar{H}]; \quad MSE(T | \bar{H}) = E[(T - \mu)^2 | \bar{H}].$$

There is also the question of how the bias affects performance. We have

$$E[(T - \mu)^2 | \bar{H}] = \text{var}(T | \bar{H})(1 + c^2)$$

where  $c$  is the ratio of the bias to the standard deviation of the estimator. Consider, for example, the performance of the sample mean  $\bar{x}$  in estimating the mean  $\mu$  of  $F: N(\mu, \sigma^2)$  when there is contamination of amount  $\lambda$  by  $G: N(\mu + a, \sigma^2)$ . With  $R$  contaminants among the  $n$  observations, we have

$$E(\bar{x} | R) = \mu + (a/n)R, \quad E(\bar{x}^2 | R) = [\mu + (a/n)R]^2 + (\sigma^2/n),$$

so that

$$E(\bar{x}) = \mu + \lambda a \quad (4.1.4)$$

and

$$\text{var}(\bar{x}) = \frac{\sigma^2}{n} + \frac{\lambda(1-\lambda)a^2}{n}. \quad (4.1.5)$$

Thus the estimator has bias  $\lambda a$ , and mean squared error

$$MSE(\bar{x}) = \frac{\sigma^2}{n} + a^2 \left[ \lambda^2 + \frac{\lambda(1-\lambda)}{n} \right]. \quad (4.1.6)$$

It follows from (4.1.4), (4.1.5) that if we want the bias of  $\bar{x}$  to be less than, say, one-half the standard deviation, the sample size must satisfy

$$n < \frac{1}{4} \left( \frac{\sigma^2}{\lambda^2 a^2} + \frac{1-\lambda}{\lambda} \right). \quad (4.1.7)$$

For example, with 1 per cent contamination, and  $a$  equal to  $5\sigma$ , the sample size must not exceed 125. For further discussion see Huber (1964, p. 83) and Jaeckel (1971a).

### (ii) Compound alternative $\bar{H}$

If in (4.1.1) we regard  $\lambda$  as a parameter with a range of possible values (say  $0 < \lambda \leq \lambda_1$ ) rather than a single known quantity, our alternative model is a family of mixture distributions, indexed by  $\lambda$ . Consider any performance measure  $M$ , for example the protection measure defined in (4.1.3)

$$M = [\text{var}(\check{\mu} | \bar{H}) - \text{var}(T | \bar{H})] / \text{var}(\check{\mu} | \bar{H}).$$

$M$  will now be a function of  $\lambda$ , and we can write it as  $M(\lambda)$ . (Similarly,  $M$  could be a function of  $b$ , or of the two parameters  $\lambda$  and  $b$ .) If our estimator  $T$  is to possess robustness in relation to  $\bar{H}$ , it must perform ‘satisfactorily’ for every distribution which can arise under  $\bar{H}$ , and the value of  $M(\lambda)$  must be ‘satisfactory’ for all values of  $\lambda$  between zero and  $\lambda_1$ . For a *single* measure of performance this naturally suggests the value of the worst possible performance under  $\bar{H}$ , in other words the extreme value of  $M(\lambda)$  (minimum or maximum as appropriate) over the range of possible values of  $\lambda$ . This measure could, for instance, take the form of the maximum variance of the estimator under  $\bar{H}$ , or on the other hand the minimum protection value, as  $\lambda$  varies between zero and  $\lambda_1$ . Correspondingly in the two-parameter case we would use the extreme value of  $M(\lambda, b)$ .

### (iii) Estimation of scale and other parameters

Robust estimation may of course be required, not only for location parameters, but for scale parameters (as with exponential distributions), dispersion or scale parameters (as with normal distributions), or shape parameters (as with Pareto distributions). Here we use the term ‘scale parameter’ for a dispersion parameter expressed in the same dimensions as the random variable, e.g. standard deviation (as opposed to variance). Our discussion of performance criteria has focused on the robust estimation of a location parameter, but obviously applies in essentials to any other kind of parameter. A scale parameter for a random variable  $X$  can in any case be regarded

as a location parameter for an appropriately transformed random variable  $Y$ : for example  $Y = |X - \theta|$ , where  $\theta$  is some location parameter for  $X$ .

#### (iv) Asymptotic measures

The variances, relative efficiencies and mean squared errors we have used so far in our discussion have been actual (i.e. finite-sample) values. For example, the variance ratio  $\text{var}(s^2 | \bar{H})/\text{var}(s^2 | H)$ , indicating the non-robustness of the normal sample variance  $s^2$  as an estimator of  $\sigma^2$ , was shown in (4.0.2) to have the value

$$M(\lambda, b) = 1 + \frac{1}{2}\lambda(b-1)(3b+1) - \frac{1}{2}\lambda^2(b-1)^2 - \frac{3\lambda(1-\lambda)(b-1)^2}{2n} \quad (4.1.8)$$

for the contamination model  $(1-\lambda)\mathbf{N}(\mu, \sigma^2) + \lambda\mathbf{N}(\mu, b\sigma^2)$ . For the typical numerical values  $\lambda = 0.05$ ,  $b = 9$ , this gave

$$M(0.05, 9) = 6.52 - \frac{4.56}{n}. \quad (4.1.9)$$

Unless  $n$  is small, this differs little in value from its limit as  $n \rightarrow \infty$ , i.e. 6.52. We can, if we wish, use this limiting value—the *asymptotic* variance ratio—as our measure, rather than the finite-sample variance-ratio (4.1.9). This choice between finite-sample and asymptotic values is obviously available for any measure involving variances, or efficiencies based on variance, or their premium and protection equivalents. With certain adjustments (see below), it is also available for measures based on mean squared errors.

As regards nomenclature,  $\text{var } T$  will in general be of the form  $(A/n) \times (1 + O(n^{-1}))$ , so that  $n \text{ var } T$  tends to a finite limit  $A$  as  $n \rightarrow \infty$ ; it is this limit which is conventionally called the asymptotic variance (and similarly for asymptotic mean squared error and asymptotic bias, see below).

Huber (1964) argues in favour of using asymptotic measures:

Since ill effects from contamination are mainly felt for large sample sizes, it seems that one should primarily optimize large sample robustness properties. . . . the asymptotic variance is not only easier to handle, but . . . even for moderate values of  $n$  it is a better measure of performance than the actual variance, because (i) the actual variance of an estimator depends very much on the behaviour of the tails of  $H$  [ $G$  in our notation] . . . . (ii) If an estimator is asymptotically normal, then the important central part of its distribution and confidence intervals for moderate confidence levels can better be approximated in terms of the asymptotic variance than in terms of the actual variance. (Huber, 1964)

On these grounds he adopts the maximum asymptotic variance (over the family of alternative distributions in a compound  $\bar{H}$ ) as a measure of performance. This measure has been widely used in the construction of robust estimators with optimal properties. The *minimax* robust estimator (Huber, 1964) is that estimator (perhaps restricted to be of a particular type)

whose maximum variance over the family of distributions in  $\bar{H}$  is as small as possible.

A related prospect (e.g. Gastwirth, 1966) is to seek the *maximin* robust estimator. Here we consider the efficiencies of an estimator relative to other estimators each of which is known to perform well for individual distributions in the compound  $\bar{H}$ . The maximin estimator is one whose minimum efficiency relative to the individually satisfactory estimators is as large as possible.

See, e.g. Huber (1964), Bickel (1965), Gastwirth (1966), Hogg (1967), Siddiqui and Raghunandanan (1967), and Jaeckel (1971a, 1971b), on use of the minimax and maximin criteria. We shall review some of the relevant results in Section 4.2.

At the same time, finite-sample variance and efficiency measures are clearly the appropriate ones to use in some situations, such as Monte Carlo studies of robustness (cf. Andrews *et al.*, 1972), or the assessment of accommodation procedures for small samples, *per se*. (Extreme instances of the latter are studies relating to samples of size 3, such as those by Anscombe and Barron, 1966, and Willke, 1966.) The choice resembles that between the finite-sample variance of an estimator  $\hat{\theta}$  and the information-function reciprocal in the straightforward estimation of a parameter  $\theta$ . For robustness studies using finite-sample measures with sample sizes greater than 3 see, e.g. Dixon (1960), Birnbaum and Laska (1967), Birnbaum, Laska, and Meisner (1971), and Guttman (1973a). The finite-sample properties of various measures, in relation to their asymptotic values, have been investigated in detail by Gastwirth and Cohen (1970); see also Crow and Siddiqui (1967).

In the case of asymmetric contamination, variances under  $\bar{H}$  are, as we have seen, replaced by mean squared errors. Now, as our example (4.1.6) illustrates, while the variance of an estimator tends to zero as  $n^{-1}$  when  $n \rightarrow \infty$ , its bias may be independent of  $n$ , or at any rate may tend to a non-zero limit. On this basis, comparisons between asymptotic mean squared errors of estimators would be meaningless. To meet this situation we might modify the contamination model (4.1.1) as follows. Instead of taking the amount of contamination to be a basic parameter,  $\lambda$ , we assume it to depend on sample size according to the relation

$$\lambda = \lambda(n) = \lambda_1 n^{-\frac{1}{2}}. \quad (4.1.10)$$

The alternative model is now

$$\bar{H}: (1 - \lambda_1 n^{-\frac{1}{2}})F + \lambda_1 n^{-\frac{1}{2}}G, \quad (4.1.11)$$

descriptive of a situation in which 'the amount of asymmetric contamination is large enough to affect the performance of the estimator, but is too small to be measured accurately at the given sample size' (Jaeckel, 1971a). In (4.1.4),

(4.1.5), (4.1.6), the estimator  $\bar{x}$  would now have bias  $\lambda_1 a n^{-\frac{1}{2}}$ , variance

$$\frac{\sigma^2}{n} [1 + O(n^{-\frac{1}{2}})],$$

and mean squared error

$$\frac{\sigma^2 + \lambda_1^2 a^2}{n} [1 + O(n^{-\frac{1}{2}})]$$

leading naturally to the definition of an *asymptotic bias*  $\lambda_1 a$  and an *asymptotic mean squared error*  $\sigma^2 + \lambda_1^2 a^2$ .

#### 4.1.2 The qualitative approach: influence curves

So far in our discussion of performance criteria, we have confined ourselves to the question ‘How does contamination affect the *precision* (and maybe the bias) of an estimator?’ This prompted the various dispersion-based and efficiency-based criteria. But precision is only one aspect. In what way, one would like to know, does contamination *influence* a given estimator? For example, is the effect on the estimator proportional to the number of contaminants present? Supposing there is just one contaminant, how is its effect related to its magnitude? What is the worst possible effect that a single contaminant can have? In particular, is this effect bounded or not? As we show below, a contaminant in a sample of  $n$  will, if large enough, shift the sample *mean*  $\bar{x}$  beyond any bound; but two contaminants in a sample of odd size  $n = 2m - 1$  can at most shift the sample *median* from  $x_{(m)}$  to  $x_{(m-1)}$  or  $x_{(m+1)}$ , however far out these two contaminants may be. Aspects such as these underlie a powerful array of tools, which we will now describe, based on the *influence function* or *influence curve*. The approach is due to Hampel (1968, 1971); for a stimulating and highly readable exposition, see Hampel (1974).

As usual, suppose we have a basic model  $F$  and a contamination model  $(1 - \lambda)F + \lambda G$ . If the contamination fraction  $\lambda$  is small enough, the number of contaminants  $R$  in a sample will effectively be either 0 or 1, so that for marginal comparisons of the performance of estimators in the neighbourhood of  $\lambda = 0$  we need only consider the case of a single contaminant. Given  $n$  ‘good’ (basic-model) observations  $x_1, \dots, x_n$  and an estimator  $T(x_1, \dots, x_n)$ , we wish in principle to examine the effect on  $T$  of substituting a contaminant for one of the  $n$  observations. Denote the contaminant, as in Section 3.2, by  $x_c$ . The effect as defined would require averaging with respect to two random elements, first the random variation in  $x_c$  as sampled from  $G$ , and second the random variation in the good value,  $x_i$  say, which has been replaced by  $x_c$ . To sidestep these sources of variation we formulate the problem, equivalently and more conveniently, as follows. The contaminating distribution  $G$  we will take to be atomic at  $\xi$ ; that is to say, the

contaminant  $x_c$  has a fixed value  $\xi$ . And then we will work in terms of the effect on  $T$  of adding the contaminant  $\xi$  to the  $n$  good observations, so that on the alternative model  $T$  is based on an enlarged sample of size  $n+1$ .

Suppose, for example, that  $T(x_1, \dots, x_n) = T$  is the sample mean  $\bar{x}$ . Write  $T(x_1, \dots, x_n, \xi) = T_c$  for the mean  $\bar{x}_c$  based on the enlarged contaminated sample. Then the effect of adjoining  $\xi$  is to change the value of the estimator by an amount

$$\bar{x}_c - \bar{x} = \frac{n\bar{x} + \xi}{n+1} - \bar{x} = \frac{\xi - \bar{x}}{n+1}. \quad (4.1.12)$$

Naturally enough this is proportional to  $1/(n+1)$ , that is to the amount of contamination in the sample; the effect standardized for the amount of contamination is

$$(n+1)(\bar{x}_c - \bar{x}) = \xi - \bar{x}. \quad (4.1.13)$$

This will, as we remarked above, exceed any bound for  $\xi$  large enough. It is a linear function of the value of the contaminant.

Again, if  $T$  is the sample variance  $s^2$  for a distribution with unknown mean and variance, we have for the enlarged sample

$$ns_c^2 = (n-1)s^2 + n\bar{x}^2 + \xi^2 - (n\bar{x} + \xi)^2/(n+1),$$

giving

$$s_c^2 - s^2 = [(\xi - \bar{x})^2/(n+1)] - (s^2/n).$$

The standardized effect is therefore

$$(n+1)(s_c^2 - s^2) = (\xi - \bar{x})^2 - \frac{n+1}{n}s^2. \quad (4.1.14)$$

The effect again exceeds any bound for  $\xi$  large enough, but this time is a quadratic function of  $\xi$ .

Effects per unit of contamination, such as (4.1.13) and (4.1.14), are called finite-sample influence functions or, following Hampel, *finite-sample influence curves*. A finite-sample influence curve depends on the argument  $\xi$ , on the estimator  $T$ , and in general (see, for example, the case of the sample median discussed below) on the basic distribution  $F$ ; it may also depend explicitly, as (4.1.14) illustrates, on the sample size  $n$ . Accordingly we write it  $IC_{T,F;n}(\xi)$ .

Equation (4.1.14) also suggests that, as with the variance and efficiency measures discussed earlier, we may wish to use the asymptotic equivalent

$$\lim_{n \rightarrow \infty} IC_{T,F;n}(\xi) = IC_{T,F}(\xi),$$

say. In fact, it is this asymptotic influence curve, or simply the *influence curve*,  $IC_{T,F}(\xi)$ , which is the really useful tool.

What is the influence curve for  $s^2$ ,  $IC_{s^2,F}(\xi)$ ? If we let  $n \rightarrow \infty$  in (4.1.14), we must not only replace  $(n+1)/n$  by 1, but also  $\bar{x}$  and  $s^2$  by  $\mu$  and  $\sigma^2$  respectively:

$$IC_{s^2,F}(\xi) = (\xi - \mu)^2 - \sigma^2. \quad (4.1.15)$$

This equation conveys the same information as the finite-sample version (4.1.14) regarding the unbounded quadratic influence of a contaminant on  $s^2$ . In deriving it, each estimator  $T(x_1, \dots, x_n)$  (e.g.  $\bar{x}$  or  $s^2$ ) on the right-hand side of (4.1.14) has been replaced by  $\lim_{n \rightarrow \infty} T(x_1, \dots, x_n)$  (e.g.  $\mu$  or  $\sigma^2$ ); this limiting form depends only on  $F$  and we will write it  $T(F)$ . For example, if  $T$  is the sample mean  $\bar{x}$ ,  $T(F) = \int x dF$ ; if  $T$  is the sample variance  $s^2$ ,  $T(F) = \int (x - \mu)^2 dF$  where  $\mu = \int x dF$ .

Encompassing our procedure in a formal definition, we say that the influence curve of an estimator  $T(x_1, \dots, x_n)$  at the basic distribution  $F$  is

$$IC_{T,F}(\xi) = \lim_{\lambda \rightarrow 0} \{[T((1-\lambda)F + \lambda G) - T(F)]/\lambda\} \quad (4.1.16)$$

where  $G$  is the atomic distribution

$$P(X = \xi) = 1. \quad (4.1.17)$$

We may also write (4.1.16) as

$$IC_{T,F}(\xi) = \frac{\partial}{\partial \lambda} \{T[(1-\lambda)F + \lambda G]\}|_{\lambda=0} \quad (4.1.18)$$

*Example 4.1 Sample mean.* We have

$$IC_{\bar{x},F}(\xi) = \lim_{\lambda \rightarrow 0} \{[(1-\lambda)\mu + \lambda\xi - \mu]/\lambda\} = \xi - \mu, \quad (4.1.19)$$

which could also have been obtained by letting  $n \rightarrow \infty$  in (4.1.13).

*Example 4.2 Sample variance*

$$\begin{aligned} IC_{s^2,F}(\xi) &= \lim_{\lambda \rightarrow 0} \{[(1-\lambda)(\mu^2 + \sigma^2) + \lambda\xi^2 - ((1-\lambda)\mu + \lambda\xi)^2 - \sigma^2]/\lambda\} \\ &= (\xi - \mu)^2 - \sigma^2 \end{aligned}$$

as in (4.1.15).

*Example 4.3 Sample median.* It is not practicable to calculate the influence curve of the sample median  $\bar{x}$  on a finite-sample basis, since the shift in sample median on moving from an odd to an even number of observations, or vice versa, is not defined. On an asymptotic basis, however, the calculation is straightforward.  $T(F)$  is now the population median  $m$ , defined by

$$\int_{-\infty}^m dF = \frac{1}{2}.$$

$T((1-\lambda)F + \lambda G)$ , the median of the mixture distribution, is equal to  $m + \Delta$ , say, where  $\Delta$  is positive or negative according as  $\xi$  is greater or less than  $m$ . We assume  $F$  to be continuous, with density  $f$ . To the first order of small quantities we have for  $\Delta < 0$ ,

$$\frac{1}{2} = (1-\lambda)F(m + \Delta) + \lambda = (1-\lambda)[\frac{1}{2} + \Delta f(m)] + \lambda = \frac{1}{2} + \frac{1}{2}\lambda + \Delta f(m),$$

giving  $\Delta = -\lambda/[2f(m)]$ ; similarly, for  $\Delta > 0$ ,  $\Delta = +\lambda/[2f(m)]$ . Hence the influence curve,  $\lim_{\lambda \rightarrow 0} (\Delta/\lambda)$ , is

$$IC_{\bar{x}, F}(\xi) = \frac{\text{sgn}(\xi - m)}{2f(m)}. \quad (4.1.20)$$

The influence of a contaminant on the sample median is thus seen to be bounded—an essential qualitative difference from, say, the sample mean!

A readily calculated finite-sample representation of the influence curve is the *sensitivity curve* introduced by Tukey (1970). For this, the contaminant  $\xi$  is added, not to a random sample  $x_1, \dots, x_n$  from  $F$  as for the finite-sample influence curve, but to a pseudo-sample  $c_1, \dots, c_n$  consisting of the order scores  $c_j = E(X_{(j)})$  for a sample of size  $n$  from  $F$ . This constructed sample may be thought of as smoothly representing the basic distribution. The sensitivity curve is given (as with the finite-sample  $IC$ ) by  $(n+1)[T(c_1, \dots, c_n, \xi) - T(c_1, \dots, c_n)]$  regarded as a function of  $\xi$ . Other convenient order statistics could of course be used in place of order scores; for example, conditional centroids or medians. All of these have been extensively tabulated in the normal case (David, Barton, Ganeshalingam, Harter, Kim, Merrington, and Walley, 1968).

Reverting to the influence curve, we note an important property. If we regard the argument  $\xi$  as a random quantity distributed according to the basic model  $F$ , it can be shown (see, for example, Huber 1972, pp. 1051–1052) that the expectation of the influence curve with respect to this variation in  $\xi$  is zero:

$$\int IC_{T,F}(\xi) dF(\xi) = 0, \quad (4.1.21)$$

and that the mean squared value of the influence curve,

$$\int [IC_{T,F}(\xi)]^2 dF(\xi),$$

is equal to the *asymptotic variance* of  $T$ . Thus we have a direct connection between the influence curve and our earlier dispersion-based performance criteria.

We now describe some further parameters of the influence curve which throw light on the robustness of an estimator.

(i) *The gross-error sensitivity*

This is the supremum of the absolute value of the influence curve,

$$\gamma_{T,F} = \sup_{\xi} |IC_{T,F}(\xi)| \quad (4.1.22)$$

The gross-error sensitivity ‘measures the worst approximate influence which a fixed amount of contamination can have on the value of the estimator (hence it may be regarded as an approximate bound for the bias of the estimator)’ (Hampel, 1974).

*Example 4.4 Suppose  $F$  is  $N(\mu, \sigma^2)$ ; for  $\bar{x}$  and for  $s^2$ ,  $\gamma_{T,F} = \infty$  (so that the effect that a contaminant can have on these estimators is unbounded), while for  $\tilde{x}$  (the median)*

$\gamma_{T,F} = 1/[2f(\mu)]$ , from (4.1.20); that is  $\gamma_{T,F} = \sigma\sqrt{(2\pi)/2} = 1.25\sigma$ .

(ii) *The local-shift sensitivity*

This is defined by

$$\beta_{T,F} = \sup_{\xi \neq \eta} |IC_{T,F}(\xi) - IC_{T,F}(\eta)|/|\xi - \eta|. \quad (4.1.23)$$

It measures the worst possible effect of ‘adjusting’ a contaminant by modifying its value, for example by Winsorizing.

(iii) *The rejection point*

Suppose that the influence curve vanishes for all points  $\xi$  outside some finite interval

$$|\xi - \mu| \leq \rho,$$

say, centred on  $\mu$ , the mean (or other appropriate location point) of  $F$ . This implies that observations outside  $[\mu - \rho, \mu + \rho]$  have no influence on the estimator  $T$ —i.e. that the estimation procedure *rejects* such observations.  $\rho = \rho_{T,F}$  is called the *rejection point* of the estimator. Examples of estimators with finite  $\rho_{T,F}$  (and which therefore reject outliers beyond some particular distance) will be encountered below in the context of  $M$ -estimation.

In some cases  $|IC_{T,F}|$  may be very small, though not zero, for  $|\xi - \mu|$  sufficiently large. The rejection point is then infinite, but outliers, though not explicitly rejected, have very little effect on the estimator.

(iv) *The breakdown aspect of performance*

We noted above that, in contrast to the sample mean, the influence of a contaminant on the sample median  $\tilde{x}$  is bounded, and indeed that two contaminants, whatever their magnitudes, added to a sample of size  $n = 2m - 1$  can at most shift the sample median from  $x_{(m)}$  to  $x_{(m-1)}$  or  $x_{(m+1)}$ .

Why stop at two? Clearly the data can absorb a greater amount of contamination than this without the sample median becoming totally unreliable. With four contaminants added, the shift from  $x_{(m)}$  is bounded (at most to  $x_{(m-2)}$  or  $x_{(m+2)}$ ); with  $2m-2$  contaminants added, it is still bounded (at most to  $x_{(1)}$  or  $x_{(2m-1)}$ ). But as soon as the number of added contaminants exceeds  $2m-2$ , it is possible for  $\bar{x}$  to take any value whatsoever. From this point of view, the contamination becomes intolerable when its proportionate amount reaches  $(2m-1)/(4m-2)$ , i.e. one-half. We say that the sample median has *breakdown point*  $\frac{1}{2}$ . The breakdown point,  $\pi_{T,F}$ , for any estimator  $T$  is the smallest proportion of contamination which can carry the value of the estimator over all bounds. It is an important measure of robustness; the idea is due to Hodges (1967) and Hampel (1971).

We have now discussed in some detail what Huber calls the *stability* aspect of robustness,

in close analogy to the stability of a mechanical structure (say of a bridge): (i) the qualitative aspect: a small perturbation should have small effects; (ii) the breakdown aspect: how big can the perturbation be before everything breaks down; (iii) the infinitesimal aspect: the effects of infinitesimal perturbations. (Huber, 1972)

#### 4.1.3 Robustness of confidence intervals

Suppose we have a sample  $x_1, \dots, x_n$  which comes, on some basic model  $H$ , from a distribution involving a location parameter  $\mu$  and a scale parameter  $\sigma$ , both unknown; typically this distribution might be  $N(\mu, \sigma^2)$ . Consider the problem of constructing a confidence interval for  $\mu$  at level  $1-\alpha$ . Essentially this is built up from the following elements:

- (i) an estimator  $T$  of  $\mu$ ;
- (ii) an estimator  $S_T$  of the standard deviation of  $T$ ;
- (iii) the distribution  $\mathbf{D}$  of  $(T-\mu)/S_T$ , assuming the basic model.

If  $u_1, u_2$  are lower and upper  $\frac{1}{2}\alpha$ -points of  $\mathbf{D}$ , the confidence interval is then

$$(T - u_2 S_T, T - u_1 S_T). \quad (4.1.24)$$

If the parent distribution is  $N(\mu, \sigma^2)$ ,  $T$ ,  $S_T$  and  $\mathbf{D}$  particularize to  $\bar{x}$ ,  $s/\sqrt{n}$  and the  $t_{n-1}$  distribution in the classical procedure, and we get the familiar confidence interval

$$\left( \bar{x} - t_{n-1}(\alpha/2) \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1}(\alpha/2) \frac{s}{\sqrt{n}} \right) \quad (4.1.25)$$

where  $t_{n-1}(\alpha/2)$  is the upper  $(\alpha/2)$ -point of  $t_{n-1}$ .

If in fact the observations  $x_1, \dots, x_n$  come, not from  $N(\mu, \sigma^2)$  as assumed, but from a contaminated distribution, the confidence interval (4.1.25) may be defective for two reasons. First, the distribution  $\mathbf{D}$  of  $(T-\mu)/S_T = (\bar{x}-\mu)/(s/\sqrt{n})$  may differ substantially from that of  $t_{n-1}$ , and the probability

that the interval (4.1.25) covers the true value  $\mu$  may thus differ substantially from  $1 - \alpha$  (and may—a particularly undesirable occurrence—be substantially less than  $1 - \alpha$ ). That is, the procedure may lack *robustness of validity*. Secondly, since  $\bar{x}$  and  $s$  are sensitive to the presence of extreme values in the sample, the confidence interval may be unnecessarily wide. It could well be preferable to use more robust estimators  $T$  and  $S_T$  in the construction (4.1.24), aiming at achieving satisfactory validity, and at the same time obtaining confidence intervals which tend to be shorter in the presence of extreme values. From this point of view the procedure leading to (4.1.25) may lack *robustness of efficiency*, or ‘robustness of performance’ in Huber’s words quoted at the beginning of this chapter. See Tukey and McLaughlin (1963), Dixon and Tukey (1968), Huber (1968, 1970).

The choice of  $T$  and  $S_T$  will be discussed in Section 4.2.4.

The following measures of performance of a confidence interval such as (4.1.24) have been proposed:

- (i) The probability that the confidence interval fails to cover the true parameter value  $\mu$  under a specified alternative model  $\bar{H}$ . This reflects the robustness of validity of the procedure. Takeuchi (1971) estimates it by the relative frequency of non-coverage of the parameter by the interval in a large number of simulations; he calls this the *error frequency* of the interval. Analogously, the measure itself may be called the *error probability*.
- (ii) Suppose we have a compound alternative  $\bar{H}$ ; a confidence interval  $(T - a, T + a)$  of given length  $2a$  will operate at different confidence levels for the different distributions that can arise under  $\bar{H}$ . For specified  $a$ , the minimum of these possible confidence levels gives a measure of ‘guaranteed’ performance. The idea is due to Huber (1968).
- (iii) If the interval (4.1.24) has robustness of validity, a natural measure of its efficiency (for specified  $\alpha$ ) is the ratio of the lengths of the intervals (4.1.25), (4.1.24); or, from another point of view, the ratio measures the relative efficiency of two procedures. Dixon and Tukey (1968, p. 86) call the *square* of this ratio the relative efficiency.

#### 4.1.4 Robustness of significance tests

This obviously bears a relation to the robustness of confidence intervals discussed above. There is an important difference, however, inasmuch as we now have a double family of alternative hypotheses. Suppose, to fix ideas, that a two-sided test of the hypothesis  $\mu = \mu_0$  on the basis of an assumed normal sample is required. We can still think in terms of a basic model

$$H: F \quad \text{where } F \text{ is } N(\mu, \sigma^2), \quad (4.1.26)$$

and a contamination alternative  $\bar{H}$  which might typically be

$$\bar{H}: (1 - \lambda)F + \lambda G \quad \text{where } G \text{ is } N(\mu, b\sigma^2). \quad (4.1.27)$$

$\bar{H}$  depends on the parameters  $\lambda$ ,  $b$  etc., which for present purposes we will denote simply by  $\lambda$ .

The null hypothesis for the significance test is

$$\bar{H}_0: F_0 \quad \text{where } F_0 \text{ is } N(\mu_0, \sigma^2). \quad (4.1.28)$$

We are concerned with the behaviour of the test, both in relation to the usual 'Type II error' family of alternatives (4.1.26) and in relation to the family of alternatives

$$\bar{H}_0: (1 - \lambda)F_0 + \lambda G_0 \quad \text{where } G_0 \text{ is } N(\mu_0, b\sigma^2) \quad (4.1.29)$$

expressing contamination under the true value of  $\mu$ ; more generally, we are concerned with the family (4.1.27), encompassing both types of departure from  $H_0$ .

For any choice of  $T$  and  $S_T$  we have, corresponding to the  $(1 - \alpha)$ -level confidence interval (4.1.24), a significance test at level  $\alpha$  of  $H_0$  against  $\bar{H}_0$  with critical region

$$\mathcal{G}: (T - \mu_0)/S_T < u_1 \quad \text{or} \quad > u_2. \quad (4.1.30)$$

We can now formulate relevant measures of performance, as follows:

(i) The conventional power of the test, as a function of  $\mu$ ; this is

$$\Pi_1(\mu) = P[(T - \mu_0)/S_T \in \mathcal{G} | H]. \quad (4.1.31)$$

(ii) The stability of the significance level under contamination, as a function of  $\lambda$ ; this is given by

$$\Pi_2(\lambda) = \Pi_2(\lambda, \alpha) = P[(T - \mu_0)/S_T \in \mathcal{G} | \bar{H}_0]. \quad (4.1.32)$$

It corresponds to the error probability of the equivalent confidence interval, as defined above.

(iii) For a compound alternative  $\bar{H}$ , the guaranteed significance level

$$\Pi_3 = \Pi_3(\alpha) = \min_{\lambda} \Pi_2(\lambda, \alpha). \quad (4.1.33)$$

This again corresponds to the guaranteed performance measure under contamination defined above for a confidence interval.

(iv) The stability of the power under contamination, as a function of both  $\mu$ , the argument of the power function, and  $\lambda$ , the measure of contamination:

$$\Pi_4(\mu, \lambda) = P[(T - \mu_0)/S_T \in \mathcal{G} | \bar{H}]. \quad (4.1.34)$$

(v) For a compound alternative  $\bar{H}$ , the guaranteed power at  $\mu$ :

$$\Pi_5(\mu) = \min_{\lambda} \Pi_4(\mu, \lambda). \quad (4.1.35)$$

These concepts apply to significance tests generally, though our discussion has been in the context of tests for the mean of a normal distribution. Veale and Kale (1972), for example, consider the testing of hypotheses for the parameter  $\sigma$  of an exponential distribution with density  $\sigma^{-1} \exp(-\sigma^{-1}x)$ , and they develop a test (described in Section 4.4) robust against a contaminant arising from an exchangeable alternative model with contamination parameter  $b$ . Three measures of performance are tabulated,  $p_m$ ,  $p_t$ , and  $p_d$ , each involving a comparison of the test with the test of the same size based on the sample sum, which is optimal under the basic model. In the notation of (4.1.31), (4.1.32), and (4.1.34), these measures can be written as

$$p_m = \check{\Pi}_1(\sigma) - \Pi_1(\sigma), \quad (4.1.36)$$

$$p_t = \check{\Pi}_2(b) - \Pi_2(b), \quad (4.1.37)$$

$$p_d = \check{\Pi}_4(\sigma, b) - \Pi_4(\sigma, b), \quad (4.1.38)$$

where  $\Pi_1$ ,  $\Pi_2$ ,  $\Pi_4$  relate to the robust test and  $\check{\Pi}_1$ ,  $\check{\Pi}_2$ ,  $\check{\Pi}_4$  to the optimal test. Interestingly, Veale and Kale call  $p_m$  the *premium* and  $p_t$  the *protection* involved in using the robust test, providing a natural extension of Anscombe's concepts of premium and protection in estimation discussed earlier; see (4.1.2) and (4.1.3).

## 4.2 GENERAL METHODS OF ACCOMMODATION

### 4.2.1 Estimation of location

We now consider some of the general methods that exist for constructing robust estimators, tests, or confidence intervals and give a brief review of the performance characteristics of selected procedures. In the main the techniques and results do not specifically relate to particular assumed forms for the basic model (special cases of normal and exponential basic models are given separate attention in Sections 4.3 and 4.4, respectively).

We start with two familiar, simple, and intuitively appealing procedures for inducing robustness, namely *trimming* and *Winsorizing*. These have already been mentioned in Section 2.6. The object is to control the variability due to the  $r$  lowest sample values  $x_{(1)}, \dots, x_{(r)}$  and the  $s$  highest ones  $x_{(n-s+1)}, \dots, x_{(n)}$ . The choice of  $r$  and  $s$  is discussed later; for the moment we suppose they are pre-chosen parameters. If these  $r+s$  observations are omitted, so that we confine ourselves to a censored sample of size  $n-r-s$ , we get the  $(r, s)$ -fold trimmed mean

$$\bar{x}_{r,s}^T = (x_{(r+1)} + \dots + x_{(n-s)}) / (n - r - s). \quad (4.2.1)$$

If on the other hand the  $r$  lowest sample values are each replaced by the value of the nearest observation to be retained unchanged, viz.  $x_{(r+1)}$ , and likewise the  $s$  highest by  $x_{(n-s)}$ , so that we work with a transformed sample

of size  $n$ , we get the  $(r, s)$ -fold Winsorized mean

$$\overset{W}{x}_{r,s} = (rx_{(r+1)} + x_{(r+1)} + \dots + x_{(n-s)} + sx_{(n-s)})/n. \quad (4.2.2)$$

Often the amounts of lower-tail and upper-tail trimming or Winsorizing are the same, i.e.  $r = s$ , and we have the  $r$ -fold symmetrically trimmed and Winsorized means

$$\overset{T}{x}_{r,r} = (x_{(r+1)} + \dots + x_{(n-r)})/(n - 2r), \quad (4.2.3)$$

$$\overset{W}{x}_{r,r} = (rx_{(r+1)} + x_{(r+1)} + \dots + x_{(n-r)} + rx_{(n-r)})/n. \quad (4.2.4)$$

The  $\alpha$ -trimmed means  $\overset{T}{m}(\alpha, \alpha)$  referred to in Section 2.6 are  $r$ -fold symmetrically trimmed means in which the amount of trimming is, for convenience, specified by the proportion  $2\alpha$  of the sample omitted rather than the number  $2r$  of observations. With an  $\alpha$ -trimming procedure in which  $\alpha$  has been specified beforehand, the number  $\alpha n$  of observations supposed to be trimmed at each end may not be an integer; suppose its integer part is  $r$ , so that  $\alpha n = r + f$  ( $0 < f < 1$ ). We then omit  $r$  observations at each end, and include the nearest retained observations,  $x_{(r+1)}$  and  $x_{(n-r)}$ , each with reduced weight  $1-f$ :

$$\overset{T}{m}(\alpha, \alpha) = [(1-f)x_{(r+1)} + x_{(r+2)} + \dots + x_{(n-r-1)} + (1-f)x_{(n-r)}]/n(1-2\alpha). \quad (4.2.5)$$

Similarly we can define  $\alpha$ -Winsorized means  $\overset{W}{m}(\alpha, \alpha)$ ; there is now no need for any fractional weighting, since the number of lower-tail observations Winsorized into  $x_{(r+1)}$  is  $r + f + 1 - f - 1 = r$ . Thus

$$\overset{W}{m}(\alpha, \alpha) = (rx_{(r+1)} + x_{(r+1)} + \dots + x_{(n-r)} + rx_{(n-r)})/n. \quad (4.2.6)$$

Clearly the 0-trimmed and 0-Winsorized means are both the same as the sample mean  $\bar{x}$ , and the  $\frac{1}{2}$ -Winsorized mean is the same as the sample median  $\tilde{x}$ ; the  $\frac{1}{2}$ -trimmed mean, by a suitable limiting argument, can also be taken to be  $\tilde{x}$ . The  $\frac{1}{4}$ -trimmed mean,  $\overset{T}{m}(\frac{1}{4}, \frac{1}{4})$ , is called the *mid-mean*.

What is the influence curve,  $IC_{T,F}(\xi)$ , for the  $\alpha$ -trimmed mean? If  $F$  is continuous with density  $f$  and mean  $\mu$ , and  $0 \leq \alpha < \frac{1}{2}$ , we have, in the notation of Section 4.1,

$$T(F) = \frac{1}{(1-2\alpha)} \int_{x_\alpha}^{x_{1-\alpha}} x dF \quad (4.2.7)$$

where  $x_\alpha$  denotes the  $\alpha$ -quantile of  $F$ :  $F(x_\alpha) = \alpha$ . Hence

$$T[(1-\lambda)F + \lambda G] = \left( \frac{1-\lambda}{1-2\alpha} \right) \int_{y_\alpha}^{y_{1-\alpha}} x dF + \left( \frac{\lambda}{1-2\alpha} \right) \int_{y_\alpha}^{y_{1-\alpha}} x dG \quad (4.2.8)$$

where  $y_\alpha$  is determined from

$$(1-\lambda)F(y_\alpha) + \lambda = \alpha \quad (\xi < x_\alpha), \quad (1-\lambda)F(y_\alpha) = \alpha \quad (\xi > x_\alpha),$$

with a similar definition for  $y_{1-\alpha}$ ; this gives

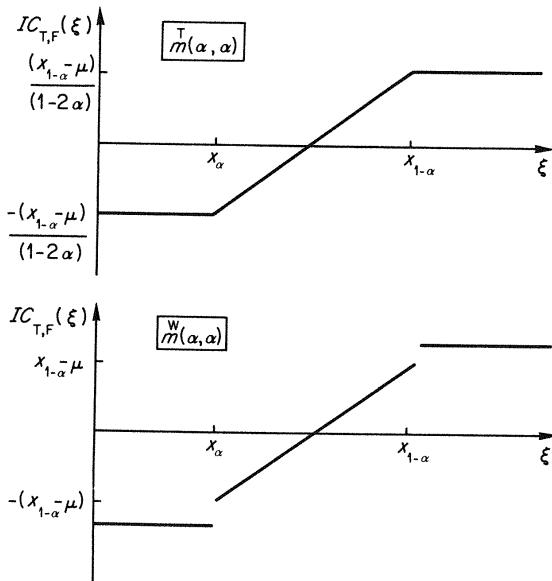
$$\begin{aligned} (\partial y_\alpha / \partial \lambda)_{\lambda=0} &= -(1-\alpha)/f(x_\alpha) \quad (\xi < x_\alpha), & \alpha/f(x_\alpha) \quad (\xi > x_\alpha) \\ (\partial y_{1-\alpha} / \partial \lambda)_{\lambda=0} &= -\alpha/f(x_{1-\alpha}) \quad (\xi < x_{1-\alpha}), & (1-\alpha)/f(x_{1-\alpha}) \quad (\xi > x_{1-\alpha}). \end{aligned}$$

For symmetric  $F$ , and in particular for normal  $F$ ,  $x_\alpha + x_{1-\alpha} = 2\mu$ , and  $f(x_\alpha) = f(x_{1-\alpha})$ . It then readily follows from (4.2.8), (4.1.16) and the relation  $\int_{x_\alpha}^{x_{1-\alpha}} x dF = (1-2\alpha)\mu$ , that the influence curve for the  $\alpha$ -trimmed mean is

$$IC_{T,F}(\xi) = \begin{cases} -(x_{1-\alpha} - \mu)/(1-2\alpha) & \text{for } \xi < x_\alpha \\ (\xi - \mu)/(1-2\alpha) & \text{for } x_\alpha \leq \xi \leq x_{1-\alpha} \\ (x_{1-\alpha} - \mu)/(1-2\alpha) & \text{for } \xi > x_{1-\alpha}. \end{cases} \quad (4.2.9)$$

This is illustrated in Figure 4.1. It shows that

the ‘influence’ of an extreme outlier on the value of a trimmed mean is *not* zero, as one would naively expect (arguing that the outlier will be ‘thrown out’); rather it is



**Figure 4.1** Influence curves for  $\alpha$ -trimmed and  $\alpha$ -Winsorized means

equal to the influence of an additional  $x$  at  $[x_\alpha \text{ resp. } x_{1-\alpha}] \dots$  the  $\alpha$ -trimmed mean does not really 'throw out' outliers, in the sense of ignoring them completely, but in effect 'brings them in' towards the bulk of the sample. But what about the  $\alpha$ -Winsorized mean which had been designed specifically to 'bring in' outliers? (Hampel, 1974)

To answer this we have, for the  $\alpha$ -Winsorized mean ( $0 \leq \alpha < \frac{1}{2}$ ),

$$T(F) = \alpha x_\alpha + \int_{x_\alpha}^{x_{1-\alpha}} x dF + \alpha x_{1-\alpha}. \quad (4.2.10)$$

Again assuming symmetric  $F$ , the influence curve now comes out to have the following form, illustrated in Figure 4.1:

$$IC_{T,F}(\xi) = \begin{cases} -[(x_{1-\alpha} - \mu) + \alpha/f(x_\alpha)] & \text{for } \xi < x_\alpha \\ \xi - \mu & \text{for } x_\alpha \leq \xi \leq x_{1-\alpha} \\ +[(x_{1-\alpha} - \mu) + \alpha/f(x_\alpha)] & \text{for } \xi > x_{1-\alpha} \end{cases} \quad (4.2.11)$$

The IC is indeed bounded, the outliers 'brought in', but there is a jump at  $[x_\alpha \text{ and } x_{1-\alpha}] \dots$ . Furthermore both slope in the center and supremum differ from that of the  $\alpha$ -trimmed mean. ... the ... point is that the mass of the tails is put on single order statistics resp. single points in the limit, and shifting them ... causes appreciable fluctuations of the Winsorized mean which are determined solely by the density in (and near) these points. A contamination in the central part, on the other hand, has the same influence as on the arithmetic mean, while the trimmed mean spreads the influence of outliers evenly over the central part, thus giving it a higher weight.

Thus ... both the trimmed mean and the Winsorized mean restrict the influence of outliers, but in different ways. While the IC of the former is always continuous, the IC of the latter is discontinuous and very sensitive to the local behaviour of the true [sic] underlying distribution at two of its quantiles. (Hampel, 1974)

A basic problem in the use of trimmed or Winsorized means is choosing the extent of trimming or Winsorization. Should we employ an asymmetric ( $r \neq s$ ), or symmetric ( $r = s$ ), scheme; how should  $(r, s)$  be chosen (or  $\alpha$  in the symmetric proportionate schemes)? No simple answers are feasible. The range (and degree of specification) of possible basic and alternative models, the variety of performance criteria which may be adopted, the dependence on sample size, and so on, all affect this choice. Some recommendations will be discussed later when we consider performance characteristics (Sections 4.2.2, 4.3, 4.4).

Some modifications of trimmed or Winsorized means change the nature of the problem of choosing the degree of trimming or Winsorization. We might, for example, contemplate eliminating (or transferring) sample values in terms of some quantitative measure of their extremeness, rather than merely on the basis of their rank order. Suppose we consider sample residuals  $z_j$  ( $j = 1, 2, \dots, n$ ), defined on some appropriate basis. For example, if the basic model is  $N(\mu, \sigma^2)$  we might use  $z_j = x_j - \bar{x}$ . Modified trimming occurs with the 'rejection rule' of Anscombe (1960a) where  $\mu$  is

estimated by

$$\begin{aligned} \bar{x} & \text{ if } |z_j| < c\sigma \text{ (all } j) \\ \overset{T}{x}_{1,0} & \text{ if } |z_{(1)}| \geq c\sigma \text{ and } |z_{(1)}| > |z_{(n)}| \\ \overset{T}{x}_{0,1} & \text{ if } |z_{(n)}| \geq c\sigma \text{ and } |z_{(n)}| > |z_{(1)}| \end{aligned}$$

for a suitable choice of  $c$ . (If  $\sigma$  is unknown it is replaced by  $s$ .)

For an alternative model of the slippage type, incorporating precisely one discordant value, the observation with maximum absolute residual is trimmed if it is *sufficiently* extreme.

Corresponding *modified Winsorization* is also contemplated: instead of trimming (rejecting) the observation with sufficiently extreme maximum absolute residual, it might be replaced by its nearest neighbour in the ordered sample. Thus Guttman and Smith (1969) suggest estimating  $\mu$  in  $\mathbf{N}(\mu, \sigma^2)$  by

$$\left. \begin{aligned} \bar{x} & \text{ if } |z_j| < c\sigma \text{ (all } j) \\ \overset{W}{x}_{1,0} & \text{ if } |z_{(1)}| \geq c\sigma \text{ and } |z_{(1)}| > |z_{(n)}| \\ \overset{W}{x}_{0,1} & \text{ if } |z_{(n)}| \geq c\sigma \text{ and } |z_{(n)}| > |z_{(1)}| \end{aligned} \right\} \quad (4.2.12)$$

for a suitable choice of  $c$  (again  $s$  replaces  $\sigma$ , if  $\sigma$  is unknown).

Another possibility, termed *semi-Winsorization* (Guttman and Smith, 1969) replaces the sufficiently extreme observation (that with largest absolute residual if this exceeds  $c\sigma$ ) with the appropriate cut-off point,  $\bar{x} - c\sigma$  or  $\bar{x} + c\sigma$ , rather than with its nearest neighbour. Again  $\mu$  is estimated by the mean of the treated sample, or by  $\bar{x}$  if  $|z_j| < c\sigma$  for all  $j$ , and  $s$  is used in place of unknown  $\sigma$ . (See Section 4.3 for further details.)

There is a growing interest in so called *adaptive methods* of statistical inference, in which the choice of inference procedures is allowed to depend in part on the actual sample to hand. Some such proposals have been made in the context of robust methods for estimation and hypothesis testing: see Hogg (1974) for a recent review of such work (together with a contributed discussion). One example, specifically concerned with trimmed means, is described by Jaeckel (1971b). Concerned with optimal choice of  $\alpha$  in the  $\alpha$ -trimmed mean for estimating the location parameter of a symmetric distribution he proposes that we choose  $\alpha$  in some permissible range  $(\alpha_0, \alpha_1)$  to minimize the sample variance  $s^2(\alpha)$  of  $\tilde{m}(\alpha, \alpha)$ . The resulting *optimal-trimmed mean*  $\tilde{m}(\hat{\alpha}, \hat{\alpha})$  is shown to be asymptotically equivalent (in terms of variance) to the best estimator  $\tilde{m}(\alpha, \alpha)$  (i.e. with minimum variance  $\sigma^2(\alpha, \alpha)$ ) provided the truly best  $\alpha$  happens to lie in the range  $(\alpha_0, \alpha_1)$ . A

modification due to Bickel is described in Andrews *et al.*, (1972). Hogg (1974) stresses the difficulty in deciding what is an appropriate measure of location for an *asymmetric* distribution and follows up a suggestion by Huber (1972) that the measure might be *defined* in terms of the limiting form of some appealing estimator. He commends the trimmed mean  $\bar{m}(\alpha_1, \alpha_2)$  for this purpose with  $\alpha_1$  and  $\alpha_2$  (which may well differ in value) chosen adaptively to minimize an estimate  $s^2(\alpha_1, \alpha_2)$  of the variance of  $\bar{m}(\alpha_1, \alpha_2)$ . Other adaptive estimators and tests will be described where appropriate in the following discussion.

Other types of robust estimator are conveniently described in terms of the three-part classification of methods for constructing estimators, outlined by Huber (1972); see Chapter 2.

### *Maximum likelihood type estimators (M-estimators)*

Huber (1964) proposes a generalization of the least squares principle for constructing estimators of (principally) location parameters. Suppose, on the basic model, that the sample comes from a distribution with distribution function  $F(x - \theta)$ . It is the location parameter,  $\theta$ , which we wish to estimate. We might estimate  $\theta$  by  $T_n = T_n(x_1, x_2, \dots, x_n)$  chosen to minimize

$$\sum_{j=1}^n \rho(x_j - T_n)$$

where  $\rho(\cdot)$  is some real valued non-constant function. As special cases we note that  $\rho(t) = t^2$  yields the sample mean,  $\rho(t) = |t|$  yields the sample median, whilst  $\rho(t) = -\log f(t)$  yields the maximum likelihood estimator (where  $f(x)$  is the density function under the basic model when  $\theta = 0$ ). If  $\rho(\cdot)$  is continuous with derivative  $\psi(\cdot)$ , equivalently we estimate  $\theta$  by  $T_n$  satisfying

$$\sum_{j=1}^n \psi(x_j - T_n) = 0. \quad (4.2.13)$$

Such an estimator is called a *maximum likelihood type estimator, or M-estimator*. Usually we restrict attention to convex  $\rho(\cdot)$ , so that  $\psi(\cdot)$  is monotone and  $T_n$  unique. Under quite general conditions  $T_n$  can be shown to have desirable properties as an estimator. If  $\rho(\cdot)$  is convex  $T_n$  is unique, translation invariant, consistent, and asymptotically normal (Huber 1964, 1967). The question of choice of  $\rho$  to achieve an ‘optimal’ *robust* estimator of  $\theta$  will be taken up at a later stage. One particular estimator with desirable properties of robustness arises from putting

$$\rho(t) = \begin{cases} \frac{1}{2}t^2 & |t| \leq \kappa \\ \kappa|t| - \frac{1}{2}\kappa^2 & |t| > \kappa \end{cases} \quad (4.2.14)$$

for a suitable choice of  $\kappa$ . It is interesting to note that this estimator is related to the *Winsorized mean*. It turns out that the estimator  $T_n$  is equivalent to the sample mean of a sample in which all observations  $x_j$  such that  $|x_j - T_n| > \kappa$  are replaced by  $T_n - \kappa$  or  $T_n + \kappa$ , whichever is the closer. (We have multiple semi-Winsorization operating at both ends of the ordered sample.)

Another  $M$ -estimator, with

$$\rho(t) = \begin{cases} \frac{1}{2}t^2 & |t| \leq \eta \\ \frac{1}{2}\eta^2 & |t| > \eta \end{cases} \quad (4.2.15)$$

can be similarly interpreted as a *trimmed mean*.  $T_n$  is now the sample mean of those observations  $x_j$  satisfying  $|x_j - T_n| < \eta$ . This extends the modified trimming above from rejection of a single extreme value to rejection of all sample values whose residuals about  $T_n$  are sufficiently large in absolute value. See Huber (1964) for details.

When the basic model involves a scale parameter (the distribution function is of the form  $F[(x - \theta)/\sigma]$ ) modified forms of  $M$ -estimator have been proposed. The estimator of  $\theta$  is a solution  $T_n$  of an equation of the type

$$\sum_{j=1}^n \psi[(x_j - T_n)/S] = 0 \quad (4.2.16)$$

where the scale parameter estimator  $S$  is robust for  $\sigma$  and is estimated either independently by some suitable scheme or simultaneously with  $\theta$  by joint solution of (4.2.16) and an equation of the form

$$\sum_{j=1}^n \chi[(x_j - T_n)/S] = 0. \quad (4.2.17)$$

Different choices for  $\psi(\cdot)$  [and for  $\chi(\cdot)$ ] yield a large assortment of  $M$ -estimators which have been discussed in the literature. One example due to Hampel (see Andrews *et al.*, 1972, or Hogg, 1974) employs Huber's  $\rho(t)$  as given by (4.2.14), i.e.

$$\psi(t) = \begin{cases} t & |t| \leq \kappa \\ \kappa \operatorname{sgn} t & |t| > \kappa \end{cases}, \quad (4.2.18)$$

with  $S$  taken as

$$\text{median } \{|x_j - \tilde{x}| \}/(0.6745) \quad (4.2.19)$$

where  $\tilde{x}$  is the sample median.

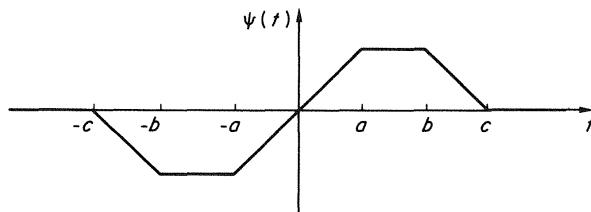
Hampel (1974) terms

$$s_m = \text{median } \{|x_j - \tilde{x}| \}$$

the *median deviation* (by analogy with the mean deviation). He outlines its earlier, but limited usage, going back as far as Gauss, and recommends it as

a quick robust scale parameter estimator and 'as a basis for the rejection of outliers'. In the present context he also advocates its use in (4.2.16) for developing *three-part descending M-estimators* where

$$\psi(t) = \begin{cases} t & |t| \leq a \\ a \operatorname{sgn} t & a < |t| \leq b \\ a(c \operatorname{sgn} t - t)/(c - b) & b < |t| \leq c \\ 0 & |t| > c. \end{cases} \quad (4.2.20)$$



A somewhat similar proposal in Andrews *et al.* (1972) employs

$$\psi(t) = \begin{cases} \sin(t/d) & |t| < d\pi \\ 0 & |t| > d\pi \end{cases} \quad (4.2.21)$$

(this is investigated for the specific choice  $d = 2.1$ .)

Related estimators based on preliminary modified Winsorization of observations whose residuals exceed  $c\sigma$  (for some choice of  $c$ ) in absolute value have also been proposed and examined (see for example, Andrews *et al.*, 1972, where they are referred to as *one-step Huber estimators*).

There is a vast range of possible *M-estimators*. Even in the cases described above a deal of choice remains in terms of how to estimate  $\sigma$  and what values to take for cut-off points such as  $\kappa$ ,  $\eta$ ,  $a$ ,  $b$ ,  $c$  and  $d$ . Many theoretical, numerical and simulation studies have been made and we shall review some of the results later (Sections 4.2.3, 4.3). Some key references are the large-scale empirical study by Andrews *et al.* (1972), also Hampel (1974), Hogg (1974), Huber (1964), Jaeckel (1971a), and Leone, Jayachandran, and Eisenstat (1967).

Before moving on to other types of estimator, however, we must note the scope here for an adaptive approach. It is reasonable to contemplate choosing, for example, relevant cut-off points in the light of the sample data. However, little work of this type seems to have been carried out to date for *M-estimators*.

Rather than employing separate estimates of  $\sigma$  in the case where  $\theta$  and  $\sigma$  are unknown, we can pursue a joint estimation process which consists of simultaneous solution of (4.2.16) and (4.2.17). The example which has received most attention is known as *Huber's proposal 2*. Huber (1964) suggested (primarily for a *normal* basic model) that we employ  $\psi(\cdot)$  as

expressed in (4.2.18) with some preliminary choice of value for  $\kappa$  and take

$$\chi(t) = \psi^2(t) - \beta(\kappa) \quad (4.2.22)$$

where

$$\beta(\kappa) = \int \psi^2(t) dt. \quad (4.2.23)$$

This form of  $\chi(t)$  was motivated by consideration of reasonable  $M$ -estimators of  $\sigma$  (see Section 4.3). The corresponding (4.2.16) and (4.2.17) need to be solved iteratively for suitably chosen starting values. The resulting estimators of  $\theta$  (and of  $\sigma$ ) have received much attention (see, for example, Andrews *et al.*, 1972; Bickel, 1965; Huber, 1964) and we shall consider them further in Sections 4.2.4 and 4.3.

#### *Linear order statistics estimators (L-estimators)*

Suppose  $x_{(1)} < x_{(2)} \dots < x_{(n)}$  denotes the ordered sample. We might estimate  $\theta$  by a linear form

$$T_n = \sum_{j=1}^n a_j x_{(j)} \quad (4.2.24)$$

of the  $x_{(j)}$  ( $j = 1, 2, \dots, n$ ). Such linear order statistics estimators have been widely studied for specific uncontaminated samples (see, for example, the lengthy review in David, 1970). Much of this work is directed to censored samples. Whilst not specifically concerned with the problem of outliers, in that the reason for censoring is seldom considered and no outlier-specific alternative model employed, some linear order statistics estimators for censored samples will possess general robustness properties which carry over to the outlier problem. There is good reason however, to consider estimators of the form (4.2.24) specifically in the context of robust estimation from possibly contaminated samples. Indeed we have already considered examples of such estimators including the sample median and trimmed and Winsorized means—all yielded by a particular choice of the  $a_i$  in (4.2.24). The modified trimmed and Winsorized means and indeed certain  $M$ -estimators have (or can be interpreted to have an analogous quasi-adaptive form in the respect that the choice of the  $a_i$  depends on the observations in the sample).

If we represent the weights  $a_j$  as

$$a_j = \int_{(j-1)/n}^{j/n} J(t) dt \quad (4.2.25)$$

for some function  $J(t)$  satisfying  $\int_0^1 J(t) dt = 1$ , we have what Huber (1972) calls *L-estimators*. Since most studies concern estimating the centre of a symmetric distribution we frequently encounter the further (natural) assumption that the weights are symmetrically valued; that is,  $a_j = a_{n+1-j}$ .

Under appropriate conditions the corresponding  $T_n$  is consistent and asymptotically normal (Chernoff, Gastwirth, and Johns, 1967; Bickel, 1967; Jaeckel, 1971a).

Let us consider some further examples of  $L$ -estimators which have been proposed and investigated.

Gastwirth and Cohen (1970) consider, primarily for a normal basic model with symmetric contamination, the estimator

$$T_n = \gamma(x_{(\lfloor pn \rfloor + 1)} + x_{(n - \lceil pn \rceil)}) + (1 - 2\gamma)\bar{x} \quad (4.2.26)$$

( $0 < p < 1$ ,  $0 < \gamma < 1$ ). This is a weighted combination of the lower and upper  $p$ th sample fractiles, each with weight  $\gamma$ , and the sample median, with weight  $1 - 2\gamma$ . It is compared with many other estimators, primarily of  $L$ -type.

A special case of (4.2.26) of the form

$$T_n = 0.3x_{(\lfloor n/3 + 1 \rfloor)} + 0.4\bar{x} + 0.3x_{(n - \lceil n/3 \rceil)} \quad (4.2.27)$$

is proposed and investigated by Gastwirth (1966). Another class of estimators based on a small number of selected ordered sample values includes the *trimean*

$$T_n = (h_1 + 2\bar{x} + h_2)/4 \quad (4.2.28)$$

where the *hinges*,  $h_1$  and  $h_2$ , are approximate sample quartiles. An adaptive form employs the notion of *skipping* (Tukey, 1977) and examples are investigated by Andrews *et al.* (1972). The hinges are taken as the lower and upper sample quartiles. Derived quantities of the form

$$\left. \begin{aligned} c_1 &= h_1 + \eta(h_2 - h_1) \\ c_2 &= h_1 - \eta(h_2 - h_1) \end{aligned} \right\} \quad (4.2.29)$$

are defined for prescribed  $\eta$  (typically 1, 1.5, or 2) and the skipping process involves deleting observations in the tails of the sample (outside the interval  $(c_1, c_2)$ ) preliminary to calculation of the *trimean* of the *retained* observations. In *iterative skipping* the process is repeated with recalculated hinges at each stage until the retained data set remains constant: *multiple skipping* repeats this process by skipping applied to the retained data set with different values of  $\eta$  at each stage.

Other adaptive  $L$ -estimators have been considered by Birnbaum and Miké (1970), Takeuchi (1971), and by Jaeckel (1971a). We should also include at this stage the '*shorth*' which is the sample mean of the shortest half of the sample (chosen as  $x_{(l)}, \dots, x_{(l + \lceil n/2 \rceil)}$  where  $l$  minimizes  $x_{(l + \lceil n/2 \rceil)} - x_{(l)}$ ), and associated estimators. (See Andrews *et al.*, 1972).

### Rank test estimators (*R*-estimators)

Within the wide range of non-parametric procedures we have methods of testing hypotheses about location parameters in (primarily symmetric) unspecified distributions, and associated estimates, which are often distribution-free and may be expected to possess various robustness properties. Specific

concern for robustness is reflected in a class of estimators (*R-estimators*) based on two-sample linear rank tests. Consider a function  $J(t)$  which is antisymmetric about 1, that is,

$$J(t) = -J(1-t).$$

For any value of  $\Delta$  we form  $x_1 - \Delta, \dots, x_n - \Delta, -x_1 + \Delta, \dots, -x_n + \Delta$  and order the  $2n$  numbers so obtained; an indicator function  $V_i$  is formed where  $V_i = 1$  if the  $i$ th smallest is of type  $x_j - \Delta$ , and  $V_i = 0$  otherwise. The *R-estimator*  $T_n$  is a solution of the equation

$$W(t) = 0 \quad (4.2.30)$$

where

$$W(t) = \sum_{j=1}^{2n} J\left(\frac{j}{2n+1}\right) V_j. \quad (4.2.31)$$

If  $J$  is monotone the solution of (4.2.30) is unique, consistent, has known variance and is asymptotically normally distributed. (See Hedges and Lehmann, 1963; Hedges, 1967; Huber, 1972; Jaeckel, 1971a.)

An asymptotically equivalent special case, based on the one-sample Wilcoxon test, which has received much attention, is the *Hedges-Lehmann* (Hedges and Lehmann, 1963) estimator. This is the median of the set of  $n(n+1)/2$  pairwise means  $(x_j + x_l)/2$  ( $j \neq l$ ;  $j = 1, 2, \dots, n$ ;  $l = 1, 2, \dots, n$ ). Whilst simple in form its calculation can be tedious if  $n$  is at all large. More easily calculable versions have been proposed, based on means of symmetrically placed ordered sample values—there are only  $[(n+1)/2]$  such means. For example we have the *folded-median* type estimators. The sample is folded by replacing  $x_1, \dots, x_n$  with  $[x_{(1)} + x_{(n)}]/2, [x_{(2)} + x_{(n-1)}]/2, \dots$ , and the median of the folded sample is chosen as the estimator (the *Bickel-Hedges estimator*). Reordering and further folding (with or without trimming) is also contemplated. See Andrews *et al.* (1972).

### *Other Estimators*

The large number of location estimators described above represents only a selection of those which have been proposed. An indication of the wider range of prospects may be found in Andrews *et al.* (1972).

Hogg (1967) comments on the use of the mean of the ‘trimmings’ in a trimmed sample (rather than the mean of the retained observations) as an estimator when the basic and alternative models all have short tails—but this is contrary to the spirit of our interest in outliers. However, his proposal for an adaptive estimator based on the form of possible models is relevant to current interests. Suppose that the data came from one of a set of possible symmetric distributions  $\{D_l\}$  ( $l = 1, 2, \dots, m$ ) each centred on  $\theta$  and that  $T_l$  is a good estimate of  $\theta$  under  $D_l$ . Hogg proposes that we adopt

$$T = \sum W_l T_l \quad (4.2.32)$$

for a suitable choice of weights  $W_i$  ( $\sum W_i = 1$ ) which are allowed to depend in value on the sample data. A special case which has received attention is the piecewise estimator

$$T = \begin{cases} \bar{m}^c(\frac{1}{4}, \frac{1}{4}) & t(\mathbf{x}) < a_1 \\ \bar{x} & a_1 \leq t(\mathbf{x}) \leq b_1 \\ \bar{m}(\frac{1}{4}, \frac{1}{4}) & b_1 < t(\mathbf{x}) \leq c_1 \\ \tilde{x} & t(\mathbf{x}) > c_1 \end{cases} \quad (4.2.33)$$

where  $\bar{m}^c(\frac{1}{4}, \frac{1}{4})$  is the mean of the set of symmetrically trimmed observations with trimming parameter  $\alpha = \frac{1}{4}$ ,  $\bar{m}(\frac{1}{4}, \frac{1}{4})$  is the corresponding trimmed mean and  $\bar{x}$  and  $\tilde{x}$  are the mean and median, respectively, of the whole sample.  $t(\mathbf{x})$  is some sample statistic and  $a_1, b_1, c_1, d_1$  are a selected set of cut-off points where we switch from one form to another. Hogg (1967) and Andrews *et al.* (1972) consider the case where  $t(\mathbf{x})$  is the sample coefficient of kurtosis and

$$a_1 = 2, b_1 = 4, c_1 = 5.5.$$

Hogg (1974) proposes a modified form replacing the sample coefficient of kurtosis with a 'better indicator of the length of the tails'. The revised estimator is felt to be better able to appropriately incorporate shorter-tailed symmetric distributions.

We shall consider in Chapter 8 some Bayesian approaches to the accommodation of outliers due to Box and Tiao (1968) for normal distributions and to Sinha (1972, 1973b) and Kale and Sinha (1971) for exponential distributions.

#### 4.2.2 Performance characteristics of location estimators

In Section 4.2.1 we reviewed the range of general robust procedures which have been proposed for estimating a location parameter, and presented a variety of special estimators. We made no attempt to examine in detail

- (i) what basic and alternative models (if any) were contemplated for any estimator;
- (ii) whether, or not, distributions and contamination were assumed to be symmetric;
- (iii) what could be claimed about the performance of estimators in respect of the variety of different performance criteria.

Space does not permit a full description of the range of published results on these matters, particularly the extensive simulation studies which have been made of the relative performances of different estimators against

different possible models and using different performance criteria. Indeed, it is doubtful whether much of the work is truly germane to a study of outliers, in that the implicit notions of robustness relate to much wider prospects (in terms of models) than we would expect to be manifest though outlying observations in a sample. We shall summarize, and give references for, those parts of the published work which have relevant general interest, or which come closest in spirit to the outlier problem.

In Section 4.1.1 we defined the notions of *minimax* and *maximin* robust estimators. The former minimizes the maximum variance over the range of possible distributions encompassed in a composite alternative model; the latter maximizes the minimum efficiency relative to corresponding 'optimal' estimators. Various types of estimators have been investigated in terms of such criteria.

Jaeckel (1971a) demonstrates an asymptotic equivalence between *M*-, *L*- and *R*-estimators for symmetric contamination of a symmetric basic model. For a given *M*-estimator, asymptotically equivalent *L*- and *R*-estimators exist. All are asymptotically normal with equal variance and share jointly in any asymptotic optimality. Huber (1964) shows that there is an *M*-estimator which minimizes the supremum, over the class  $\mathcal{C}$  of distributions of the form  $(1-\lambda)F + \lambda G$ , of the asymptotic variances. This turns out to be the optimal *M*-estimator for a particular (exhibited) member  $F_0$  of  $\mathcal{C}$ . Jaeckel (1971a) shows that this optimality extends to asymptotically equivalent *L*- and *R*-estimators. With *asymmetric* contamination of a basic symmetric distribution, *F*, the estimators are typically biased and do not converge to the centre of symmetry of *F* (Huber, 1964). Jaeckel (1971a) attempts to overcome this difficulty by using a mixture model in which  $\lambda$  is a decreasing function of *n* in the sense of (4.1.10). Under rather specific conditions he exhibits a minimax optimality result (in terms of asymptotic mean square error) analogous to that of Huber for the symmetric contamination case. See also Bickel (1965), Gastwirth (1966) and Huber (1972) on associated asymptotic behavioural and optimality considerations for *M*-, *L*-, and *R*-estimators.

Gastwirth (1966) and Crow and Siddiqui (1967) consider different classes of estimator from the *maximin* (rather than *minimax*) standpoint. For location estimators in the case of symmetric contamination it proves to be useful to identify the two 'extreme' distributions in  $\mathcal{C}$ , and the *maximin* estimator typically takes the form of a weighted average of the respective 'reasonable estimators' for these two extreme distributions, provided each estimator is reasonably efficient for the alternative extreme distribution. Gastwirth and Rubin (1969) consider *maximin* robust estimators in the specific class of linear order statistics estimators. They show under quite generous conditions that over a large class  $\mathcal{C}$  of distributions (*not* restricted to mixture type alternatives), for each of which an asymptotically efficient linear estimator exists, we can find a linear estimator which maximizes the

minimum (asymptotic) efficiency. They adopt a Bayesian decision-theoretic approach. Usually the maximin estimator is difficult to determine explicitly. One case, where  $\mathcal{C}$  has just two members (double-exponential and logistic distributions), is examined in detail and an explicit maximin estimator exhibited. This is of interest to the outlier problem only in as far as it provides an example of using an *inherent* type of alternative model. Gastwirth and Rubin further show that maximin linear estimators do not possess conspicuously higher asymptotic efficiency than simpler linear estimators and suggest further limiting the field of estimators in the cause of simplicity and at little cost. They determine the maximin estimators (for a location parameter in the case of symmetric contamination under prescribed conditions) within the classes of trimmed means and of linear combinations of sample percentiles. Special cases considered yield high trimming factors: for example, for Cauchy versus normal distributions we need  $\alpha = 0.275$  so that only the middle 45 per cent of the sample is retained! Crow and Siddiqui (1967) present other numerical comparisons.

The minimax (or maximin) approach is not ideal in two respects. It is bound to be a highly pessimistic policy—we pay a lot to protect against the most extreme prospects. Its asymptotic nature gives little clue to the finite sample behaviour of estimators.

In a different vein Hampel (1974) discusses in some detail the various measures of robustness based on the influence curve illustrating how different types of estimator stand up on such criteria. See also Hampel (1971).

Hogg (1974) reviews and considers some performance characteristics of various *adaptive* robust estimators. See also Jaeckel (1971b) on adaptive L-estimators.

Comparisons of the finite-sample, and asymptotic, behaviour of a large variety of specific location estimators have been widely undertaken. Various performance characteristics are employed, often examined by simulation methods. Some results centred on the normal distribution as basic model will be summarized later (Section 4.3). References to work not specific to the normal distribution include Bickel (1965), Crow and Siddiqui (1967), Gastwirth and Cohen (1970), Hogg (1967), and Siddiqui and Raghunandan (1967). But undoubtedly the *tour de force* is the study by Andrews *et al.* (1972) of 68 different location estimators in terms of various asymptotic characteristics as well as a variety of finite sample characteristics for samples of sizes 5, 10, 20, and 40 for a large range (c. 20) of different possible data generating models. They also consider such matters as the relative ease of computation of the estimators.

We make no attempt to present a short recommended list of robust location estimators for general use. Some sets of ‘best buys’ are offered: see for example Andrews *et al.* (1972; in Chapter 7 each contributor tackles the unenviable task of summarizing the vast amount of information—some are even brave enough to specify tentative choices of ‘best estimator’) or Hogg

(1974). But wealth of detail and inconsistencies of relative performance are not the major reasons for hesitating to recommend particular robust estimators at this stage. The problem is that general studies of robustness are seldom specific to our theme: *outliers*. They reflect a more general concept of robustness where alternative models encompass widely differing distributions not promoted solely or even primarily by a desire to reflect outliers. Undoubtedly the procedures that have been advanced will include many which will prove valuable on more specific investigation (which we hope will materialize) from the outlier standpoint. This belief explains the lengthy review we present in this chapter. In Section 4.3 some more detailed prescriptions will be offered for a *normal* basic model with specific mixture or slippage types of alternative model. These are intrinsically closer in relevance to the problem of accommodating outliers.

#### 4.2.3 Estimation of scale or dispersion

Far less attention has been given to robust estimation of a scale or dispersion parameter,  $\sigma$ , than to robust estimation of location parameters. What few contributions exist again are not specific to the outlier accommodation issue.

We have remarked in Section 4.2.1 on one interesting estimator based on the *median deviation*,

$$s_m = \text{median} \{ |x_i - \bar{x}| \}.$$

By its nature we might expect it to provide reasonable protection against the influence of discordant values in the sample—this is touched on by Hampel (1974) who discusses the pedigree and performance of the estimator. It arises ‘as the *M*-estimate of scale with the smallest possible gross-error-sensitivity at the normal (and many other) models’. Hampel regards it as the counterpart of the median as location estimator, and he discusses the form of its influence function: in particular the gross-error-sensitivity and breakdown point. In these respects the median deviation is seen to be more desirable than the *semi-interquartile-range*,  $Q$  (although an equivalence exists for symmetric samples and, asymptotically, for a symmetric basic model).  $Q$  might also be expected to protect against outliers, but it may well be over-protective.

In spite of the relatively low efficiency of  $s_m$  for uncontaminated data (c. 40 per cent in the normal case) it appears to possess a robustness absent from estimators which are more efficient for homogeneous data. See Tukey (1960) and Stigler (1973b).

Andrews *et al.* (1972), in studying the characteristics of some rather attractive robust location estimators (such as the Huber ‘proposal 2’ estimate with  $\kappa = 1.5$ , the Bickel one-step modification with Winsorization of residuals at  $\pm \kappa s$ , where  $s$  is a robust scale estimate, and Hampel’s three-part descending *M*-estimator with one of the cut-off points determined from a

robust scale estimator), provide evidence to suggest that using a scale estimate based on  $s_m$  (their ‘robust scale estimate’) is preferable to using one based on  $Q$ .

A variety of scale estimators based on a Winsorized sample have been proposed both for robust estimation of a scale parameter,  $\sigma$ , *per se*, and, more often to obtain a robust studentized test statistic (see Section 4.2.4).

Thus if we effect  $(r, r)$  Winsorization of the ordered sample  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  we obtain

$$\underbrace{x_{(r+1)}, x_{(r+1)}, \dots, x_{(r+1)}}_{r+1 \text{ times}}, \underbrace{x_{(r+2)}, \dots, x_{(n-r-1)}, x_{(n-r)}, x_{(n-r)}, \dots, x_{(n-r)}}_{r+1 \text{ times}}.$$

Rewriting these as  $y_{(1)}, y_{(2)}, \dots, y_{(n)}$ , we might consider  $\frac{w}{S_{r,r}^2} = \frac{w}{S_{r,r}^2}/(n-1)$  with

$$\frac{w}{S_{r,r}^2} = \sum_{j=1}^n (y_{(j)} - \bar{y})^2 \quad (4.2.34)$$

as an estimator of  $\sigma^2$  where  $\bar{y} = \frac{w}{S_{r,r}}$ : the  $(r, r)$  Winsorized mean. See for example Dixon and Tukey (1968). Clearly the propriety of such a procedure will depend strongly on how many discordant outliers there are at each end of the sample relative to the prescribed  $r$ . Choice of  $r$  (as for location-estimation) is crucial, although likely to be even more so if we contemplate using a trimmed rather than a Winsorized sample. But here we have added risk of under-estimation due to deletion of ‘respectable’ extreme values in the sample. We might also consider asymmetric Winsorization in the outlier context, especially for a scale parameter in an asymmetric basic model.

#### General linear forms

$$\tilde{\sigma} = \sqrt{\sum_1^n b_i x_{(i)}} \quad (4.2.35)$$

have also been studied, including censored (trimmed) or Winsorized equivalents (see, for example, Sarhan and Greenberg, 1962, pp. 218–251; David, 1970, pp. 109–124). The interest in this work lies predominantly in what loss occurs relative to the full sample equivalent for a *prescribed distribution*. What little reference there is to robustness is not specific, nor particularly relevant, to accommodating outliers. It is a little surprising that no study of (4.2.35) seems to have been made in relation to accommodation of outliers, when the approach proves so fruitful for censored data from a homogeneous source. Tukey (1960) points out that even for small contaminations ( $\lambda \sim 0.01$ ) of a basic distribution  $N(\mu, \sigma^2)$  with a contaminating distribution  $N(\mu, b\sigma^2)$ , where  $b > 1$ , the relative advantage of the sample standard deviation over the *mean deviation* which holds in the uncontaminated situation is dramatically reversed. See also Section 4.3.

Dixon (1960) describes estimation of  $\sigma$  based on the range of trimmed samples and shows (in the normal case) that they can have relative linear efficiency in excess of 96 per cent. But there is again no consideration of their robustness (no alternative model is contemplated).

Huber (1970) considers estimators of dispersion based on rank tests, on sums of squares of order statistics and on his earlier ‘proposal 2’—all with a view to determining an appropriate studentized form for a robust location estimator. (See Section 4.2.4.) The latter approach does merit further comment here. In Huber’s proposal 2 we have a method of simultaneously estimating  $\theta$  and  $\sigma$  via (4.2.16–18) and (4.2.22). This proposal was prompted by the desire to obtain a robust estimator  $S$  of  $\sigma$ , as well as a robust estimator  $T$  of  $\theta$ , in the context of a mixture-type alternative model  $(1-\lambda)F + \lambda G$  with symmetric  $F$ . Huber (1964) remarks of the procedure

It corresponds to Winsorizing a variable number of observations: slightly more if ...  $[(1-\lambda)F + \lambda G]$  has heavier tails, slightly less if ... [it] has lighter tails, and if ...  $[G]$  is asymmetric, more on the side with heavier contamination.

$S$  (and  $T$ ) are readily determined iteratively but typically possess no simple explicit form.

Huber (1964) also considers robust estimation of  $\sigma$  alone for the mixture-type model. He restricts his detailed proposals to the case where  $F$  is normal, and the details are best deferred until Section 4.3: likewise methods due to Guttman and Smith (1971) based on the Anscombe (1960a) premium-protection approach.

#### 4.2.4 Studentized location estimates, tests, and confidence intervals

If  $x_1, x_2, \dots, x_n$  is a random sample from  $N(\mu, \sigma^2)$  where  $\mu$  and  $\sigma^2$  are unknown we have the familiar test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (4.2.36)$$

for testing the hypothesis  $H: \mu = \mu_0$  against  $\bar{H}: \mu \neq \mu_0$ . The null-distribution of  $t$  is of course Student’s  $t$  with  $n - 1$  degrees of freedom. It is natural to ask how reasonable the  $t$ -test would be if we were wrong in attributing the sample to the normal distribution. This is an example of the robustness of significance tests. The  $t$ -test is not the only test we might wish to examine from this standpoint, but it is illustrative and has been given much attention in the literature. Various matters need to be more fully specified—in particular, we need to declare the types of departure from normality to be entertained and what criteria of robustness might be appropriately applied. As in all aspects of the study of robustness only certain types of alternative model are relevant to the outlier problem.

One particular approach (among many possible ones) has been widely studied. Given that the data may have arisen from a variety of possible

distributional sources, can we replace  $\bar{x}$  and  $s$  in (4.2.36) by (robust) estimators  $T$  and  $S$  in such a way that the corresponding statistic is still distributed essentially as Student's  $t$  over the range of contemplated distributions? This requires a decision on what is to be our criterion of the 'Student's  $t$ -ness' of the statistic. (The broader issue of *ab initio* generation of a robust test of location does not seem to have received much attention.)

On the particular approach it seems almost inevitable that the basic distribution  $F$  should be normal, or at very least symmetric. Neither is it surprising that the range of distributions contemplated often consists of a set of symmetrically contaminated versions of  $F$  of a mixture type with normal (or symmetric) contaminating distributions. This, of course, bears on our interest in outliers—although occasionally wider families of distributions have been entertained with little outlier relevance.

Huber (1970) summarizes the basic aims in examining a studentized version of a robust estimator  $T$  of a location parameter  $\theta$ , and examines various possibilities. We are interested in

$$(T - \theta)/S(T) \quad (4.2.37)$$

where, in the main, we would hope to achieve high *robustness of performance* for  $T$ , and to then 'match'  $T$  with an estimated standard error  $S(T)$  yielding high *robustness of validity* over a range of possible distributions for the sample. *Jackknifing* may assist in obtaining an estimated variance for  $T$  (see Section 2.6). But our interest in determining the approximate distributional form of quantities (4.2.37) under a range of distributions supports the use of a simpler, tractable, estimator  $S(T)$ . We can often expect asymptotic normality for (4.2.37) but the question is when does this occur and how fast is the approach to normality. These are complicated matters from the technical standpoint and highly dependent on the contemplated range of distributions. Asymptotic normality of the numerator  $(T - \theta)$  is commonly encountered as we have already noted; finite sample approximate normality has been examined by Hodges (1967) and by Leone, Jayachandran, and Eisenstat (1967). For the studentized form it is also important to investigate the consistency of  $S^2(T)$  as an estimator of the variance of  $T$  and also whether  $S^2(T)$  and  $T$  become essentially independent. See Huber (1970, 1972). In the former work Huber considers such problems in relation to *L*-, *M*-, and *R*-estimators of  $\theta$ , providing some prescriptions and conjectures. See also Hodges and Lehmann (1963) (rank test procedures).

By analogy with the normal theory case it is also relevant to ask if  $(T - \theta)/S(T)$  can have its distribution approximated in finite samples by the *t*-distribution, and if so what form  $S(T)$  should take and what is an appropriate number of degrees of freedom. Tukey and McLaughlin (1963) suggest a symmetrically *trimmed* mean  $\bar{x}_{r,r}^T$  for  $T$  and tentatively conclude that this is best matched with  $S^2(T)$  based on the related *Winsorized* sample, in an attempt to give greater (but not excessive) importance to the more

extreme observations. Specifically they suggest using

$$\stackrel{T}{(x_{r,r} - \theta) / \{ \stackrel{W}{S}_{r,r} / \sqrt{[h(h-1)]} \}} \quad (4.2.38)$$

as Student's  $t$  on  $(h-1)$  degrees of freedom, where  $h = n - 2r$ . This proposal is briefly examined in relation to (mainly) normal and uniform samples. A similar proposal, with support in terms of the influence function, is advanced by Huber (1972). Huber (1970) augments Tukey and McLaughlin's small-sample investigations of (4.2.38) by demonstrating useful asymptotic properties. He reinforces the relative disadvantages of *common* matching of *trimmed* means and scale estimates, or *Winsorized* means and scale estimates (but see Section 4.3 on the latter combination). Huber also shows that lower degrees of freedom than  $(h-1)$  for (4.2.38) would be appropriate for long-tailed distributions.

As ever, choice of the trimming factor  $r$  is a problem. Tukey and McLaughlin (1963) suggest an adaptive approach, choosing  $r$  to minimize  $\stackrel{W}{S}_{r,r}^2 / [h(h-1)]$  or some allied quantity but they warn about loss of stability of the denominator term in (4.2.38) with an extensive amount of trimming.

We shall consider in more detail (Section 4.3) the use of a studentized form of Huber's 'proposal 2' and of the Hodges-Lehmann estimator for a normal model symmetrically contaminated by mixing with a more disperse normal distribution.

Huber (1968) considers some fundamental matters to do with robust testing of monotone likelihood ratio alternative hypotheses, and develops minimax test procedures for various classes of problem. He extends his tests in a natural way to the construction of confidence limits.

The determination of robust confidence intervals for location or scale parameters can be approached in terms of almost all the robust estimation methods that we have considered, *provided* we know the sampling distribution of the estimator. But here lies the major obstacle: seldom do we know enough about the sampling distribution, nor is it sufficiently constant in form over the range of contemplated data-generating models. Although not specifically directed to outlier-type contamination the range of non-parametric procedures do yield direct confidence intervals which may in their general robustness provide some basis for accommodating outliers. For a general discussion see Noether (1974). Further details are included in Lehmann (1975) and Noether (1967, 1973).

The asymptotic normality properties of  $L$ -,  $M$ -, and  $R$ -estimators all provide means for determining approximate confidence intervals, but their finite sample properties are little understood. The more accessible forms for location parameters are typified by that which derives from studentized location estimates. From (4.2.38), for example, we obtain a central confidence interval symmetric about  $\stackrel{T}{x}_{r,r}$  with width an appropriate multiple (in terms of the approximating  $t$ -distribution) of  $\stackrel{W}{S}_{r,r} / \sqrt{[h(h-1)]}$ .

### 4.3 ACCOMMODATION OF OUTLIERS IN UNIVARIATE NORMAL SAMPLES

From the vast amount of material on general robustness reviewed above we now select for more detailed comment some results which relate most closely to our theme: the accommodation of outliers. We shall be concerned with robustness in relation to families of distributions which correspond with the alternative models for outliers previously discussed. The information below sometimes represents specific study of outlier-type families of distributions—more frequently it is obtained by judicious selection from more embracing robustness studies. It will sometimes prove convenient to use abbreviated notation to describe certain estimators and in such cases we adopt that used by Andrews *et al.* (1972).

In the present section we concentrate on the case of a basic *normal* model  $F$ , with either a *mixture type* alternative

$$(1 - \lambda)F + \lambda G$$

( $G$  is also usually normal with the same mean as  $F$  but larger variance) or a *slippage type* alternative (where the slippage occurs either in the mean or in the variance). In the concluding section we review results for the exponential distribution with an *exchangeable* type of alternative model.

#### *Mixture model; estimation of the mean*

Suppose our sample arises from  $(1 - \lambda)F + \lambda G$  when  $F$  is  $N(\mu, \sigma^2)$  and we wish to estimate  $\mu$ . Huber (1964) shows that the  $M$ -estimator with

$$\psi(t) = \begin{cases} t & |t| \leq \kappa \\ \kappa \operatorname{sgn} t & |t| > \kappa \end{cases} \quad (4.3.1)$$

is minimax among translation invariant estimators for symmetric  $G$ . Tabulated values suggest that choice of  $\kappa$  is not highly critical—performance being fairly insensitive and reasonable over the range  $1 \leq \kappa \leq 2$  for  $\lambda < 0.2$ . For asymmetric  $G$  the above estimator proves to be biased. Huber discusses the extent of the bias and concludes that attempts at substantially reducing the bias may be quite costly in terms of asymptotic variance. The straightforward  $M$ -estimators implicitly depend on knowing the values of  $\sigma^2$  and of  $\lambda$ . Scale invariant versions (allowing more realistically for unknown  $\sigma^2$ ) have been discussed in Section 4.2.1. These include estimators where  $\sigma^2$  is estimated robustly, perhaps in terms of median deviation or interquartile range and possibly involving associated Winsorization or having multi-part form.

Alternatively,  $\mu$  may be estimated simultaneously with  $\sigma^2$ , as for example in Huber's 'proposal 2' (that is solving (4.2.16) and (4.2.17) with  $\chi(t)$  given by (4.2.22) and (4.2.23) and with  $\psi(t)$  as in (4.3.1)). Some quantitative features of this approach, for symmetric  $G$ , are given by Huber (1964) for a

range of values of  $\kappa$  and  $\lambda$ . Minimax optimality, and asymptotic variances, are highly limited criteria, but it is of some relevance (if not specific to outliers) to note that various  $M$ -type estimators find broad support on various bases in the study by Andrews *et al.* (1972). Although no simple prescription of 'best estimator' is feasible, the different contributors tend to include among their recommendations  $M$ -estimators and *one-step Huber estimators* (both using (4.3.1), median deviation to estimate  $\sigma$  and with  $\kappa$  in the vicinity of 1.5), *Huber's proposal 2* (with somewhat smaller  $\kappa$ ) and the *three-part descending estimators* (with  $a, b, c$  in the regions of 2, 4, and 8, or with  $a$  chosen adaptively and  $b, c$  in the regions of 4 and 8). From Siddiqui and Raghunandanan (1967) we can make a limited asymptotic comparison of the *Hodges-Lehmann estimator, trimmed and Winsorized means* and the estimator (4.2.26): that is,

$$T_n = \gamma(x_{\lfloor pn \rfloor + 1} + x_{(n - \lfloor pn \rfloor)}) + (1 - 2\gamma)\bar{x}.$$

With the mixture model where  $F$  and  $G$  are  $N(\mu, \sigma^2)$  and  $N(\mu, 9\sigma^2)$ , respectively, and the mixing parameter  $\lambda$  is restricted to at most 0.05, there is little to choose between the first three estimators (the best trimming factor has the value in the region of  $\alpha = 0.20$ ) with minimum efficiency about 95 per cent, almost 10 per cent higher than that of the best form of (4.2.26). (It is worth noting that Gastwirth's version (4.2.27) achieves an efficiency of about 80 per cent or more for the set of *inherent* alternatives: normal, Cauchy, double-exponential, logistic. See Gastwirth, 1966.)

Asymptotic properties need however to be augmented with information on finite sample behaviour.

In Gastwirth and Cohen (1970) we find tables of means, variances, and covariances of order statistics for samples of sizes up to 20 from contaminated normal distributions

$$(1 - \lambda)F + \lambda G$$

where  $F$  is  $N(0, 1)$ ,  $G$  is  $N(0, 9)$  and  $\lambda = 0.01, 0.05, 0.10$ . These are useful for comparing the performance of order-statistics-based robust linear estimators of the mean in the corresponding range of contaminated normal distributions. The authors tabulate some results from which we see that over the types of estimator they consider (including mean, median, trimmed means, Winsorized means, combinations of the median with equally weighted fractiles of the form (4.2.26), and the Hodges-Lehmann estimator), it is again the trimmed means which perform well in terms of minimax variance both asymptotically and at the different finite sample sizes: typically (for  $n = 20$ ) needing trimming factor  $\alpha$  in the regions of

$$\begin{aligned} 0.15-0.20 & \quad (0 \leq \lambda \leq 0.1), \\ 0.10-0.15 & \quad (0 \leq \lambda \leq 0.05). \end{aligned}$$

(Note that published support for  $\alpha$ -values as high as 0.25–0.30 is based on

minimax performance over wide-ranging families of *distinct* distributions—normal to Cauchy. Such a catholic situation does not accord with the outlier models we have been considering.)

Some Monte Carlo results reported by Huber (1972), from the Andrews *et al.* (1972) work, are illuminating. For samples of size 20 he compares 20 estimators in terms of their estimated variance when  $(20 - k)$  observations come from  $N(0, 1)$  and  $k$  come from  $N(0, 9)$ . This amounts to a scale-slippage type model. When  $k = 1$  some estimators show up better than others: the  $\alpha$ -trimmed mean with  $\alpha = 0.05$  or  $0.10$ , H20 and H15 (Huber's 'proposal 2' with  $\kappa = 2.0$  or  $1.5$ ) A15 and P15 (Huber *M*-estimates, with  $\kappa = 1.5$  and median dispersion scale estimate, in direct form, and in one-step form starting with the median) and 25A (Hampel's three-part descending estimator with  $a = 2.5$ ,  $b = 4.5$ ,  $c = 9.5$ ). For  $k = 2, 3$  the  $\alpha$ -trimmed means remain impressive with  $\alpha$  advancing to  $0.1$  or  $0.15$ , and  $0.15$ , respectively, as do H15 and H10, respectively, and, for  $k = 2$  alone, A15, P15, 25A. With 18 observations from  $N(0, 1)$  and two from  $N(0, 100)$ , 25A stands out as better than most other estimators. Often the estimated variances show only small differences and to put the above recommendations in perspective it is useful to examine Table 4.2 extracted from the tabulated results in Huber (1972).

We should also bear in mind the computational effort involved in constructing the estimators. Trimmed means are fairly easily determined, and whilst possibly needing some iteration the various Huber-type estimators (such as H15, A15, P15) are not unreasonable. In comparison, the Hodges-Lehmann estimator can be most time-consuming.

Hodges (1967) uses Monte Carlo methods to examine the extent to which some simple location estimators are efficient with respect to estimating the mean of a normal distribution and are able to 'tolerate extreme values' in the sense of not being influenced by the  $r$  lowest and  $r$  highest extremes. Thus  $\bar{x}_{r,r}^T$  and  $\bar{x}_{r,r}^W$  have 'tolerance'  $r$ , the median  $\tilde{x}$  has tolerance  $[(n-1)/2]$ ,  $\bar{x}$  has tolerance 0. A modified more easily calculated type of Hodges-Lehmann estimator, BH, is the median of the means of symmetrically chosen pairs of ordered observations (the *Bickel-Hodges folded median*; see also Bickel and Hodges, 1967). BH has tolerance  $[(n-1)/4]$  and is shown by sampling experiments with  $n = 18$  to have efficiency about 95 per cent relative to  $\bar{x}$ . (But this needs careful interpretation—there is no contemplation of an alternative outlier generating model, we are merely estimating  $\mu$  with reduced consideration of extreme values whether or not they are discordant. We do not learn how BH compares with other estimators for a prescribed mixture- or slippage-type model).

Other performance characteristics are also important, for example those based on the influence curve including such features as gross-error-sensitivity, local-shift-sensitivity, and rejection point. Hampel (1974) tabulates such quantities for many of the estimators we have discussed assuming

Table 4.2 Monte Carlo variances of  $n^{\frac{1}{2}}T_n$  for selected estimators and distributions; sample size  $n = 20$ 

		$N(0, 1)$		$(n - k)N(0, 1)$ plus $kN(0, 9)$ , $n = 20$			$18N(0, 1)$ plus $2N(0, 100)$
		$n = \infty$	$n = 20$	$k = 1$	$k = 2$	$k = 3$	
Trimmed mean	Mean	1.00	1.00	1.40	1.80	2.20	10.90
	$\alpha = 0.05$	1.026	1.02	1.16	1.39	1.64	2.90
	$\alpha = 0.10$	1.060	1.06	1.17	1.31	1.47	1.46
	$\alpha = 0.15$	1.100	1.10	1.19	1.32	1.44	1.43
	$\alpha = 0.25$	1.195	1.20	1.27	1.41	1.50	1.47
	median	1.571	1.50	1.52	1.70	1.75	1.80
Huber (1964) prop. 2	$\kappa = 2.0$	1.010	1.01	1.17	1.41	1.66	1.78
	$\kappa = 1.5$	1.037	1.04	1.16	1.32	1.49	1.50
	$\kappa = 1.0$	1.107	1.11	1.21	1.34	1.44	1.43
	$\kappa = 0.7$	1.187	1.20	1.27	1.42	1.49	1.47
Hodges-Lehmann Gastwirth (1966) Jaekel (1969) Hogg (1967) Takeuchi (1971)		1.047	1.06	1.18	1.35	1.50	1.52
		1.28	1.23	1.30	1.45	1.52	1.50
		1.000	1.10	1.21	1.37	1.47	1.45
		1.000	1.06	1.28	1.56	1.79	1.79
		1.000	1.05	1.19	1.38	1.53	1.32
A15		1.037	1.05	1.17	1.33	1.47	1.49
		1.037	1.05	1.17	1.33	1.47	1.49
Hampel 25A Hampel 12A		1.025	1.05	1.16	1.32	1.49	1.26
		1.166	1.20	1.26	1.40	1.47	1.32

an uncontaminated normal distribution. Though we would be better armed for current purposes if the distribution were of a mixture or slippage type, the results are interesting. One comment, in particular, adds to the summary above:

three-part descending  $M$ -estimators pay a small premium in asymptotic variance of gross-error-sensitivity, as compared with Huber estimators, in order to be able to reject outliers completely. (Hampel, 1974)

A detailed study of the accommodation of outliers in slippage models is presented by Guttman and Smith (1969, 1971) and Guttman (1973a). They consider three specific methods based on modified trimming, modified Winsorization, and semi-Winsorization (the 'A-rule', after Anscombe, 1960a; the 'W-rule' and the 'S-rule') for estimating the mean (Guttman and Smith 1969; Guttman 1973a) and the variance (Guttman and Smith, 1971) for a normal basic model  $N(\mu, \sigma^2)$  with a location, or scale, slippage alternative model to explain the behaviour of one or two observations. Performance characteristics are restricted to the premium-protection ideas of Anscombe

(1960a), that is, variance ratios or relative efficiencies. (See Sections 2.6 and 4.1.1.)

Consider first the case where we wish to estimate  $\mu$  robustly under Ferguson's *model A* for slippage of the mean or *model B* for slippage of scale. Here we assume that  $x_1, x_2, \dots, x_n$  arise from  $N(\mu, \sigma^2)$ , but entertain the prospect that at most one observation may have arisen from  $N(\mu + a, \sigma^2)$  (*model A*) or from  $N(\mu, b\sigma^2)$  with  $b > 1$  (*model B*). In either case the three robust estimators considered are those described in Section 4.2.1, namely the *modified trimmed mean*, the *modified Winsorized mean* and the *semi-Winsorized mean* which we denote  $T_A$ ,  $T_W$ , and  $T_S$ . Guttman and Smith (1969) determine and compare the finite-sample premium and protection measures for these estimators. Detailed results are presented for the case where  $\sigma^2$  is known. When  $\sigma^2$  is unknown (and replaced by the full-sample unbiased variance estimate  $s^2$ ) computational difficulties restrict the amount of information readily obtainable.

Under the basic model  $\bar{x}$  is optimal for  $\mu$ . Putting  $\check{\mu} = \bar{x}$  in (4.1.2) and (4.1.3) (with obvious modification of the latter to allow for any bias in the typical candidate estimator  $T$ ) we can determine the premium and protection measures for  $T_A$ ,  $T_W$ , and  $T_S$ . The premiums have the general form:

$$\text{Premium} = nE(U^2)/\sigma^2 \quad (4.3.2)$$

when  $T$  is re-expressed as  $\bar{x} + U$ . The protection is

$$\{E[(\bar{x} - \mu)^2] - E[(T - \mu)^2]\}/E[(\bar{x} - \mu)^2]$$

evaluated under the alternative hypothesis, and for *model A* and *model B*, respectively, we have:

$$\text{Protection} = \begin{cases} -n^2 E[U(U+2a\sigma)/n]/[\sigma^2(n+a^2)], \\ -n^2 E[U(U+2\bar{x}-2\mu)]/[\sigma^2(n+b-1)]. \end{cases} \quad (4.3.3)$$

$$(4.3.4)$$

To determine (4.3.2) and (4.3.3) or (4.3.4) we have to investigate the first two moments of the incremental estimators  $U_A$ ,  $U_W$ , and  $U_S$  under the basic and alternative models. Simple closed form expressions are not available, but Guttman and Smith (1969) develop an appropriate computational (Monte Carlo type) procedure and provide graphs and tables for comparing  $T_A$ ,  $T_S$ , and  $T_W$  for sample sizes up to 10 at premium levels of 5 per cent and 1 per cent and for different values of the slippage parameters  $a$  and  $b$ .

The general conclusions are that under *model A*  $T_S$  is best for small  $a$ ,  $T_W$  for intermediate  $a$ , and  $T_A$  for large  $a$ . Under *model B*,  $T_A$  is not a contender;  $T_S$  is best for small  $b$ ,  $T_W$  for large  $b$ . We need to recognize the limitations of these results. Comparisons are purely relative within the set  $\{T_A, T_W, T_S\}$ ; we do not know how these estimators compare with others. The values of the slippage parameters  $a$  or  $b$  will be unknown, so choice within the set is problematical. Only sample sizes up to  $n = 10$  are considered in detail. The assumption that  $\sigma^2$  is known is unrealistic; only for the

case  $n = 3$  are any results available when  $\sigma^2$  is unknown. Restriction to 5 per cent and 1 per cent premium levels is arbitrary; we need to learn from experience and intercomparison of different measures if these levels are reasonable in practical terms. The estimators are defined in terms of cut-off values  $c$  (modification of residuals takes place if they exceed  $c\sigma$  in absolute value).  $c$  needs to be determined in any situation. It depends on the chosen premium, the sample size and the type of estimator. Guttman and Smith (1969) tabulate approximate values of  $c$  for premiums of 5 per cent and 1 per cent and  $n = 3, 4(2)10$ .

Extensions to larger sample sizes (encompassing the prospect of one or two discordant values) are considered by Guttman (1973a), again principally for the case of known  $\sigma^2$ . An interesting feature of this work is the replacement of the residuals by adjusted (independent) residuals to facilitate the calculation of premium and protection in larger sample sizes. See also Tiao and Guttman (1967). (The multivariate case is considered in Section 7.3.1.) Under the basic model the residuals  $x_j - \bar{x}$  ( $j = 1, 2, \dots, n$ ) have common variance  $(n-1)\sigma^2/n$ , and covariance  $-\sigma^2/n$ . Thus if  $u$  is an observation from  $N(0, 1)$ , independent of the  $x_j$ , the *adjusted residuals*

$$z_j = x_j - \bar{x} + \sigma u / \sqrt{n} \quad (4.3.5)$$

are *independent* observations from  $N(0, \sigma^2)$ . For reasonable sample sizes the  $z_j$  will differ little from the true residuals  $x_j - \bar{x}$ ; the induced independence, however, renders the determination of performance measures of the corresponding  $T_A$ ,  $T_S$ ,  $T_W$  more tractable and enables some quantitative comparisons to be made. See Guttman (1973a) for details.

Another area in which some detailed numerical studies have been made specifically for a normal basic model is that of studentized location estimators of the form  $\sqrt{n}(T - \mu)/S$ . Dixon and Tukey (1968) study by qualitative arguments and Monte Carlo methods the sampling behaviour of

$$t_W = (\bar{x}_{r,r} - \mu) / \sqrt{\{S_{r,r}^2 / [n(n-1)]\}}. \quad (4.3.6)$$

They conclude that

$$(h-1)t_W / (n-1)$$

has a distribution which is well approximated by Student's  $t$  distribution with  $h-1$  degrees of freedom ( $h = n-2r$ ). No consideration is given to how the distribution of  $t_W$  changes over, say, a mixture model  $(1-\lambda)F + \lambda G$  where  $F$  and  $G$  are similarly centred but differently scaled normal distributions, which would be germane to the outlier problem.

Just such a mixture model is examined, however, by Leone, Jayachandran, and Eisenstat (1967). Again by Monte Carlo methods, they examine the sampling behaviour of studentized forms,  $\sqrt{n}(T - \mu)/S$ , of robust location estimators, for the mixture model  $(1-\lambda)F + \lambda G$  with  $\lambda = 0.05$  and  $0.10$  and  $F$  and  $G$  both normal. Symmetric and asymmetric contamination

are considered: specifically  $F$  is  $\mathbf{N}(\mu, \sigma^2)$  and  $G$  is  $\mathbf{N}(\mu + a, b\sigma^2)$  with  $a = 0, \frac{1}{2}, 1$  and  $b = 1, 9, 25$ . The estimators considered are various joint estimators ( $T, S$ ) obtained under Huber's proposal 2. The goodness-of-fit of  $\sqrt{n}(T - \mu)/S$  to a Student's  $t$  distribution was examined for samples of size  $n = 20$  (also studied is the extent to which their  $T$ -estimators, and the Hodges-Lehmann estimator, have approximate normal distributions).

Broad conclusions from a mass of empirical results include the following.

- (i) For proximity of  $\sqrt{n}(T - \mu)/S$  to Student's  $t$  distribution over the contemplated range of models, reasonable choice of  $\kappa$  in Huber's proposal 2 is in the region of 1.8 or 1.9. The fit is reasonable except for extreme cases such as  $\lambda = 0.1, b = 25$ .
- (ii) The Huber ( $\kappa = 1, 1.5, 2$ ) and Hodges-Lehmann location estimators are reasonably normal, except again in extreme cases, e.g.  $\lambda = 0.1, a = 1, b = 25$ .

### *Dispersion estimators*

There has been little detailed study of robust dispersion estimators *per se*, either in relation to a general family of distinct distributional models or in relation to specific cases such as families of mixed, or slipped, normal distributions. We have *en passant* referred to particular estimators based on the interquartile range,  $Q$ , the median deviation (e.g. Hampel, 1974)

$$s_m = \text{median} \{ |x_{(j)} - \bar{x}| \}$$

use of quasi-ranges (but specifically for the *uncontaminated* normal-distribution: Dixon, 1960) and quadratic measures, using trimmed or Winsorized, samples, such as  $s_{r,r}^{w_2}$  (e.g. Dixon and Tukey, 1968) or, with an analogous notational interpretation,  $s_{r,r}^{T_2}$ . These have been used (with especial support for  $s_m$ ) in an auxiliary role in constructing robust location estimators such as Huber-type estimators, also in considering robust *studentized* location estimators.

Specific proposals for mixture and slippage models with normal distributions are made by Huber (1964) and Guttman and Smith (1971).

Huber (1964) re-expresses the estimation of a scale parameter  $\sigma$  for a random variable  $X$  in terms of estimation of a location parameter for  $Y = \log(X^2)$ . Thus we are estimating  $\tau = \log(\sigma^2)$ .

He shows that for the contaminated normal case (the mixture model) there is a minimax  $M$ -estimator  $\tilde{\tau}$  of  $\tau$  satisfying

$$\sum \chi(y_i - \tilde{\tau}) = 0$$

where

$$\chi(t) = \begin{cases} \frac{1}{2}(e^t - 1) & \frac{1}{2}|e^t - 1| < c \\ c \operatorname{sgn}(e^t - 1) & \frac{1}{2}|e^t - 1| \geq c \end{cases} \quad (4.3.7)$$

for an appropriate choice of  $c$ . The estimator minimizes the maximal asymptotic variance within the class of all estimators of  $\tau = \log(\sigma^2)$  which are invariant under changes of scale of the  $x_i$ . Unfortunately, the maximization process takes place only over a set of contaminating distributions concentrated on  $\{|X| > q\}$  where (if  $c \geq \frac{1}{2}$ )  $q^2 = 2c + 1$ . This restricts the range of prospects in the mixture model.

In terms of the  $x_i$  the robust estimator of  $\sigma$  is  $\tilde{\sigma} = e^{\frac{1}{2}\tau}$  and arises from solving

$$\sum_{i=1}^n \psi^2(q, x_i) = n \quad (4.3.8)$$

where

$$\psi(q, t) = \begin{cases} t & |t| < q \\ q \operatorname{sgn} t & |t| \geq q \end{cases} \quad (4.3.9)$$

See Huber (1964) for more details, including discussion of some limitations of this approach.

It is also reasonable to contemplate dispersion estimators based on samples subjected to modified trimming, modified Winsorization, or semi-Winsorization. Guttman and Smith (1971) define robust dispersion estimators  $\tilde{S}_A^2$ ,  $\tilde{S}_W^2$ , and  $\tilde{S}_S^2$  of  $\sigma^2$  analogous to their location estimators  $T_A$ ,  $T_W$ , and  $T_S$  for a normal slippage model where all but possibly one observation arise from  $N(\mu, \sigma^2)$  and at most one discordant value arises either from  $N(\mu + a, \sigma^2)$  or from  $N(\mu, b\sigma^2)$  with  $b > 1$ . The same principle applies of rejecting or modifying the observation with largest absolute residual, should this be sufficiently large. The proposed estimators take the following forms shown in Table 4.3.

In each estimator  $\kappa$  must be prescribed, and  $d$  is then chosen to ensure unbiasedness in the null case (no discordant value). Forms given in Table 4.2 apply to the more usual case where  $\mu$  is unknown. If  $\mu$  were known we would merely replace  $\bar{x}$ ,  $\bar{x}_{(1)}$  etc. in  $s^2$ ,  $s_{(1)}^2$  etc. by the true mean  $\mu$ . (Subscript indices in brackets refer to omitted ordered observations.)

Table 4.3 Forms of  $\tilde{S}_A^2$ ,  $\tilde{S}_W^2$ ,  $\tilde{S}_S^2$

$\tilde{S}_A^2$	$\tilde{S}_W^2$	$\tilde{S}_S^2$	Condition
$ds^2$	$ds^2$	$ds^2$	$z_{(1)}^2 < \kappa s^2$ and $z_{(n)}^2 < \kappa s^2$
$ds_{(1)}^2$	$d \max[s_{(2,1)}^2 s_{(n,1)}^2]$	$\frac{d}{n-1} [(n-2)s_{(1)}^2 + \kappa s^2]$	$z_{(1)}^2 \geq \kappa s^2$ and $z_{(1)}^2 > z_{(n)}^2$
$ds_{(n)}^2$	$d \max[s_{-(1,n)}^2 s_{(n-1,n)}^2]$	$\frac{d}{n-1} [(n-2)s_{(n)}^2 + \kappa s^2]$	$z_{(n)}^2 \geq \kappa s^2$ and $z_{(n)}^2 > z_{(1)}^2$

The general forms of the premium and protection measures under the two types of alternative model—mean-slippage and variance-slippage—are exhibited by Guttman and Smith, who also consider the problem of their numerical determination. Results are given *only* for the primitive case  $n = 3$ . The facts that here  $\tilde{S}_W^2$  is not worth considering when  $\mu$  is known, likewise both  $\tilde{S}_W^2$  and  $\tilde{S}_A^2$  when  $\mu$  is unknown, thus provide highly limited practical guidance on the issue of the robustness of the estimators  $\tilde{S}_A^2$ ,  $\tilde{S}_W^2$ , and  $\tilde{S}_S^2$  for reasonable sample sizes.

#### 4.4 ACCOMMODATION OF OUTLIERS IN EXPONENTIAL SAMPLES

Suppose our sample  $x_1, x_2, \dots, x_n$  comes from an exponential distribution with density

$$f(x, \sigma) = \frac{1}{\sigma} \exp(-x/\sigma) \quad (4.4.1)$$

but for the prospect that one observation may be discordant: arising from a distribution with density  $f(x, b\sigma)$  for some  $b > 1$ . We have a basic model  $H$ :  $f(x, \sigma)$  and apparently a scale-slippage alternative model  $\bar{H}$  where the slippage relates to just one observation. Since any discordant value under  $\bar{H}$  is likely to be one of the upper extremes it seems sensible to consider estimators of  $\sigma$  which minimize the influence of the higher-ordered sample values. Among *restricted L-type* estimators

$$S(\mathbf{l}) = \sum_{j=1}^m l_j x_{(j)} \quad (4.4.2)$$

(which ignore the  $n - m$  largest ordered observations) Kale and Sinha (1971) show that the one-sided Winsorized mean

$$S_{m,n} = \frac{1}{m+1} \left[ \sum_{j=1}^{m-1} x_{(j)} + (n-m+1)x_{(m)} \right] \quad (4.4.3)$$

is ‘optimal’ in the sense of minimizing  $MSE(S(\mathbf{l}) | b = 1)$ . This is demonstrated under the *exchangeable* version of the slippage model (see Section 2.3). Note that there is no suggestion of optimality under the alternative model where  $b > 1$ . In the null case (4.4.3) has efficiency  $(m+1)/(n+1)$  relative to the optimal full-sample version:  $S_{n,n}$ . It is argued, however, that this loss of efficiency under the basic model (where we do not need to protect against a discordant value) may be offset by a corresponding gain under  $\bar{H}$  (where we do need to do so). Accordingly  $MSE(S_{m,n} | b > 1)$  is investigated. The cases  $n = 3, 4$  are studied in detail and this gain (relative to  $S_{n,n}$ ) is confirmed for sufficiently large  $b$ . Typically if  $n = 4$ ,  $m = 3$  the relative efficiency rises from 0.8 at  $b = 1.1$  to 1.0 at  $b = 2$ , 4 at  $b = 5$ , 15 at  $b = 10$ , and ultimately becomes infinite.

The questions of choice of  $m$ , and performance for larger  $n$ , are taken up by Joshi (1972b). When  $b < 2$  it turns out that no  $m \neq n$  improves on  $S_{n,n}$ .

But for more extreme discordancy (larger  $b$ ) substantial gains in relative efficiency are available. Table XX on page 323 (extracted from Joshi, 1972b) presents the optimal choice  $m^*$  for  $m$  and associated relative efficiency  $e_{m^*}$  for values of  $b$  in the range 2–20. Specifically, if  $b = 1/h$  the table presents results for  $h = 0.05(0.05)0.50$ .

Of course,  $b$  will not be known and Joshi suggests an ad hoc procedure which consists of first calculating  $S_{n-1,n}$  as a provisional estimate  $\tilde{\sigma}$ , then estimating  $b$  from

$$nS_{n,n} = (n + b - 1)\tilde{\sigma} \quad (4.4.4)$$

for the purpose of determining  $m^*$  from Table XX. The corresponding  $S_{m^*,n}$  is used for  $\tilde{\sigma}$  in (4.4.4), and a new  $m^*$  is determined from the table. The process is repeated until  $m^*$  becomes stable at which stage  $\sigma$  is estimated by the corresponding  $S_{m^*,n}$ .

*Example 4.5 Failures of a critical electronic component occur from time to time in a navigational aid. On failure the component is replaced by a new one. Records show the ordered values of lifetime for 9 components to be*

$$1.6 \ 2.8 \ 2.9 \ 4.1 \ 9.8 \ 14.1 \ 16.7 \ 22.1 \ 54.3$$

Here  $n = 9$  and we have  $S_{9,9} = 12.84$  and  $S_{8,9} = 10.689$ . From (4.4.4) we get  $b = 2.811$ . Thus from Table XX,  $m^* = 8$ . We do not need to proceed further. We estimate  $\sigma$  by 10.689 for an efficiency gain (if  $b$  is truly 2.811) of about 28 per cent.

Other aspects of this approach are discussed by Sinha (1973a; moment properties and limiting form of the MSE), Sinha (1973c; refinements for the two-parameter, location shifted, case), Sinha (1973d; some exact distributional results, including lengths of confidence intervals for  $n = 4$ ,  $m = 3$ ).

Veale and Kale (1972) consider a corresponding hypothesis test. Under  $H$  the UMP size- $\alpha$  test of  $H_0: \sigma = 1$  versus  $H_1: \sigma > 1$  has critical region of the form:

$$S_{n,n} > C_{\alpha,n}$$

The robustness of this test is examined by considering its performance under the contaminated model  $\bar{H}$ . An expression for the power function  $\beta(b, \sigma)$  is obtained. Sinha proceeds to examine tests based on  $S_{m,n}$  ( $m < n$ ). For any  $m$ , a UMP size- $\alpha$  test again exists with rejection for sufficiently large  $S_{m,n}$ . However, consideration of power shows not surprisingly that for the basic (uncontaminated) model we are best to take  $m = n - 1$  if we cannot take  $m = n$ . Robustness properties of this test are discussed in terms of ‘premium’ and ‘protection’ measures (but see Section 4.1.4), and some tabulated values are presented.

Kale (1975c) presents a wider study of robust estimation of scale parameters under an exchangeable model consisting of two components in the

exponential family, with  $(n - k)$  observations from one and  $k$  (not necessarily just one) from the other. Employing a maximum likelihood approach he obtains in the case of an exponential distribution a *trimmed mean* estimator

$$S'_{k,n} = \sum_{j=1}^{n-k} x_{(j)} / (n - k) \quad (4.4.5)$$

rather than the Winsorized mean of the Kale and Sinha (1971) approach. (Trimmed means also arise for location estimates in the normal case with known, common,  $\sigma^2$ .)

The trimmed, and Winsorized, means are compared for  $k = 1$  using premium-protection measures.  $S'_{1,n}$  provides greater protection (lower MSE as  $b \rightarrow \infty$ ) than  $S_{n-1,n}$  but at a higher premium (higher MSE as  $b \searrow 1$ ). We should recall, however, that  $S_{n-1,n}$  is not necessarily the optimal form of  $S_{m,n}$ .

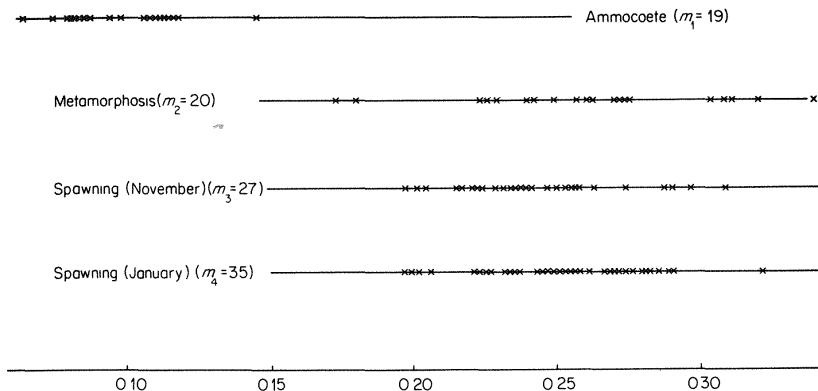
## CHAPTER 5

# *Outlying Sub-Samples: Slippage Tests*

A different type of outlier problem may arise in situations where a set of data can be divided into distinct sub-samples. The sub-samples may correspond with different levels (qualitative or quantitative) of some factor of classification, or with different combinations of some set of factors. The subdivision of the sample into the sub-samples may take place after we have collected a random sample from some overall population in which case sub-sample sizes,  $m_i$ , are random quantities. Alternatively, and more likely, we may choose random samples of prescribed sizes,  $m_i$ , at different factor levels or under different circumstances: these samples in combination serve as sub-samples in the overall data set. This is the case in many designed experiments and one interest (examinable by analysis of variance techniques on the customary assumptions of normality, additivity, and homoscedasticity) is in the comparison of the means of the populations from which the 'sub-samples' arise. The presence and effect of individual outlying observations in such sub-samples, in the context of an analysis of variance, will be discussed in Chapter 7.

Analysis of variance techniques serve to test the homogeneity of the means of the different factor-level populations against the alternative that not all means are equal: they accord to some pattern expressed by a linear model reflecting factor effects. A more general consideration can, however, be contemplated. Examination of the sub-samples (or some summary sub-sample measures, such as their means or variances) may manifest *individual* outlying sub-samples, and appropriate tests (analogous to tests of discordancy for individual outlying observations in a single sample) are of interest.

A special case might be where we believe that the sub-samples have arisen from populations with identical means, with the alternative prospect that in just one (or a few) populations the mean has 'slipped' up or down from the predominant level. Or perhaps the variance of one (or a few) populations is larger than the common variance of the majority of the



**Figure 5.1** Heart ratios of river lampreys at different stages of development

sub-populations. Tests exist for such problems. Termed *slippage tests* they are germane to a general study of outliers.

Consider the data shown diagrammatically in Figure 5.1. This shows measurements by Claridge and Potter (1974 and personal correspondence) of the heart ratios of river lampreys (*Lampetra fluviatilis*) for random samples at different stages of development: ammocoete (larva), during metamorphosis, and during the spawning run (two dates). We have a sample of 101 observations made up of four independent samples of predetermined sizes 19, 20, 27, and 35.

Whilst the results may not be surprising from the biological standpoint, they illustrate the statistical matters discussed above. The ammocoete sub-sample might well be regarded as an outlier, and a test for slippage of its mean below the means for the more adult populations could be informative. Then again, some might adjudge the metamorphosis sample an outlier in terms of its larger dispersion.

In discussing below the basis and nature of slippage tests in general we shall, *inter alia*, refer to specific methods for testing for slippage of the mean or of the variance and can later examine the lamprey data in more detail.

The slippage problem is clearly closely related to the earlier study of tests of discordancy for individual outlying observations in a single sample. We shall see that in some circumstances the statistical methods developed for individual outliers may be immediately carried over. This is trivially so if all sub-sample sizes are unity. But it can arise in other cases. For example, in testing for upward slippage of the mean in one of a set of normal distributions with common known variance it is plausible that with equal-sized samples the sufficiency of the sub-sample means will effectively reduce the problem to a test of an upper outlier in a single sample where the basic observations are the sample means. But in other situations this direct reinterpretation will not hold and new methods will result.

Any relationship with earlier single-sample results will of course only arise in the context of the slippage-type alternative hypothesis for outlier generation. Indeed, this is why the slippage-type model was so termed. Again, we must expect to encounter the usual spectrum of distributional, inferential, methodological, and dimensional distinctions. For example, slippage tests have been developed for normal, gamma (and other) populations; identification of a slipped population is of prime interest but robust estimation in the presence of slippage is also contemplated; non-parametric, Bayesian, and multiple-decision approaches have been used; slippage of multivariate populations has been considered. Paralleling the study of outliers in single samples we shall again need to distinguish techniques designed for slippage of a single population from those which contemplate (not necessarily similar) slippage of several populations; we also find once more some informality in the expression of the alternative hypothesis which makes precise classification of aim or principle rather difficult on occasions. This leads to some confusion in the literature on what is meant by 'the slippage problem' and on how it differs from analysis of variance (for designed experiments), from problems with an alternative hypothesis expressing an order relationship among the population parameters or from identification and ranking procedures for means, variances, or distributions. We shall comment further on this matter in Section 5.2.

Since the slippage problem is relatively self-contained, we shall break with the principle we have adopted in other parts of the book and draw together all the various distinctions within the present chapter rather than, for example, deferring until Chapter 8 discussion of the relevant non-parametric or Bayesian methods.

## 5.1 NON-PARAMETRIC SLIPPAGE TESTS

The earliest work on slippage, and apparently the first use of the term, is by Mosteller (1948). He considers a situation in which  $n$  equal-sized random samples, each of size  $m$ , arise from continuous distributions which are unspecified but identical, except for the possibility that one of them may have slipped in location to the right. Mosteller's non-parametric test has stimulated others. In reviewing these it is convenient to distinguish the situation where at most one population has slipped from that where several populations may have slipped.

### 5.1.1 Non-parametric tests for slippage of a single population

Mosteller (1948) approaches the problem of slippage of a single population in the following way.

On the working hypothesis,  $H$ , each of the  $n$  samples of  $m$  observations arises independently from a distribution with density function  $f(x)$ ; on the alternative hypothesis the  $i$ th sample (with  $i$  unspecified) comes from a distribution with density function  $f(x - a)$  ( $a > 0$ , unknown).

A non-parametric test of  $H$  is proposed based on rank orders of the observations in the combined sample of  $nm$  observations.

Suppose that the samples are ordered in terms of their maximum observations and denoted  $\mathcal{S}_{(1)}, \mathcal{S}_{(2)}, \dots, \mathcal{S}_{(n)}$  where  $\mathcal{S}_{(1)}$  contains the overall maximum observation,  $\mathcal{S}_{(2)}$  the second largest maximum, and so on. We shall refer to  $\mathcal{S}_{(i)}$  as the sample of rank  $i$ . Let  $M(i, j)$  represent the number of observations in  $\mathcal{S}_{(i)}$  which exceed all those in  $\mathcal{S}_{(j)}$  (and, of course, in  $\mathcal{S}_{(j+1)}, \dots, \mathcal{S}_{(n)}$ ).

Mosteller's test uses as test statistic  $M(1, 2)$ : the number of observations in  $\mathcal{S}_{(1)}$  which exceed all observations in the other samples. If  $M(1, 2)$  is sufficiently large we *reject*  $H$  and conclude that  $\mathcal{S}_{(1)}$  comes from a population which has slipped in location to the right. (A corresponding test for slippage to the left has the obvious form based on ranking the samples in terms of their minimum observations.)

The null distribution of  $M(1, 2)$  is easily determined. Let

$$\left. \begin{aligned} F(r; s, t) &= \frac{s! (t-r)!}{(s-r)! t!} = \binom{s}{r} / \binom{t}{r} & r \leq s \leq t. \\ &= 0 & r > s \leq t. \end{aligned} \right\} \quad (5.1.1)$$

Then, when  $H$  is true,

$$P\{M(1, 2) \geq h\} = F(h-1; m-1, mn-1) \quad (5.1.2)$$

$$= n \binom{mn-h}{m-h} / \binom{mn}{m} = nm^{(h)} / (mn)^{(h)} \quad (5.1.3)$$

where  $s^{(r)} = s!/(s-r)!$ ; Mosteller (1948) gives a brief table of these tail probabilities for  $n = 2(1)6$ ,  $h = 2(1)6$ ,  $m = 3(2)7, 10(5)25, \infty$ . He also presents an asymptotic form for  $P\{M(1, 2) \geq h\}$  as  $n^{-h+1}$ , but this requires  $m$  to be quite large (in excess of about 25) for reasonable accuracy.

With unequal sample sizes  $m_1, m_2, \dots, m_n$  ( $M = \sum_i^n m_i$ ) no change of principle arises. We now have

$$P\{M(1, 2) \geq h\} = \sum_1^n m_i^{(h)} / M^{(h)}. \quad (5.1.4)$$

In discussing this case, Mosteller and Tukey (1950) give an improved asymptotic form in the *equal-sized* sample case as

$$P\{M(1, 2) \geq h\} = n^{-h+1} \exp[-h(h-1)(n-1)/(2mn)] \{1 + (2h-1)/6m\} \quad (5.1.5)$$

and suggest that we allow for different-sized samples by assuming that we have  $n^*$  effective equal-sized samples of size  $M/n^*$  where

$$n^* = \left( \sum_1^n m_i \right)^2 / \left( \sum_1^n m_i^2 \right). \quad (5.1.6)$$

The only difficulties that arise with unequal sample sizes are the extra calculation effort and the unmanageable scale of any useful tabulation of tail probabilities or critical values. Accordingly this case has received little detailed consideration and we shall henceforth assume that the sample sizes are equal unless specifically stated otherwise.

Table XXI on page 324 presents critical values for 5 per cent and 1 per cent Mosteller tests (that is, the smallest values of  $h$  for which (5.1.2) is less than 0.05 and 0.01) for  $n = 2(1)6$  and  $m = 3(1)10(5)25, 100, \infty$ , compiled from published tables of tail probabilities (Mosteller, 1948; Bofinger, 1965) and additional calculations.

Since  $M(1, 2)$  is discrete we inevitably find here, and in other rank tests below, that significance levels cannot be attained exactly. Thus the size of the tests may be rather smaller than the stated significance level. For example, using the Mosteller test with  $m = 5$ ,  $n = 3$  the critical value of  $h$  for a 5 per cent test is given in Table XXI as 4. But in this case

$$P\{M(1, 2) \geq 4\} = 0.011$$

so that the size of the test is only 1.1 per cent. For  $m = 5$ ,  $n = 3$

$$P\{M(1, 2) \geq 5\} = 0.001$$

so that  $h = 5$  is the critical value for a test with significance level 1 per cent; the test size is only 0.1 per cent however. As  $m$  increases the sizes of the 5 per cent and 1 per cent tests rise only to 3.7 per cent and 0.9 per cent, respectively. Thus in operating tests of any particular level the safeguard (in terms of the probability of incorrect rejection) may be much greater than is superficially implied by the significance level. This effect is common, of course, to any non-randomized test with a discrete test statistic. In the context of non-parametric slippage tests, Neave (1972) proposes a resolution of this problem by considering two types of critical value in any situation:  $h_\alpha$  is the *best conservative* critical value at level  $\alpha$  based on a test statistic  $T$  if

$$P(T \geq h_\alpha) \leq \alpha$$

$$P(T \geq h_\alpha - 1) > \alpha$$

whereas  $h'_\alpha$  is the *nearest* critical value if

$$|P(T \geq h) - \alpha|$$

is minimized for  $h = h'_\alpha$ . Common statistical practice supports the use of the *best conservative* critical value, and this is employed throughout this text. However, care is needed in interpreting other tabulated values (particularly in Neave, 1972). Neave gives an example which highlights the problem. In the example,  $P(T \geq 2) = 0.0543$ ,  $P(T > 3) = 0.0099$  so that for a 5 per cent test  $h_{0.05} = 3$ ,  $h'_{0.05} = 2$  and the adoption of the (best conservative) critical value 3 appears to be rather wasteful. Surely the most reasonable policy is to

merely take action in relation to the observed *critical level* of the test; if we observe a test statistic value of  $t$  then we would consider the upper-tail probability  $P(T \geq t)$  evaluated under the null hypothesis, rather than be constrained by particular (arbitrary) significance levels. Thus, in practice, tables of critical values provide only a rough-and-ready guide and it is sensible to augment their information with appropriate upper-tail probabilities which are tabulated by the respective authors for most of the non-parametric tests described in this chapter and in Chapter 8.

*Example 5.1. For the Lamprey data  $n^* = 3.76$  so that the critical level for slippage to the left of the Ammocoete sample is approximately  $(3.76)^{-18} e^{-5.66}$ : overwhelming evidence of such slippage, as might be anticipated.*

Various modifications and extensions of the Mosteller-Tukey test have been published. Bofinger (1965) considers the unequal sample size case where the sample sizes  $m_1, m_2, \dots, m_n$  arise as observations from a multinomial distribution with parameters  $M = \sum_i m_i$  and  $p_j$  ( $j = 1, 2, \dots, n$ ) and we wish to test for slippage of a single population. He also presents some results on the power, against particular types of alternative hypothesis, of tests for both equal and random sample sizes, and extends the tabulation of upper-tail probabilities for tests with equal-sized samples beyond that given by Mosteller (1948).

Neave (1972) shows how the test might be modified to take account of different alternative hypothesis interests expressed by the double dichotomy:

$$\left. \begin{array}{l} A_1: \text{a specified population} \\ A_2: \text{any single population} \end{array} \right\}$$

has slipped

$$\left. \begin{array}{l} B_1: \text{in a specified direction (say, upwards)} \\ B_2: \text{in either direction.} \end{array} \right\}$$

Mosteller's (1948) test corresponds with case  $A_2B_1$ . Denoting by  $T_{A_iB_j}$  ( $i = 1, 2$ ;  $j = 1, 2$ ) the appropriate test statistics, their respective forms and explicit and asymptotic null distributions (given in terms of the function  $F(r; s, t)$ ) are as follows.

**A<sub>1</sub>B<sub>1</sub>**  $T_{A_1B_1}$  = number of observations in the *specified* sample which exceed all observations in the other samples.

$$\begin{aligned} P(T_{A_1B_1} \geq h) &= F(h; m, mn) \\ &\rightarrow n^{-h} \quad \text{as } m \rightarrow \infty \end{aligned} \tag{5.1.7}$$

**A<sub>1</sub>B<sub>2</sub>**  $T_{A_1B_2}$  = *larger* of the number of observations in the specified sample which exceed, or are less than, all observations in the other samples.

$$\begin{aligned} P(T_{A_1B_2} \geq h) &= 2F(h; m, mn) - F(2h; m, mn) \\ &\rightarrow n^{-h}(2 - n^{-h}) \quad \text{as } m \rightarrow \infty. \end{aligned} \tag{5.1.8}$$

$$A_2B_1 \quad T_{A_2B_1} = M(1, 2)$$

$$\begin{aligned} P(T_{A_2B_1} \geq h) &= F(h-1; m-1, mn-1) \\ &\rightarrow n^{-h+1} \quad \text{as } m \rightarrow \infty \end{aligned} \quad (5.1)$$

**A<sub>2</sub>B<sub>2</sub>**     $T_{A_2B_2}$  = larger of  $M(1, 2)$  or the complementary quantity:  
number of observations in the sample of rank  $n$  w  
are less than all observations in the other samples.

$$\begin{aligned} P(T_{A_2B_2} \geq h) &= F(h-1; m-1, mn-1) \\ &\times \{2 - F(h; m-h, mn-h) \\ &- F(1; (n-1)m, mn-h) \\ &\times F(h-1; m-1, mn-h-1)\} \\ &\rightarrow n^{-h+1}\{2 - n^{-h+1}\} \quad \text{as } m \rightarrow \infty. \end{aligned} \quad (5.1.10)$$

Neave (1972) presents tables of  $P(T \geq h)$  based on the asymptotic forms for all four tests with  $h = 1(1)8$ ,  $n = 3(1)10(5)20$ , which he states are ‘quite close approximations’ to the true values provided  $m \geq 20$ . He also gives the corresponding four tables of best conservative, and nearest, critical values for 5 per cent, 1 per cent, and 0.1 per cent significance levels (including any necessary corrections for finite  $m$ ).

In a later paper, Neave (1973) gives the results of a Monte Carlo investigation of the power of his four tests in comparison with an analysis of variance procedure, its Kruskal–Wallis type non-parametric counterpart and a quick test proposed by Granger and Neave (1968) (this latter test is described below). The results are based on normally distributed data with slippage in the mean of one population to the extent of 0.5, 1, 2, and 3 standard deviations, and cover the cases  $n = 3, 5, 8$ ;  $m = 10, 20, 50$ . The power of the four Neave (1972) tests is not impressive in comparison with the analysis of variance or Kruskal–Wallis tests (normality, of course, favours the former, and in both cases the alternative hypothesis is not specific to slippage of a single population). Even the Granger and Neave (1968) test is usually much better than the four Neave (1972) tests. The Granger and Neave (1968) test was offered as a quick test avoiding laborious sorting of data or special tables. It amounts to considering (for testing upward slippage) the  $k$  largest observations in the combined ordered sample of  $mn$  observations. If

$X_j$  = number of observations in the selected  $k$  which come from the  $j$ th sample,

the test statistic is

$$S = \sum_{j=1}^n X_j^2 \quad (5.1.11)$$

with critical values of the form (asymptotically)

$$\frac{k}{n} \left( \frac{mn - k}{mn} \right) C_\alpha(n) + \frac{k^2}{n} + 1$$

for a level- $\alpha$  test, where  $C_\alpha(n)$  is the upper  $\alpha$ -point of  $\chi_{n-1}^2$ .

Granger and Neave use heuristic arguments in support of this proposal. We need to choose a value of  $k$  and they recommend  $\max(12, 2n)$  (or as a 'safer' prescription,  $\max(20, 3n)$ )!

The implicit alternative hypothesis is non-specific with respect to slippage—it merely denies homogeneity of distribution for all  $n$  populations. Speculative proposals are made to take account of the direction of slippage of the means, slippage of variances and unequal sample sizes.

This test needs to be viewed as a 'quick and simple' member of the vast range of parametric and non-parametric tests of homogeneity of distribution which have non-specific alternative hypotheses. Thus in spite of the confusion in the literature on this matter it is not really a slippage test in the sense of our discussion in this chapter.

A more sophisticated non-parametric slippage test based on ranks is described by Doornbos and Prins (1958) and Karlin and Truax (1960) and is investigated by Odeh (1967). It takes the form of an  $m$ -sample version of the two-sample Wilcoxon test for equality of distributions, and has been shown to possess certain statistical optimality properties.

Given  $n$  random samples of size  $m$  the overall sample of  $mn$  observations is ranked. Suppose  $r_{jl}$  is the (overall) rank of the  $l$ th observation from the  $j$ th sample, and define

$$T_j = \sum_{l=1}^m r_{jl} \quad (j = 1, 2, \dots, n) \quad (5.1.12)$$

as the rank sum for the  $j$ th sample. To test for slippage to the right for a single population we consider

$$T_{\max} = \max_{j=1, 2, \dots, n} T_j \quad (5.1.13)$$

and conclude that the population yielding  $T_{\max}$  has slipped if  $T_{\max}$  is sufficiently large.

Specifically we reject the basic hypothesis of homogeneity of distribution at level  $\alpha$  if

$$T_{\max} > \lambda_\alpha$$

when, under the basic hypothesis,  $\lambda_\alpha$  is as small as possible subject to

$$P(T_{\max} > \lambda_\alpha) \leq \alpha.$$

(A corresponding test for slippage to the left is based appropriately on  $T_{\min} = \min_{j=1, 2, \dots, n} T_j$  with rejection if  $T_{\min}$  is sufficiently small.)

Doornbos and Prins (1958) present the  $n$  alternative hypotheses in the non-parametric forms

$$\bar{H}_i : P(X_i > X_j) \geq \frac{1}{2}$$

with  $X_j$  ( $j \neq i$ ) identically distributed (with the  $>$  sign applying to slippage to the right; the  $<$  sign applying to slippage to the left).

Karlin and Truax (1960) adopt a different form for the alternative hypotheses expressing slippage. The  $i$ th population has slipped to the right if the distribution functions  $F_j$  satisfy  $F_j = F$  ( $j \neq i$ ) and

$$F_i = (1 - \gamma)F + \gamma F^2 \quad (0 < \gamma < 1).$$

Applying results of Lehmann (1953) on rank tests they demonstrate that the above test is locally most powerful invariant (for small  $\gamma$ , and with respect to monotone transformations). They also present a different slippage test for the case where the slipped population has the different type of aberrant distribution function:

$$F_i = F^{1+\xi} \quad (\xi > 0).$$

Odeh (1967) has calculated the null distribution of  $T_{\max}$  and has presented a table of critical values  $\lambda_\alpha$  for  $m = 2(1)8$ ,  $n = 2(1)6$  and  $\alpha = 0.20, 0.10, 0.05, 0.025, 0.01, 0.005, 0.001$ . The 5 per cent and 1 per cent critical values are reproduced as Table XXII on page 325. The italicized figures are obtained from an asymptotic form and should be accurate for  $\alpha \geq 0.01$  (that is, for the values quoted in Table XXII). For  $\alpha = 0.005, 0.001$  the error is at most one unit. (It is interesting to note that asymptotically the quantities

$$V_i = \frac{T_i - m(mn+1)/2}{[nm^2(mn+1)/12]^{\frac{1}{2}}} \quad (i = 1, 2, \dots, n)$$

are jointly distributed as the set  $W_i = U_i - \bar{U}$  ( $i = 1, 2, \dots, n$ ) where the  $U_i$  are independent  $N(0, 1)$ . See Odeh, 1967.)

*Note.* Using Odeh's tables a result is significant if it is *strictly greater than* the relevant entry in the table; using the earlier Table XXI a result was significant if it was greater than or equal to the appropriate tabulated value; Table XXII re-expresses Odeh's results in this latter form.

There is a simple relationship between the critical level  $\lambda_\alpha$  and the corresponding value  $\lambda'_\alpha$  for a test for slippage to the left. We have

$$\lambda'_\alpha + \lambda_\alpha = m(mn+1) \quad (5.1.14)$$

so that Table XXII also serves for testing slippage to the left for a single population.

The principle of ranking samples in terms of the maximum observations, which was described above, appears to have been proposed independently at about the same time by Bofinger (1965) and Conover (1965). Conover (1968) uses it in proposing a 'slippage test' for equal-sized samples based on the test statistic  $M(1, n)$ . This is the number of observations in the rank 1

sample which exceed the maximum observation in the rank  $n$  sample. The null hypothesis of homogeneity of location is *rejected for sufficiently large  $M(1, n)$*  and Conover gives a short table of upper-tail probabilities and a comprehensive table of critical values for 5 per cent, 1 per cent, and 0.1 per cent tests over the ranges  $m = 4(1)10(2)20(5)40, \infty; n = 2(1)20$ . Table XXIII on page 326 reproduces the critical values for 5 per cent and 1 per cent tests for the slightly restricted range of values of  $n = 2(1)6(2)20$ . The tabulated values here are the smallest integral  $h$  (if such an  $h \leq m$  exists) for which  $P\{M(1, n) \geq h\} \leq \alpha$  for  $\alpha = 0.05$ , and 0.01. (There will of course be an obvious dual version of this test based on ranking samples in terms of their minimum observations.)

As with the test of Granger and Neave (1968), Conover's test does not employ an alternative hypothesis which is specific with regard to slippage (in spite of the author's description of the test as a 'slippage test'). The alternative hypothesis is expressed in the form: 'the distribution functions differ, at least with respect to their location parameters'. Conover reports on an empirical power comparison of the  $M(1, n)$  test with the traditional  $F$ -test for an analysis of variance, in cases where the underlying distributions have different forms: concluding that the  $M(1, n)$  test is more powerful for uniform distributions or normal distributions with *unequal* variances. But it would seem inappropriate to consider the  $F$ -test in such cases!

Slippage tests based on rank ordering of the samples, and the statistics  $M(i, j)$ , clearly place great emphasis on the extent to which extreme values in a sample reflect its location (or dispersion). Conover (1968), for example, warns against the use of his test when 'a shift in the population means does not affect the upper tail of the distribution in the same way'.

Indeed, we encounter another prospect here which does not seem to have been investigated. It could be that the sample extremes are themselves individual outliers within the particular samples, perhaps indicating contamination of a form different from any slippage of one population relative to another. The resolution of intra-sample outliers and slipped samples seems fraught with difficulty, and many non-parametric extreme rank slippage tests will be far from robust against individual outliers. Neave (1975) suggests that intra-sample outliers will tend to reduce the power of slippage tests rather than to induce wrong decisions about slippage, but this is by no means obvious and merits much wider study. With parametric models we can of course attempt to test individual outliers using the results of Chapters 3 or 7 on outliers in single samples or in designed experiments.

### **5.1.2 Non-parametric tests for slippage of several populations: multiple comparisons**

We have remarked on the non-specific form of the alternative hypothesis in many slippage tests. It is natural that such tests should be accompanied by proposals for multiple comparisons between the populations, and this is

often so. The aim is to group the populations into sub-groups within which populations are similar with respect to location, dispersion, or distribution (although the informality of structure of many published tests usually makes it impossible to attribute a particular similarity criterion).

Before considering such general grouping methods it is interesting to enquire if there are any slippage tests which directly parallel work on multiple outliers either in the sense of consecutive tests or of tests for a prescribed number of slipped populations.

Bofinger (1965) provides an example of the latter interest. To test if a *prescribed* number,  $k$ , of  $n$  populations have slipped to the right he proposes the test statistic

$$\sum_{i=1}^k M(i, k+1)$$

based on ranking  $n$  samples of  $m$  observations. If the test statistic is sufficiently large we conclude that the  $k$  populations which yield the  $k$  highest-ranking samples have slipped. (There will be an obvious dual test for slippage to the left of a prescribed set of  $k$  populations, based on sample ordering with respect to minimum observations in the samples.) An expression is given for the upper-tail probabilities in the null distribution, and selected values are tabulated ( $k = 2, 3$ ;  $n = 3(1)6$ ;  $m = 3, 5, 7, 10(5)25, \infty$ ).

Conover (1968) suggests two 'consecutive' procedures. In the first we consider  $M(1, 2)$ ,  $M(1, 3)$ ,  $M(1, 4)$  and so on until a significant result is first obtained. Suppose this happens with  $M(1, j)$ ; we conclude that populations yielding the samples of rank  $1, 2, \dots, j-1$  are indistinguishable but they have all slipped relative to the others. Alternatively we might consider  $M(1, n)$ ,  $M(1, n-1)$ ,  $M(1, n-2)$  and so on until the first insignificant result is obtained. If this arises with  $M(1, j)$ ; we conclude that the populations yielding the  $j-1$  lowest ranking samples are indistinguishable and have slipped relative to the others. Other possible consecutive schemes might be to consider

$$M(1, n), M(2, n), M(3, n) \dots$$

or

$$M(n-1, n), M(n-2, n), M(n-3, n) \dots$$

or

$$M(1, 2), M(2, 3), M(3, 4) \dots$$

In all these cases the aim is to dichotomize the set of populations into two subsets where one has slipped relative to the other.

To obtain a more refined grouping of the populations Conover (1968) suggests consecutive examinations of the members of the last of the above sets of statistics:

$$M(1, 2), M(2, 3), M(3, 4) \dots$$

In this way we look for significant differences between samples of consecutive rank. Using an approximation to the upper-tail probabilities for the

null distribution of  $M(j, j+1)$ , Conover presents algebraic expressions, tabulated values and illustrations of the technique (avoiding the restriction of equal sample sizes). It is tentatively suggested that an overall significance level  $\alpha$  can be achieved by employing a significance level of  $\alpha/(n-1)$  for each of the individual comparisons.

Another multiple comparison approach to slippage is proposed by Neave (1975). Again the alternative hypothesis is not specific: it merely states that not all location parameters are equal, but suggests that non-equality is usually manifest in a minority of the location parameters 'straying in one direction or another'. This attitude once more supports the quest for a means of dividing the populations into distinct groups with regard to their location.

A consecutive-type procedure is proposed, with no restriction on sample sizes. Firstly we examine  $M(1, 2)$  to test for upward slippage of precisely one population. We then consider

$$M(1, 3) + M(2, 3) - M(1, 2)$$

i.e. the number of observations in the samples with the *two* greatest maxima which exceed all other observations *less* the number in the highest rank sample which exceed all others. Then we examine

$$M(1, 4) + M(2, 4) + M(3, 4) - M(1, 3) - M(2, 3),$$

and so on. Assessing significance appropriately at each stage using conditional probability distributions (described by Neave, 1975, and exhibited in tables of critical values for 5 per cent and 1 per cent tests) we continue until stage  $[n/2]$ . Beyond this we would be implicitly examining slippage to the left (downwards) of the residual smaller group of lower rank samples and it is suggested that this be approached directly from the opposite end using the complementary argument for study of downward slippage (up to stage  $[(n-1)/2]$ ). The switchover is admittedly arbitrary in its effect.

The type of conclusion that might be drawn from the Neave (1975) test is illustrated by an example he discusses. Using six samples each of size 20, he finds from study of

$$M(1, 3) + M(2, 3) - M(1, 2)$$

strong evidence of slippage to the right in two populations.  $M(1, 2)$  provides less convincing evidence that one of the two has slipped to the right further than the other. At the third stage the result is insignificant, but the complementary analysis reveals significant evidence of slippage to the left for one population. Thus the six populations become partitioned in the form

\* \*\*\* \* \*

All the non-parametric slippage procedures that we have considered have the double advantage of being distribution-free and very simple to implement. Set against these advantages, however, are a variety of disadvantages. Non-parametric tests may have relatively low power. Alternative hypotheses

are often unspecified so that it is not clear what are the implications of rejecting the null hypothesis. Tests based on ranks will not necessarily reflect in any immediate sense slippage of location, dispersion, or distribution; they are also likely to be affected in an unpredictable way by individual intra-sample outliers. Multiple comparison methods have a renowned difficulty in relating the overall significance level of the battery of tests to the significance levels of individual comparisons. Finally, we might also query the extent to which the masking effect in consecutive tests of outliers has a counterpart in the Conover, or Neave, consecutive multiple comparison slippage procedures. There is clearly much work to be done to resolve the difficulties and to effect a valid comparison of the advantages and disadvantages.

## 5.2 THE SLIPPAGE MODEL

We have already encountered a degree of confusion in the literature on what is meant by slippage. Before proceeding to study parametric slippage tests it is necessary to spend some time defining what we mean in this book by the slippage problem, how it is distinguished from the wider study of outliers and where we draw the line in relation to the general statistical problem of examining homogeneity of distribution.

We suppose that the data set can be represented as  $n$  samples of sizes  $m_1, m_2, \dots, m_n$  where *at least one* of the  $m_i$  exceeds one (usually most of the sample sizes will be greater than one). Our interest is in testing if one (or a few) of the *complete* samples come from population(s) which have *slipped* relative to the others either in having different location, or dispersion, from that of the majority of the populations. This will be reflected in an alternative hypothesis which expresses such change of location or dispersion either explicitly, or implicitly through a change of distribution. The working (null) hypothesis declares homogeneity of distribution for all  $n$  populations. Thus we extend the one-sample outlier problem naturally to a problem of outlying samples in a set of samples.

If interest centres on outliers *within* particular samples we can use results in other parts of the book on individual or multiple outliers in single samples, designed experiments, or regression or time-series models, as appropriate.

The distinction is least clear if *all*  $m_i = 1$ , where results on outliers in single samples with a slippage-type alternative hypothesis would seem to be just a special case of the slippage problem defined above. This is formally true, but to distinguish cases where at least one of the  $m_i$  exceeds one has operational importance in two respects.

(i) The slippage problem with all  $m_i = 1$  is just one version of the single-sample outlier problem, where the alternative hypothesis is of a slippage type. Many other forms of alternative hypothesis are feasible as we have

seen, and it seems best to consider *en masse* the variety of single sample outlier problems to reasonably represent the range of work on outliers.

(ii) The  $n$ -sample slippage problem (without the restriction  $m_j = 1$ , all  $j$ ) figures in the literature as a topic in its own right. It is this body of work which is covered in the present chapter.

Notwithstanding (ii), emphasis in the literature is somewhat confused. For example, in a seminal paper 'on slippage', Karlin and Truax (1960) consider observations  $x_1, x_2, \dots, x_n$  from populations  $\pi_1, \pi_2, \dots, \pi_n$  and examine a means of determining whether one population has slipped. Thus it appears as if the term 'slippage' is used in relation to samples each of size 1. However, the authors explain that the  $x_i$  are typically sufficient statistics based on a *sample* from each population. The text entitled *Slippage Tests* by Doornbos (1966) is, in spite of its title, largely concerned with testing outliers in single samples: it is again only to the extent that 'observations' may be interpreted as summary statistics from samples that the term 'slippage' corresponds with our usage. Both these works, and many other important contributions, are discussed in Section 5.3 on parametric slippage tests.

Other tangential topics which we will not include further under the 'slippage' label include the following.

(a) Tests (parametric or non-parametric) of homogeneity of distribution against highly structured alternative hypotheses, such as analysis of variance, and generalizations of two-sample Wilcoxon and Mann-Whitney tests (see, for example, Friedman, 1940; Kruskal and Wallis, 1952; Downton 1976).

(b) Inference under an *ordered* alternative hypothesis. For example, if the populations have means  $\mu_1, \mu_2, \dots, \mu_n$ ;

$$H: \mu_1 = \mu_2 = \dots = \mu_n$$

$$\bar{H}: \mu_1 \geq \mu_2 \geq \dots \geq \mu_n.$$

See Barlow, Bartholomew, Bremner, and Brunk (1972) for general results in this area.

(c) Identification and ranking procedures for distributions or means. For example, the selection of a subset of  $l$  populations which contain those with the  $k$  ( $\leq l$ ) largest means. To a degree the informal classification methods of Conover, and Neave, described in Section 5.1.2 above, are in this category but we shall not further pursue such work. Bechhofer, Kiefer, and Sobel (1968) expound a sequential approach.

### 5.3 PARAMETRIC SLIPPAGE TESTS

If we know, or are prepared to make precise assumptions, about the form of the populations from which our sub-samples arise then it becomes relevant to use population-specific slippage tests.

### 5.3.1 Normal samples

The typical situation might be one in which we consider  $n$  independent normal random samples, each of size  $m$ , from distributions  $\mathbf{N}(\mu_j, \sigma^2)$  ( $j = 1, 2, \dots, n$ ) where the  $\mu_j$ , and  $\sigma^2$ , are unknown and we are interested in testing for upward slippage in the mean of at most one distribution. The working hypothesis is

$$H: \mu_j = \mu \quad (j = 1, 2, \dots, n)$$

with  $\mu$  unspecified, whilst the alternative hypothesis is

$$\bar{H}: \mu_j = \mu \quad (j \neq i)$$

$$\mu_i = \mu + a$$

where  $i$ ,  $\mu$ , and  $a$  are unspecified, with  $a > 0$ .

An appropriate test statistic is

$$T = \frac{m(\bar{x}_{\max} - \bar{x})}{[\sum_{j=1}^n \sum_{l=1}^m (x_{jl} - \bar{x})^2]^{\frac{1}{2}} \quad (5.3.1)}$$

where  $x_{jl}$  is the  $l$ th observation in the  $j$ th sample ( $j = 1, 2, \dots, n$ ;  $l = 1, 2, \dots, m$ ),  $\bar{x}$  is the overall sample mean

$$\left( \frac{1}{mn} \sum_{j,l} x_{jl} \right)$$

and  $\bar{x}_{\max}$  is the *largest* of the individual sample means,

$$\bar{x}_j = \left( \sum_{l=1}^m x_{jl} \right) / m.$$

We see that  $T$  is an inclusive statistic in the sense of Section 3.1.1 since the denominator involves *all* the data. We accept  $H$  if  $T$  is sufficiently small, otherwise we reject  $H$  and conclude that the sample yielding  $\bar{x}_{\max}$  comes from a distribution whose mean  $\mu_M$  has slipped upwards in the sense of  $\bar{H}$  (that is,  $i = M$ ).

We shall examine the credentials of the test statistic (5.3.1) later. For the moment we note its intuitive appeal as a standardized measure of aberrance of the sample with the largest mean. Specifically,  $H$  is rejected at level  $\alpha$  if

$$T > h_\alpha$$

where  $h_\alpha$  is the upper  $\alpha$ -point of the null distribution of  $T$ .

If  $m = 1$ , (5.3.1) is just the test statistic for the Pearson and Chandra Sekar (1936) single outlier test (see Section 3.4.3; Test N1) and has the optimality properties of that test against a slippage-type alternative hypothesis. We shall see that analogous properties hold for the ( $m \neq 1$ ) slippage test. Another connection with single outlier tests arises as follows. If  $x_1, x_2, \dots, x_n$  is (under a null hypothesis) a random sample from  $\mathbf{N}(\mu, \sigma^2)$ ,

and  $s_\nu^2$  is an independent external estimate of  $\sigma^2$  with  $ms_\nu^2/\sigma^2 \sim \chi_\nu^2$ , then we have considered (Section 3.4.3; Test Nv2) testing the discordancy of a single upper outlier by means of the test statistic

$$t = \frac{(x_{(n)} - \bar{x})\sqrt{(n + \nu - 1)}}{\left[\sum_{j=1}^n (x_j - \bar{x})^2 + ms_\nu^2\right]^{\frac{1}{2}}}. \quad (5.3.2)$$

But in the slippage problem described above, the *sample means* are, under  $H_0$ , independent  $\mathbf{N}(\mu, \sigma^2/m)$ , and we could test for discordancy of the upper mean outlier by using the corresponding form,

$$\frac{(\bar{x}_{\max} - \bar{x})\sqrt{(n + \nu - 1)}}{\left[\sum_{j=1}^n (\bar{x}_j - \bar{x})^2 + ms_\nu^2\right]^{\frac{1}{2}}}$$

where  $\bar{x}$  is now the *overall sample mean* and  $s_\nu^2$  is some appropriate external statistic where  $ms_\nu^2/\sigma^2 \sim \chi_\nu^2$ . Now (5.3.1) is (apart from a factor  $\sqrt{[m/(n + \nu - 1)]}$ )

of precisely this form with  $\nu = n(m - 1)$  since

$$\begin{aligned} \frac{1}{m} \sum_{j=1}^n \sum_{l=1}^m (x_{jl} - \bar{x})^2 &= \sum_{j=1}^n (\bar{x}_j - \bar{x})^2 + \frac{1}{m} \sum_{j=1}^n \sum_{l=1}^m (x_{jl} - \bar{x}_j)^2 \\ &= s_1 + s_2 \end{aligned}$$

where  $s_2$ , independent of  $s_1$ , is such that  $ms_2/\sigma^2 \sim \chi_{n(m-1)}^2$  irrespective of the values of the  $\mu_j$ .

Thus the slippage test based on the test statistic  $T$  in (5.3.1) is just a familiar single outlier test applied to the *means* of the equal-sized samples. The test can be implemented using appropriately the tabulated critical values for the single outlier test of discordancy based on (5.3.2). Specifically we need to refer  $T$  to critical values  $h_\alpha = d_\alpha \sqrt{[m/(mn - 1)]}$  where  $d_\alpha$  are the tabulated critical values in Table VIIIC on page 302. Alternatively the test can be conducted by referring  $T\sqrt{[(mn - 1)/m]}$  to the values in Table VIIIC.

*Example 5.2.* Stress trials were conducted on five processes producing castings for oil pipes. Five random samples of ten castings were chosen from the outputs of the processes, yielding the (linearly transformed) breaking strains shown in Table 5.1, expressed in appropriate units.

Sample sizes are small so that any formal test of normality is not feasible; probability plots provide no serious contra-indication. Proceeding to a Bartlett homogeneity of variance test we obtain a test statistic value of 1.48 which is not significant as  $\chi_4^2$ . Thus we shall model the data set by normal distributions of constant variance. The null hypothesis of equality of mean breaking strain could be tested against a global alternative by means of a one-way analysis of variance. This yields a mean square ratio of 5.06: highly significant as  $F_{4,45}$ .

So the means do not seem to be equal. But special interests might dictate a more specific alternative hypothesis of a slippage type. Suppose at least four processes are to be used in the mass-production of the castings. A prudent policy might be to test for downward slippage in the mean of at most one

Table 5.1

Process	1	2	3	4	5
	52	64	80	33	68
	58	47	74	58	52
	49	50	64	43	75
	45	44	84	51	56
	54	37	44	25	60
	40	51	60	40	49
	66	30	55	55	53
	67	52	47	37	41
	73	76	63	50	62
	46	56	70	15	56
means, $\bar{x}_i$	55.0	50.70	64.10	40.70	57.20
variances, $s_i^2$	116.67	169.12	175.43	186.90	93.51

population. If we reject the null hypothesis in favour of this alternative it might be best to operate just four (rather than five) processes, omitting the one which seems to yield an inferior mean breaking strain. Employing (5.3.1) suitably modified for downward slippage we calculate

$$T = \frac{m(\bar{x} - \bar{x}_{\min})}{[\sum (x_{jl} - \bar{x})^2]^{\frac{1}{2}}} = \frac{10(53.54 - 40.70)}{(9674.42)^{\frac{1}{2}}} = 1.305$$

where  $\bar{x}_{\min}$  denotes the smallest of the sample means. Thus  $T\sqrt{[(mn-1)/m]} = 2.89$ .

Table VIIIIC on page 302 with  $n=5$  and  $v=45$  shows that this result is highly significant (the upper 1 per cent point is about 2.5) and it is sensible to exclude process 4.

(In passing we might note that process 3 supports an alternative hypothesis of upward slippage of the mean of one process, to almost the same degree. But this is less relevant to the practical interest described above. However, the two significant extreme means presumably account for such a highly significance F-ratio in the analysis of variance.)

Another case of special interest is in testing for slippage of the variance of one population on the basis of  $n$  equal-sized samples of  $m$  observations from normal distributions,  $N(\mu_j, \sigma_j^2)$  ( $j = 1, 2, \dots, n$ ). Irrespective of the values of the means we can test

$$H: \sigma_j^2 = \sigma^2 \text{ (unspecified)} \quad (j = 1, 2, \dots, n)$$

against

$$\begin{aligned} \bar{H}: \sigma_j^2 &= \sigma^2 & (j \neq i) \\ \sigma_i^2 &= b\sigma^2 & (b > 1) \end{aligned}$$

using the statistic, proposed by Cochran (1941),

$$s_{\max}^2 / \sum_{j=1}^n s_j^2$$

where  $s_{\max}^2$  is the largest of the (unbiased) sample variance estimates,  $s_j^2$ .

For downward slippage in the variance, that is where under  $H: \sigma_j^2 = \sigma^2$  ( $j \neq i$ ),  $\sigma_i^2 = b\sigma^2$  ( $0 < b < 1$ ), we would use

$$s_{\min}^2 / \sum_{j=1}^n s_j^2$$

where  $s_{\min}^2$  is the smallest of the  $s_j^2$ .

The distributional properties of the test statistics  $s_{\max}^2 / \sum s_j^2$  and  $s_{\min}^2 / \sum s_j^2$  in the null cases have been quite widely studied (Cochran, 1941; Chandra Sekar and Francis, 1941; Eisenhart and Solomon, 1947; and Chambers, 1967, for the first statistic, and Doornbos, 1956, for the second statistic). Tables I and XXIV on pages 290–291 and 327 give approximate critical values for 5 per cent and 1 per cent tests in the two cases, respectively. These are reproduced from Eisenhart, Hastay, and Wallis (1947), and from Doornbos (1956).

*Note.* Table I is styled for testing discordancy in gamma samples. For present purposes it needs to be entered with  $r = (m - 1)/2$ .

*Example 5.3 (continuation of Example 5.2).* A sixth process yields a random sample of ten castings with breaking strains:

$$87, 39, 55, 30, 72, 77, 51, 44, 42, 94.$$

This gives  $\bar{x}_6 = 59.10$ ,  $s_6^2 = 481.88$ . The value of the sample variance is disquieting. For the earlier five samples we accepted the homogeneity of their population variances after an appropriate test. Perhaps we should now conduct a test for upward slippage of the variance in at most one of the six populations. We consider the test statistic

$$\frac{s_{\max}^2}{\sum_1^6 s_j^2} = \frac{481.88}{1223.51} = 0.394.$$

From Table I with  $n = 6$  and  $r = 4.5$ , we see that this is significant at the 5 per cent level and we would be advised to act cautiously in relation to process 6, in view of the manifest relative lack of consistency of the standard of castings it produces as reflected by the excess variability of their breaking strains. (Note that a test for downward slippage of variance does not show process 5 as being significantly less variable than the others.)

Having illustrated slippage tests for means and variances in normal samples we now proceed to a more systematic review of the nature and properties of such tests.

### *Slippage in the mean*

The first formal study of the problem of slippage in the mean for samples from normal distributions is that of Paulson (1952b). This work remains a cornerstone for such study.

Concerned with  $n$  independent samples, each of size  $m$ , from normal distributions  $N(\mu_j, \sigma^2)$  ( $j = 1, 2, \dots, n$ ), Paulson adopts a multiple decision approach by defining  $n+1$  possible decisions  $\mathcal{D}_0, \mathcal{D}_1, \dots, \mathcal{D}_n$ .  $\mathcal{D}_0$  declares that all the means are equal.  $\mathcal{D}_i$  declares that  $\mu_i = \mu + a$  ( $a > 0$ ),  $\mu_j = \mu$  ( $j \neq i$ ). He seeks a statistical procedure which is in some sense optimal for judging which of the  $n+1$  decisions to take, given the sample data. Subject to the restrictions:

- (i) that when  $\mathcal{D}_0$  is appropriate (that is, all means are equal) it will be selected with probability  $(1 - \alpha)$ ,
- (ii) that the decision procedure is invariant with respect to changes in the origin, or positive changes in the scale, of the measurement basis,
- (iii) the probability of adopting the *correct* decision,  $\mathcal{D}_i$ , about slippage of the mean, does not depend on  $i$ ,

a decision procedure is sought which maximizes the probability of making the correct decision when one population has slipped to the right in the sense described above.

It turns out that the optimal procedure is to use the statistic  $T$  of (5.3.1) concluding either that all the means are equal ( $\mathcal{D}_0$ ) if

$$T \leq h_\alpha$$

or that the population yielding the largest sample mean has slipped ( $\mathcal{D}_i$ ) if

$$T > h_\alpha$$

where  $i$  is the subscript of the sample yielding  $\bar{x}_{\max}$  and  $h_\alpha$  is some constant chosen to ensure that  $P(T \leq h_\alpha) = \alpha$  in the null case. This procedure is optimal irrespective of the (unknown) values of  $a$  and  $\sigma$ .

Paulson (1952b) demonstrates that the procedure is the Bayes solution for a uniform prior distribution over  $(\mathcal{D}_0, \mathcal{D}_1, \dots, \mathcal{D}_n)$ ; David (1970) modifies the proof by using the Neyman-Pearson lemma.

Operationally the determination of the critical values  $h_\alpha$ , for a procedure of size  $\alpha$ , is vitally important. The precise null distribution of  $T$  is complicated, but Paulson uses a first-order Bonferroni inequality (see, for example, Feller, 1968, page 110) to approximate  $h_\alpha$  by

$$h'_\alpha = \sqrt{\frac{m(n-1)F_\alpha}{n(mn-2+F_\alpha)}} \quad (5.3.3)$$

where  $F_\alpha$  is the upper  $2\alpha/n$  point of the  $F$ -distribution with degrees of freedom  $n_1 = 1$ ,  $n_2 = mn - 2$ . Equivalently,  $F_\alpha$  is  $(t_\alpha)^2$  where  $t_\alpha$  is the upper

$2\alpha/n$  point of the  $t$ -distribution with  $mn - 2$  degrees of freedom. Paulson shows that (5.3.3) is conservative in that  $P(T > h'_\alpha) \leq \alpha$  in the null case, and that for large  $n$  the discrepancy from the true probability is less than  $\alpha^2/2$ .

This is typical of the accuracy to which the size of slippage tests can be controlled. General results on this topic, with various applications, are presented by Doornbos (1966) and will be reviewed in Section 5.3.2 below.

There have been proposed many extensions of Paulson's result, new methods of proof and refined approximations to  $h_\alpha$ , and it is useful to summarize some of these.

Considering first of all the determination of  $h_\alpha$ , Quesenberry and David (1961) develop a computational procedure which is relevant to the calculation of approximate values of  $h_\alpha$ . Specifically they consider the distribution of  $(x_{(n)} - \bar{x})/\tilde{S}$  for a normal random sample  $x_1, \dots, x_n$  from  $N(\mu, \sigma^2)$  having sample mean  $\bar{x}$  and where

$$\tilde{S}^2 = (n-1)s^2 + \nu s_v^2$$

with  $s^2$  as the unbiased sample variance estimate  $\sum (x_i - \bar{x})^2/(n-1)$  and  $s_v^2$  an independent estimate of  $\sigma^2$  where  $\nu s_v/\sigma^2 \sim \chi_{\nu}^2$ . They use a similar Bonferroni-type approximation to that employed by Paulson (1952b). If  $c_\alpha$  is the upper  $\alpha$ -point of the distribution of  $(x_{(n)} - \bar{x})/\tilde{S}$  they show that this approach leads to an exact value of  $c_\alpha$  whenever  $c_\alpha > [(n-2)/(2n)]^{1/2}$  which holds for low values of  $n$  and  $\nu$ ; they present tables of upper 5 per cent and 1 per cent points for  $n = 3(1)10, 12, 15, 20$  and  $\nu = 0(1)10, 12, 15, 20, 24, 30, 40, 50$ . For small  $n$  and  $\nu$  these are exact to the 4 d.p. accuracy quoted, otherwise the lower and upper bounds yielded by their approach 'agree so well... that only one value had to be tabulated'. Their tables are reproduced as Tables 26a in Pearson and Hartley (1966) and, in modified form, are presented as Table VIIIC on page 302. On the relationship shown earlier in this section between the corresponding slippage and single outlier tests, we have that

$$h_\alpha = c_\alpha \sqrt{m}. \quad (5.3.4)$$

Quesenberry and David (1961) discuss an interesting example on rocket motor ignitors in which they first apply the single outlier discordancy test based on the appropriate form of (5.3.1) (that is, the Pearson and Chandra Sekar test) to each of a set of samples, and then proceed to apply the mean slippage test to the set of samples.

The optimal multiple-decision procedure of Paulson (1952b) arises also as a special case of the general Bayesian decision-theoretic approach to slippage problems described by Karlin and Truax (1960; see also Section 5.3.2). Kudo (1956b) shows that Paulson's procedure retains its optimality even if slippage in the mean of one population is accompanied by reduction in the variance of the 'slipped' population. Kapur (1957) demonstrates an unbiasedness property of the Paulson procedure when used to decide which

population mean,  $\mu_i$ , is largest (rather than to decide between the more restricted  $\mathcal{D}_i$  of the slippage model).

Certain modifications of aim or model need to be considered.

(i) *Variance known* If  $\sigma^2$  is known (unlikely as this may be) the slippage test reduces to use of  $\sqrt{m(\bar{x}_{\max} - \bar{x})}/\sigma$  as test statistic with rejection of the null hypothesis of equality of means in favour of upward slippage of  $\mu_{\max}$  if  $\sqrt{m(\bar{x}_{\max} - \bar{x})}/\sigma > \lambda_\alpha$ , where the  $\alpha$ -point can be found for  $\alpha = 0.05, 0.01$ , from Table VIIe on page 299.

(ii) *Slippage to the left* Clearly all that is needed is to replace  $\bar{x}_{\max}$  in  $T$  by  $\bar{x}_{\min}$ : the *lowest* of the sample means, with rejection at level  $\alpha$  if  $T < -h_\alpha$ .

(iii) *Slippage in an unspecified direction* Kudo (1956a) points out that Paulson's optimum procedure is easily modified for this situation yielding a decision procedure based on

$$T' = m \max_{j=1, 2, \dots, n} |\bar{x}_j - \bar{x}| / \sqrt{\sum_{j=1}^n \sum_{l=1}^m (x_{jl} - \bar{x})^2} \quad (5.3.5)$$

where if  $T' \leq h_\alpha^*$  we conclude that all means are equal; if  $T' > h_\alpha^*$  that the mean of population  $M$  has slipped, where  $M$  is the index of the sample for which  $|\bar{x}_j - \bar{x}|$  is a maximum. The critical values  $h_\alpha^*$  can again be obtained from corresponding single sample tables. Table VIIId on page 303 presents, for  $\alpha = 0.05, 0.01$  and the same range of values of  $n$  and  $\nu$  as for the one-sided test discussed above, approximate values of  $h_\alpha^* \sqrt{[(mn-1)/m]}$ . Thus  $h_\alpha^*$  is obtained as  $\sqrt{[m/(mn-1)]}$  times the entry for the appropriate number of samples,  $n$ , and for  $\nu = n(m-1)$ .

(iv) *Additional information about  $\sigma^2$*  If in addition to the set of samples we have an independent estimate of  $\sigma^2$  in the form of a quantity  $s_\eta^2$  where  $\eta s_\eta^2/\sigma^2 \sim \chi_{\nu}^2$ , then this can be profitably incorporated in  $T$  or  $T'$  by changing the denominator to

$$\left[ \sum_{j=1}^n \sum_{l=1}^m (x_{jl} - \bar{x})^2 + \eta s_\eta^2 \right]^{\frac{1}{2}}.$$

The appropriate  $h_\alpha$  or  $h_\alpha^*$  now correspond with  $\nu = n(m-1) + \eta$ .

(v) *Unequal sample sizes* Suppose the samples are of sizes  $m_1, m_2, \dots, m_n$ . This important prospect can be accommodated by modifying Paulson's procedure as follows. Put

$$y_i = m_i(\bar{x}_i - \bar{x}) / \left[ \sum_{j=1}^n \sum_{l=1}^{m_i} (x_{jl} - \bar{x})^2 \right]^{\frac{1}{2}}$$

and use as test statistic

$$T'' = \max_{j=1, 2, \dots, n} y_j \quad (\text{or} \quad \max_{j=1, 2, \dots, n} |y_j|) \quad (5.3.6)$$

in place of  $T$  (or  $T'$ ). Pfanzagl (1959) demonstrates that this procedure is locally optimum (for small  $\alpha$ ). New critical values will now be needed and do not appear to have been tabulated. Approximate values can be determined using, as before, Bonferroni-type inequalities (see Doornbos, 1966; Doornbos and Prins, 1958; also Section 5.3.2). Quesenberry and David (1961) suggest that if the  $m_i$  are not too dissimilar we can obtain an approximate test by putting

$$\bar{\bar{x}} = \frac{1}{n} \sum_{j=1}^n \sqrt{m_j} \bar{x}_j$$

and using in the numerator of the test statistic

$$\max_{j=1, 2, \dots, n} (\sqrt{m_j} \bar{x}_j - \bar{\bar{x}}) \quad \left( \text{or} \quad \max_{j=1, 2, \dots, n} |\sqrt{m_j} \bar{x}_j - \bar{\bar{x}}| \right)$$

with reference to the tabulated  $h_\alpha$  (or  $h_\alpha^*$ ) with  $\nu = \sum_{j=1}^n m_j - n$ . They do not examine the properties of this approximate test.

(vi) *Other modifications* Kudo (1956a) examines a slippage problem in a rather special situation. We have three sets of data: in the first set observations arise at random from  $m_1$  normal distributions  $N(\mu_j, \sigma^2)$ , in the second set  $m_2$  observations arise from a common normal distribution  $N(\mu', \sigma^2)$ , in the third set  $m_3$  observations arise from a common normal distribution  $N(\mu'', \sigma^2)$ , with all parameters unknown. No assumptions are made about  $\mu''$  or  $\sigma^2$ , but the null hypothesis declares that  $\mu_j = \mu'$  ( $j = 1, 2, \dots, m_1$ ) whilst as an alternative hypothesis we postulate that one of the  $\mu_j$  (in the first set) has slipped upward from  $\mu'$ . He derives a multiple decision procedure under similar restrictions and with a similar optimality property to those of Paulson's procedure. The decision rule utilizes the value of

$$\frac{x_M - \bar{x}}{S}$$

where  $x_M$  is the largest observation in the first set of data,  $\bar{x}$  is the overall mean for the first two sets of data and  $S^2$  is an overall sum of squares made up of the sum of squares about  $\bar{x}$  of the observations in the first two sets of data, plus the sum of squares about its mean of the observations in the third set. No percentage points of the null distribution have been calculated for general  $m_1$ ,  $m_2$ ,  $m_3$ . Special cases (e.g.  $m_2 = 0$ , or  $m_2 = m_3 = 0$ ) are of course easily handled in terms of single-sample outlier tests.

Karlin and Truax (1960) include in their general Bayesian study of slippage tests a particular case of normal samples where there are  $n-1$

samples of size  $m$  from  $\mathbf{N}(\mu_j, \sigma^2)$  ( $j = 1, 2, \dots, n-1$ ) and an additional control sample (index 0) of size  $m$  from  $\mathbf{N}(\mu_0, \sigma^2)$ . The basic (null) model postulates  $\mu_j = \mu_0$  ( $j = 1, 2, \dots, n-1$ ), whilst under an alternative model one of the  $\mu_j$  has slipped to some value in excess of  $\mu_0$ . They show that when  $\sigma^2$  is unknown the optimum (Bayes) procedure uses the statistic

$$\max_{j=1, 2, \dots, n-1} (\bar{x}_j - \bar{x}) / \left[ \sum_{j=0}^{n-1} \sum_{l=1}^m (x_{jl} - \bar{x})^2 \right]^{\frac{1}{2}}$$

(that is, of a form similar to (5.3.1) but where the control sample mean has been excluded from the maximization process).

### *Slippage in the variance*

Paulson's multiple-decision approach has been applied by Truax (1953) to determine an optimal procedure for (upward) slippage in the variance of one of a set of normal distributions. Again we have  $n$  samples, each of size  $m$ , from normal distributions,  $\mathbf{N}(\mu_j, \sigma_j^2)$ . If  $\mathcal{D}_0$  is the decision that

$$\sigma_j^2 = \sigma^2 \text{ (unspecified)} \quad (j = 1, 2, \dots, n)$$

and  $\mathcal{D}_i$  the decision that

$$\begin{aligned} \sigma_j^2 &= \sigma^2 \text{ (unspecified)} & j \neq i \\ \sigma_i^2 &= b\sigma^2 & (b > 1) \end{aligned}$$

then with similar invariance assumptions to those employed by Paulson, and with the same optimality principle of maximizing the probability of *correctly* adopting  $\mathcal{D}_i$  ( $i = 1, 2, \dots, n$ ) subject to  $\mathcal{D}_0$  being chosen with probability  $1-\alpha$  when no slippage has occurred, the optimal procedure turns out to be based on the statistic considered earlier by Cochran (1941),

$$U = s_{\max}^2 / \sum_{j=1}^n s_j^2. \quad (5.3.7)$$

Here  $s_{\max}^2$  is the largest of the unbiased sample variance estimates,  $s_j^2$  ( $j = 1, 2, \dots, n$ ). If  $U \leq d_\alpha$  we adopt  $\mathcal{D}_0$  and conclude that no slippage has taken place. If  $U > d_\alpha$  we adopt  $\mathcal{D}_M$ , concluding that the population yielding the largest sample variance  $s_{\max}^2$  has slipped upwards in its variance in comparison with the rest. The optimality of the procedure is demonstrated, in the proof by Truax (1953), by considering the Bayes solution relating to a prior distribution over the  $\mathcal{D}_i$  which assigns probability  $p < 1/n$  to  $\mathcal{D}_i$  ( $i = 1, 2, \dots, n$ ) and probability  $1-np$  to  $\mathcal{D}_0$ .

The determination of the percentage points,  $d_\alpha$ , has again been approached using Bonferroni-type inequalities. Cochran (1941) used this approach; some tables are published in Eisenhart, Hastay, and Wallis (1947). Table 31a in Pearson and Hartley (1966) presents the upper 5 per cent and 1 per cent points of  $U$  for  $n = 2(1)10, 12, 15, 20$  and  $m = 2(1)11, 17, 37$ ,

145,  $\infty$ , reproduced from Eisenhart, Hastay, and Wallis (1947). This table is presented as Table I on pages 290–291.

For unequal sample sizes, Pfanzagl (1959) derives a locally optimal procedure ( $b$  near 1) which uses the statistic

$$\max_{j=1, 2, \dots, n} (m_j - 1) \left[ \frac{s_j^2}{s^2} - 1 \right]$$

where  $s^2 = \sum_{j=1}^n \sum_{l=1}^{m_j} (x_{jl} - \bar{x}_j)^2 / (\sum_{j=1}^n m_j - n - 1)$ , with rejection of the hypothesis of equality of the  $\sigma_j^2$  for sufficiently large values of the statistic, but no tabulation appears to be available in easily accessible form.

For downward slippage ( $0 < b < 1$ ) the corresponding statistic is

$$U' = s_{\min}^2 / \sum_{j=1}^n s_j^2$$

where  $s_{\min}^2$  is the smallest sample variance. We conclude that slippage has occurred (in the population yielding  $s_{\min}^2$ ) if  $U'$  is sufficiently small, and approximate lower 5 per cent and 1 per cent points of  $U'$  are tabulated in Doornbos (1956). These are reproduced as Table XXIV on page 327.

Further relevant results are to be found in discussion of slippage in gamma distributions. (See Doornbos, 1966; Doornbos and Prins, 1958; and Section 5.3.3).

A quick variance slippage test based on ranges of equal-sized samples rejects the null hypothesis of equality of variances in favour of upward (downward) slippage of one variance if

$$W_{\max} / \sum W_i \quad (\text{or} \quad W_{\min} / \sum W_i)$$

is sufficiently large (small). Here  $W_{\max}$  ( $W_{\min}$ ) is the largest (smallest) of the  $n$  sample ranges  $W_i$ . Bliss, Cochran, and Tukey (1956) give approximate upper 5 per cent points for  $W_{\max} / \sum W_i$  in the null case; for  $n = 2(1)10, 12, 15, 20$  and  $m = 2(1)10$ . These are reproduced as Table 31b in Pearson and Hartley (1966) (note the use of  $k$  for  $n$ , and  $n$  for  $m$  in this table).

### 5.3.2 General slippage tests

In his book entitled *Slippage Tests*, Doornbos (1966) reviews a general method for constructing parametric slippage tests based on earlier work by himself and Prins (Doornbos and Prins, 1958).

Independent random samples of sizes  $m_1, m_2, \dots, m_n$  arise from distributions within the same family, indexed by a (vector) parameter  $\theta$ . One component,  $\theta_1$ , is of prime interest from the slippage viewpoint, the others are nuisance parameters. We want to test the working hypothesis

$$H: \theta_{11} = \theta_{12} = \dots = \theta_{1n},$$

where  $\theta_{1j}$  is the value that the parameter  $\theta_1$  assumes in the population yielding the  $j$ th sample, against one- (or two-)sided slippage alternatives for

the parameter  $\theta_1$ . Thus we contemplate the prospect, if  $H$  is untrue, that  $\theta_{1j}$  may be larger than (smaller than, different from) some common value taken by  $\theta_{1j}$  ( $j \neq i$ ). We consider a transformation of the original data where each sample is now represented by a single sufficient statistic

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

in such a way that the  $X_j$  are identically distributed with a *known* distribution not depending on the nuisance parameters and if  $H$  is true independent of  $\theta_1$  also.

To test for slippage to the right of a single population we consider

$$d_j = P(X_j > x_j) \quad (j = 1, 2, \dots, n) \quad (5.3.8)$$

and reject  $H$  if

$$D = \min_j d_j \leq \alpha/n. \quad (5.3.9)$$

If (5.3.9) holds we conclude that slippage has occurred in the population yielding the minimum value  $D$ .

For slippage to the left, or slippage in either direction, we consider also

$$e_j = P(X_j \leq x_j) \quad (j = 1, 2, \dots, n) \quad (5.3.10)$$

and conclude that slippage has taken place if

$$E = \min_j e_j \leq \alpha/n \quad (5.3.11)$$

or

$$\min(D, E) \leq \alpha/2n, \quad (5.3.12)$$

respectively.

To see that these proposals lead to tests of level  $\alpha$  we need to study the probabilities of type I errors. Consider, for an arbitrary set of  $n$  real numbers  $u_1, u_2, \dots, u_n$

$$\begin{aligned} p_j &= P(X_j \leq u_j) & (q_j &= P(X_j > u_j)) \\ p_{jl} &= P(X_j \leq u_j, X_l \leq u_l) & (j \neq l) \\ q_{jl} &= P(X_j > u_j, X_l > u_l) & (j \neq l) \end{aligned}$$

determined under  $H$ . Suppose  $P$  is the probability that at least one of the  $X_j$  does not exceed the corresponding  $u_j$ . If  $A_j$  is the event:  $X_j \leq u_j$ , then

$$P = P\left(\bigcup_1^n A_j\right)$$

and by the first Bonferroni inequality (Feller, 1968, page 110)

$$\sum_{j=1}^n p_j - \sum_{j < l} p_{jl} \leq P \leq \sum_{j=1}^n p_j. \quad (5.3.13)$$

Equivalently,

$$\sum_{j=1}^n q_j - \sum_{j < l} q_{jl} \leq Q \leq \sum_{j=1}^n q_j \quad (5.3.14)$$

where  $Q$  is the probability that the least one of the  $X_i$  exceeds the corresponding  $u_j$ .

Now if it happened that

$$p_{jl} \leq p_j p_l, \quad (5.3.15)$$

or equivalently, that

$$q_{jl} \leq q_j q_l, \quad (5.3.16)$$

we have immediately, from (5.3.13) and (5.3.14), that

$$\sum_{j=1}^n p_j - \sum_{j < l} p_j p_l \leq P \leq \sum_{j=1}^n p_j, \quad (5.3.17)$$

$$\sum_{j=1}^n q_j - \sum_{j < l} q_j q_l \leq Q \leq \sum_{j=1}^n q_j. \quad (5.3.18)$$

Aggregating the  $p_j$ , and  $q_j$ , and putting  $p = \sum_{j=1}^n p_j$ ,  $q = \sum_{j=1}^n q_j$  leads immediately to

$$p - \frac{1}{2}p^2 \leq P \leq p \quad (5.3.19)$$

$$q - \frac{1}{2}q^2 \leq Q \leq q \quad (5.3.20)$$

which enable bounds to be placed on the probabilities of type I error for the general slippage tests. For if (in the case of continuous  $X_i$ ) we choose values  $u_{j\alpha}$  for the  $u_j$  where

$$q_{j\alpha} = P(X_j > u_{j\alpha}) = \alpha/n \quad (5.3.21)$$

then on the basis of the test described in terms of (5.3.8) and (5.3.9) we reject  $H$  in favour of the alternative hypothesis of slippage to the right if at least one of the  $x_i$  exceeds  $u_{j\alpha}$  (for a common distribution of the  $X_i$ , as postulated in the structure above, all the  $u_{j\alpha}$  will of course be equal). Thus  $Q_\alpha$ , the probability that at least one  $X_i$  exceeds the corresponding  $u_{j\alpha}$ , is the probability of type I error, or significance level, of the test of slippage to the right. So from (5.3.20)

$$\alpha - \alpha^2/2 \leq Q_\alpha \leq \alpha \quad (5.3.22)$$

and the test has level  $\alpha$  and size which is not less than  $\alpha - \alpha^2/2$ .

A similar result can be readily demonstrated for the test of slippage to the left.

Certain important points should be noted.

- (i) If the  $X_i$  are discrete, precise  $\alpha$ -points may not exist but we can take  $u_{j\alpha}$  in (5.3.21) as the smallest integer for which  $q_{j\alpha} \leq \alpha/n$ . Then (5.3.22) holds with  $\alpha$  replaced by  $\alpha' = \sum_{j=1}^n q_{j\alpha}$ .

(ii) We remarked on (5.3.22) in relation to the Paulson (1952b) test for slippage of a normal mean with *large* equal-sized samples. In fact, it holds under the above conditions even if the sample sizes differ and are not necessarily large.

(iii) (5.3.22) depends crucially (in the above proof) on the inequality (5.3.15), or (5.3.16), holding. Doornbos (1966) shows this to be true for the case of normal means (see also Quesenberry and David, 1961) and demonstrates its propriety for other parametric slippage tests (including normal variance, tests and some others we discuss in Section 5.3.3 below). When (5.3.15) does not hold, the upper bound in (5.3.22) still prevails so that the test has level  $\alpha$ . The lower bound does not, so that we are uncertain of how close is the size of the test to the level,  $\alpha$ . Doornbos (1966) discusses conditions for the validity of (5.3.15) in the continuous case.

Following a similar multiple decision approach to that advanced by Paulson (1952b) for normal means, and employed by Truax (1953) for normal variances, Doornbos (1966) shows that certain discrete slippage tests have a similar optimality. These include tests for slippage of means of Poisson distributions, and of proportions in binomial or negative binomial distributions. See below.

Karlin and Truax (1960) also present a rather general approach to slippage tests, with particular illustrative cases discussed in detail. Their approach is entirely Bayesian decision-theoretic, exhibiting optimum procedures in the form of Bayes solutions for loss structures which do not need to be specified in great detail. Some of their specific proposals are discussed elsewhere in this chapter. At this stage we consider briefly the methodological basis of their approach. They consider a one-parameter problem in which decision rules are required to be invariant with respect to permutation of the labels of the samples. The loss functions  $h_i(\theta)$ , relating to the taking of decision  $\mathcal{D}_i$  when the values of the parameter  $\theta$  for all populations are specified, are assumed to have certain desirable properties including permutation invariance and reduced values when slippage occurs for population  $i$ . The special form of the Bayes solution when sufficient statistics exist is given particular attention. It amounts to accepting  $\mathcal{D}_0$  (no slippage) unless the maximum discrepancy between the sample point  $x_i$  and the MLE  $\hat{\theta}$  of the parameter  $\theta$  (under the assumption of no slippage) is particularly large. The use of a *sample point*  $x_i$  highlights the transference from an initial statement of the problem in terms of *samples* from each of a set of populations to detailed study involving a single observed value from each population. In this respect (as remarked in Section 5.2) the Karlin and Truax (1960) Bayesian analysis is more in keeping with single outlier study than with slippage based on multi-observation samples—in spite of the authors' description of their work.

Karlin and Truax show that their symmetric invariant Bayes solutions are

(largely) uniformly most powerful among symmetric invariant procedures with a prescribed probability of incorrectly adopting  $\mathcal{D}_0$ . In this part of their work a simple  $(0, 1)$  loss structure is employed.

### 5.3.3 Non-normal samples

We have implicitly considered non-normal samples in the test of slippage of variance based on normal samples and described in Section 5.3.1. For if  $s_j^2$  ( $j = 1, 2, \dots, n$ ) are the sample variances based on samples of size  $m$  they are, in the null case, independent observations from  $\sigma^2 \chi_{m-1}^2 / (m-1)$ . Thus in testing for slippage of a population variance we are testing for discordancy of an outlier in a  $\chi^2$  (gamma) sample, against a slippage-type alternative hypothesis. It is but a minor modification to proceed to test for slippage of the scale parameter in a gamma distribution on the basis of a set of samples from gamma populations.

#### *Gamma samples*

Here, and throughout most of this section, we use or extend some of the results of Doornbos (1966).

Suppose we have  $n$  samples, of sizes  $m_j$  ( $j = 1, 2, \dots, n$ ), from gamma distributions  $\Gamma(r_j, \lambda_j)$  ( $j = 1, 2, \dots, n$ ). Under the null hypothesis all the  $r_j$  are equal, also the  $\lambda_j$ , with common (unknown) values  $r$  and  $\lambda$  respectively. Under the alternative hypothesis of upward (downward) slippage of the scale parameter we have

$$\lambda_i = c\lambda \quad (i \text{ unspecified})$$

$$\lambda_j = \lambda \quad (j \neq i)$$

with  $c > 1$  ( $c < 1$ ). In view of the additivity property of the gamma distribution, the sample sums  $t_1, t_2, \dots, t_n$  are independent  $\Gamma(m_j r_j, \lambda_j)$  and we can test for upward (downward) slippage by using a test statistic

$$\max t_j \left/ \sum_{l=1}^n t_l \right. \\ \left( \min t_j \left/ \sum_{l=1}^n t_l \right. \right).$$

The null distribution of this test statistic is complicated. But following the general approach to slippage tests described in Section 5.3.2 we conclude (equivalently) that slippage has occurred, for a test with significance level  $\alpha$ , if

$$\min_j P \left\{ t_j \left/ \sum_1^n t_l \right. > t_{j0} \left/ \sum_1^n t_{l0} \right. \right\} \leq \alpha/n$$

$$\left( \min_j P \left\{ t_j \left/ \sum_1^n t_l \right. \leq t_{j0} \left/ \sum_1^n t_{l0} \right. \right\} \leq \alpha/n \right)$$

where  $t_{j0}$  is the observed value of the  $j$ th sample sum.

This test has level  $\alpha$  on the usual Bonferroni-type inequality argument and, again, size within  $\alpha^2/2$  of  $\alpha$  (for details see Doornbos, 1966; Doornbos and Prins, 1956). The probabilities  $p(x) = P(t_i/\sum_1^n t_i > x)$  (or  $q(x) = 1 - p(x)$ ) can of course be expressed in terms of incomplete Beta functions and evaluated from published tables or nomograms (Pearson, 1968; Pearson and Hartley, 1966, Table 17).

Doornbos (1966) also considers slippage tests for some discrete distributions.

### *Poisson samples*

We have  $n$  samples of sizes  $m_j$  ( $j = 1, 2, \dots, n$ ) from Poisson distributions  $\mathbf{P}(\mu_j)$  with means  $\mu_j$  ( $j = 1, 2, \dots, n$ ). The sample totals  $t_i$  have Poisson distributions  $\mathbf{P}(m_j \mu_j)$  and the null hypothesis,  $H$ , of equality of the  $\mu_j$  (at a value  $\mu$ ) is equivalent to declaring specific values for the ratios of the means of the  $t_i$ . Under  $H$  the  $t_i$  constitute independent single observations from Poisson distributions  $\mathbf{P}(\mu_j^*)$  where  $\mu_j^* = m_j \mu$  so that

$$\mu_j^* / \sum_1^n \mu_i^* = m_j/M = \pi_j \quad (5.3.23)$$

(say) where  $M = \sum_1^n m_j$ . Note that this structure arises also from observing numbers of events in different time intervals in independent Poisson processes where as a working hypothesis we declare that the processes all have the same rate.

The alternative hypothesis of upward (downward) slippage of one of the means  $\mu_i$  becomes transformed to

$$\frac{\mu_i^*}{\sum_1^n \mu_i^*} = c\pi_i; \quad \frac{\mu_j^*}{\sum_1^n \mu_i^*} = \frac{1-c\pi_i}{1-\pi_i} \pi_j \quad (j \neq i) \quad (5.3.24)$$

with  $1/\pi_i > c > 1$  ( $0 < c < 1$ ).

Since the joint distribution of  $t_1, t_2, \dots, t_n$ , conditional on the sum  $t = \sum_1^n t_i$ , is multinomial a simple conditional slippage test is easily constructed. Given  $t$ , each  $t_i$  has a binomial distribution, so under  $H$  we can readily express

$$\begin{aligned} d_i &= P\left\{t_i \geq t_{i0} \mid \sum_1^n t_i = t\right\} = \sum_{s=t_{i0}}^t \binom{t}{s} \pi_j^s (1-\pi_j)^{t-s} \\ &= I_{\pi_j}[t_{i0}, t - t_{i0} + 1] \end{aligned} \quad (5.3.25)$$

where  $I_p(\alpha, \beta)$  is the incomplete Beta function satisfying

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} I_p(\alpha, \beta) = \int_0^p u^{\alpha-1} (1-u)^{\beta-1} du.$$

So on the now familiar argument we again reject  $H$  at level  $\alpha$  in favour of upward slippage of one of the Poisson means if

$$\min_i d_i \leq \alpha/n$$

concluding that the population yielding the minimum probability  $d_i$  is the one that has slipped. (For downward slippage we employ the corresponding argument in terms of

$$e_j = P\left[ t_j \leq t_{j0} \mid \sum_1^n t_i = t \right].$$

Note that the probability calculations involve the well-tabulated incomplete Beta function (or tail probabilities for the binomial distribution). The test can obviously be applied in an unconditional form for testing slippage in *multinomial* populations.

When the sample sizes  $m_i$  are equal the  $\pi_i$  will also have equal values and the minimum value of  $d_i$  occurs at the maximum value of  $t_{j0}$ , and we will reject  $H$  if this value is sufficiently large. Table XXV (from Doornbos, 1966) gives corresponding (approximate) critical values for 5 per cent and 1 per cent upward slippage tests for  $n = 2(1)10$  and  $t = 2(1)25$ , together with test sizes (which in view of the discreteness of the data may differ noticeably from the significance levels). The upper limit of 25 on the value of  $t$  is of course a serious practical restriction on the use of Table XXV (page 328).

### *Binomial samples*

Sets of  $m_i$  independent Bernoulli observations arise from  $n$  populations where the probabilities of 'success' are  $p_i$  ( $i = 1, 2, \dots, n$ ). The respective numbers of successes are  $r_1, r_2, \dots, r_n$ . The null hypothesis  $H$  declares that all  $p_i$  are equal. Under the alternative hypothesis of upward (downward) slippage of one of the  $p_i$  we have

$$p_i = cp \quad (i \text{ unspecified})$$

$$p_j = p \quad (j \neq i)$$

with  $1 < c \leq 1/p$  ( $0 < c < 1$ ). The slippage test is again of a conditional form based now on the *hypergeometric* distribution.

Under  $H$  we have

$$P\left\{ r_j = r_{j0} \mid \sum_1^n r_i = r \right\} = \binom{m_j}{r_{j0}} \binom{M - m_j}{r - r_{j0}} / \binom{M}{r} \quad (5.3.26)$$

where  $M = \sum_1^n m_i$ , and we can determine the tail probabilities

$$d_j = P\left[r_j \geq r_{j0} \mid \sum r_i = r\right]$$

$$\left(e_j = P\left[r_j \leq r_{j0} \mid \sum r_i = r\right]\right)$$

and hence conclude, at level  $\alpha$ , that upward (downward) slippage has occurred in the population yielding  $\min_j d_j$  ( $\min_j e_j$ ) if

$$\min_j d_j \leq \alpha/n \quad \left(\min_j e_j \leq \alpha/n\right).$$

Tables of hypergeometric probabilities for the determination of critical values can be found in Owen (1962) up to  $M = 20$  or in more extensive form in Lieberman and Owen (1961).

Again if the sample sizes  $m_i$  are equal, detection of the slipped population, and assessment of its discordancy, are based simply on  $\max r_{j0}$  ( $\min r_{j0}$ ). See Tests B1 and B2 in Chapter 3 and associated tables.

Doornbos (1966) also considers a slippage test for *negative binomial* samples.

### 5.3.4 Group parametric slippage tests

We may wish to test if a group of  $k (> 1)$  out of  $n$  populations have slipped, based on samples of  $m_j$  observations from the  $j$ th population ( $j = 1, 2, \dots, n$ ). Doornbos (1966) extends the Bonferroni-type inequality approach to this more general situation. We lose the facility of setting a lower bound to the size of the slippage test but can obtain a straightforward (if on occasions somewhat time-consuming) slippage test of a given significance level.

The approach is applied to derive specific tests for sets of normal, gamma, Poisson, binomial (and other) samples. We shall illustrate it only for the normal case with upward slippage in the mean.

We have  $n$  normal samples of sizes  $m_1, m_2, \dots, m_n$  from distributions with means  $\mu_1, \mu_2, \dots, \mu_n$  and common unknown variance  $\sigma^2$ . Under  $H: \mu_j = \mu$  (unspecified) whilst under  $\bar{H}$  we have

$$\begin{aligned}\mu_i &= \mu + a & (i \in I, \text{ unspecified}) \\ \mu_j &= \mu & (j \notin I)\end{aligned}$$

where  $I$  is some subset of subscripts:  $I = (i_1, i_2, \dots, i_k)$ .

We combine the data into two sets, one consisting of  $k$  samples with total sample size  $M_I$ , the other of  $n - k$  samples with total sample size  $M - M_I$ ,

where  $M = \sum m_i$ , corresponding with a particular choice of  $I$ . For any  $I$  we determine the critical level for the usual one-sided two-sample  $t$ -test for equality of two normal means against the alternative that the *first* sample comes from a distribution with larger mean. Let the critical level be  $d_I$ .

We repeat this calculation for all  $\binom{n}{k}$  choices of the set  $I$  and reject  $H$  in favour of  $\bar{H}$  if

$$\min_I d_I \leq \alpha / \binom{n}{k}.$$

Such a test is easily seen to have significance level  $\alpha$ : the probability of incorrectly rejecting  $H$  is less than or equal to  $\alpha$ .

Note however, that with  $n = 20$ ,  $k = 3$  we would have to determine 1140 values  $d_I$ !

A more specific group test for normal samples of equal size  $m$ , with  $k = 2$  but where the alternative hypothesis specifies slippage of one mean upwards, and the other downwards, is described by Ramachandran and Khatri (1957). It amounts to considering a test statistic

$$\frac{m(\bar{x}_{\max} - \bar{x}_{\min})}{[\sum_{j=1}^n \sum_{l=1}^m (x_{jl} - \bar{x})^2]^{\frac{1}{2}}}$$

and concluding that slippage has occurred if this statistic is sufficiently large; slippage then being attributed to the two populations yielding  $\bar{x}_{\min}$  and  $\bar{x}_{\max}$ . The test has the Paulson-type multiple decision optimality, but critical values do not seem to have been published.

#### 5.4 OTHER SLIPPAGE WORK

In the earlier sections of this chapter we have reviewed and categorized the published work on slippage tests, viewed as a natural generalization of tests of discordancy for outliers. Not all outlier considerations have been paralleled in the slippage ('outlying sub-sample') context. The absence of detailed comment on multivariate slippage, accommodation of slipped populations in the framework of testing or estimating parameters, effects of slippage in studies of structured models such as regression or designed experiments, all reflect the apparent lack of much published work in these areas. (Although in a general sense the vast heritage of linear model study utilizing analysis of variance techniques can be viewed, in the case of replicated observations, as concerned with slippage of the means of a set of populations.)

Some isolated topics under the above headings merit brief mention. Karlin and Truax (1960) include comment on multivariate slippage in their Bayesian decision-theoretic examination of slippage problems, but they are concerned more particularly with samples of size 1 (outliers rather than slipped samples).

Naik (1972) has considered a Bayesian analysis of 'contaminated samples' in which each of  $n$  samples arises either from a population  $\mathcal{P}$  or from a population  $\mathcal{P}^*$ . The populations  $\mathcal{P}$  and  $\mathcal{P}^*$  are from a common family of populations, distinguished only by different values  $\theta$  and  $\theta^*$  of some important parameter. The model is typified by a situation in which certain samples correspond with 'good runs' of some system for which the parameter value is  $\theta$ , the others with 'bad runs' for which the parameter value is  $\theta^*$ . An event  $a^{(r)}$ , with probability  $p^{(r)}$ , is defined under which a particular  $r$  of the  $n$  samples arise from good runs, and starting with a non-informative prior distribution for  $(a^{(r)}, \theta, \theta^*)$  its posterior distribution is determined.

In the special cases considered in detail, the populations are either normal or exponential, and squared error loss functions for  $\theta$  are assumed in the process of estimating the (uncontaminated) value  $\theta$  *a posteriori*. For the normal case three situations are discussed.  $\mathcal{P}$  is  $N(\mu, \sigma^2)$  and  $\mathcal{P}^*$  is  $N(\mu^*, \sigma^{*2})$ , and it is assumed either that

$$\mu^* = \mu + a, \quad \sigma^* = \sigma$$

where  $a$  is either known or unknown, or else that  $\mu^*$  and  $\sigma^*$  are arbitrarily different from  $\mu$  and  $\sigma$ . For  $\mu^* = \mu + a$ ,  $\sigma^* = \sigma$ , with  $a$  specified a numerical example is presented for  $n = 3$  and equal sample sizes  $m_j = 10$  ( $j = 1, 2, 3$ ). Parallel situations for exponential samples are also considered.

This work by Naik (1972) is an example of the accommodation of slipped samples in the estimation of crucial parameters in the unslipped population.

A sequential approach to testing for upward slippage in the mean of one of  $n$  normal populations is proposed by Srivastava (1973). The normal distributions have common unknown variance and the aim is to decide, on the basis of observations taken one by one from each population, which decision  $\mathcal{D}_i$  to adopt where under  $\mathcal{D}_i$  ( $i = 1, 2, \dots, n$ )

$$\begin{aligned}\mu_i &= \mu + a \\ \mu_j &= \mu \quad (j \neq i)\end{aligned}$$

with  $\mu$  known and  $a > 0$  specified. Under  $\mathcal{D}_0$  no slippage has occurred:  $\mu_j = \mu$  ( $j = 1, 2, \dots, n$ ).

An asymptotically efficient sequential procedure is presented under which the probability of correctly adopting  $\mathcal{D}_0$  exceeds  $1 - \alpha$ , and the probability of correctly adopting  $\mathcal{D}_i$  ( $i \neq 0$ ) exceeds  $1 - \beta$ , for prescribed  $\alpha$  and  $\beta$ .

A numerical study is presented, for small  $n = 2(2)6$ , of the error probabilities and expected sample numbers for the case  $\alpha = \beta = 0.05$ . Both depend strongly on the value of  $\lambda = \sigma/a$ , and  $\lambda$  needs to be quite large for the error probabilities to become close to  $\alpha$  and  $\beta$  (that is, for the 'size' of the procedure to approach its 'level'). When  $n = 2, 4, 6$  we need  $\lambda$  to exceed about  $n/2$  to obtain error probabilities in excess of about 0.046.

This sequential procedure, whatever its potential efficiency, has crucial practical limitations in respect of the need to know  $\mu$  and  $a$  and the computational effort needed to carry it out.

A sequential procedure utilizing an independent estimate of  $\sigma^2$  is described by Paulson (1962).

An 'optimal' distribution-free slippage test based on normal scores rather than ranks is described by Paulson (1961).

## CHAPTER 6

# *Outliers in Multivariate Data*

Tests of discordancy of outliers have as much relevance and importance for multivariate data as they do for univariate samples. Many factors carry over immediately. The two basic notions, of an outlier as an observation which engenders surprise owing to its extremeness, and of its discordancy in the sense of that ‘extremeness’ being statistically unreasonable in terms of some basic model, are not constrained by the dimensionality of the data. But their expression is by no means as straightforward in more than one dimension.

Any formal, or indeed even subjective, idea of extremeness is obtuse. As Gnanadesikan and Kettenring (1972) remark, the multivariate outlier no longer has a simple manifestation as the observation which ‘sticks out at the end’ of the sample. But, notably in bivariate data, we may still perceive an observation as suspiciously aberrant from the data mass, particularly so if the data is represented in the form of a scatter diagram.

The idea of extremeness inevitably arises from some form of ‘ordering’ of the data. Unfortunately, no unique unambiguous form of total ordering is possible for multivariate data, although different types of sub- (less than total) ordering principle may be defined and employed. Barnett (1976) surveys the role of sub-ordering in multivariate analysis. Thus in attempting to express mathematically the subjective stimulus to the declaration of an outlier in a multivariate sample we will have to employ some *sub-ordering* principle.

Suppose we were to represent a multivariate observation  $\mathbf{x}$  by means of some univariate metric, or distance measure,

$$R(\mathbf{x}; \mathbf{x}_0, \Gamma) = (\mathbf{x} - \mathbf{x}_0)' \Gamma^{-1} (\mathbf{x} - \mathbf{x}_0)$$

where  $\mathbf{x}_0$  reflects the location of the data set or underlying distribution and  $\Gamma^{-1}$  applies a differential weighting to the components of the multivariate observation related to their scatter or to the population variability. For example,  $\mathbf{x}_0$  might be the zero vector  $\mathbf{0}$ , or the true mean  $\boldsymbol{\mu}$ , or the sample mean  $\bar{\mathbf{x}}$ , and  $\Gamma$  might be the variance-covariance matrix  $V$  or its sample equivalent  $S$ , depending on the state of our knowledge about  $\boldsymbol{\mu}$  and  $V$ . One

means of ordering the sample  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ , termed *reduced ordering*, is achieved by ordering the values  $R_i(\mathbf{x}_0, \Gamma) \equiv R(\mathbf{x}_i; \mathbf{x}_0, \Gamma)$  and this may be employed as a framework on which to express the extremeness of certain members of the sample. Needless to say, we will (except in special circumstances) be sacrificing certain information on the multivariate structure by employing such a reduction of the data. This would be true also if we order in terms of, say, the first principal component or restrict attention to *marginal ordering* based on some favoured single component of  $\mathbf{x}$ .

Ideally we would hope that an appropriate sub-ordering principle would implicitly emerge in outlier analysis from a specification of a basic model, an alternative hypothesis, and a statistical test principle. But as with much of the work on univariate outliers, this ideal is not often realized. Again tests for discordancy have emerged from 'intuitively reasonable' test statistics and criteria which may indeed be more 'reasonable' in one context than in some other. We must acknowledge the degree of arbitrariness which arises from having to employ sub-ordering, rather than the unattainable total ordering, as the framework for expressing extremeness. But this need not always be too serious. For example, when the basic model is multivariate normal we find that reduced ordering of the distances

$$R(\mathbf{x}; \boldsymbol{\mu}, V) = (\mathbf{x} - \boldsymbol{\mu})' V^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

has substantial appeal in terms of probability ellipsoids (an appeal less evident for non-normal data) and also arises naturally from a likelihood ratio approach to outlier discordancy tests.

Once we have set up some pragmatic test statistic for a test of discordancy of a multivariate outlier no *fundamental* obstacle remains to its application. The only problems are computational and manipulative: in determining and tabulating appropriate percentage points in the null distribution of the test statistic for assessment of significance, or in the non-null distribution as an expression of the power of the test. But these problems can be severe. Manipulation of multivariate distributions is notoriously complicated and it is little wonder that few detailed calculations have been published.

In this chapter we shall be examining, and illustrating, the rather sparse amount of material extant for testing discordancy of multivariate outliers. Many proposals remain in qualitative rather than quantitative form but it is informative to consider some of the recent qualitative proposals for outlier assessment which rest on various forms of graphical representation of the data. The dual problem of *accommodation* of outliers in robust multivariate inference procedures is correspondingly ill-represented in the literature. A few comments on this topic appear later, in Section 6.3.

## 6.1 OUTLIERS IN MULTIVARIATE NORMAL SAMPLES

As may be anticipated, most of the sparse study of outliers in multivariate data deals either with the case of an underlying normal distribution, or is

non-specific in relation to the form of the basic model. We maintain this distinction by considering first the results available for normal samples, and in the subsequent section examine the less formal proposals that have been made for data from a probabilistically unspecified source.

Suppose  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  is a sample of  $n$  observations of a  $p$ -component normal random variable  $\mathbf{X}$ . We initially assume these to have arisen at random from a  $p$ -dimensional normal distribution,  $\mathbf{N}(\boldsymbol{\mu}, V)$ , where  $\boldsymbol{\mu}$  is the  $p$ -vector of means, and  $V$  the  $p \times p$  variance-covariance matrix. A possible alternative model which would account for a single outlier is the slippage alternative, obtained as a multivariate adaptation of the univariate *models A* (slippage of the mean) and *B* (slippage of the variance) discussed by Ferguson (1961a)—see Section 2.3. Specifically, the alternative hypotheses are:

$$\begin{aligned} \text{model A} \quad E(\mathbf{X}_i) &= \boldsymbol{\mu} + \mathbf{a} \quad (\text{some } i) \\ E(\mathbf{X}_j) &= \boldsymbol{\mu} \quad (j \neq i) \\ \text{with variance-covariance matrix } V(\mathbf{X}_j) &= V \quad (j = 1, 2, \dots, n). \end{aligned}$$

$$\begin{aligned} \text{model B} \quad V(\mathbf{X}_i) &= bV \quad (\text{some } i) \quad (b > 1) \\ V(\mathbf{X}_j) &= V \quad (j \neq i) \\ \text{with mean vector } E(\mathbf{X}_j) &= \boldsymbol{\mu} \quad (j = 1, 2, \dots, n). \end{aligned}$$

We shall consider later (in Chapter 8) the Bayesian analysis of the *model A* presented by Guttman (1973b); reference has already been made to the corresponding work of Karlin and Truax (Section 5.4).

For the present we adopt a more traditional viewpoint, in considering a test of discordancy based on the *two-stage maximum likelihood ratio principle* as explained in Section 3.1. We consider *models A* and *B* separately, with and without the assumption that parameter values are known.

### *Model A, V known*

On the basic model the likelihood of the sample  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  is proportional to

$$P_{\boldsymbol{\mu}}(\mathbf{x} \mid V) = \frac{1}{|V|^{n/2}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^n (\mathbf{x}_j - \boldsymbol{\mu})' V^{-1} (\mathbf{x}_j - \boldsymbol{\mu}) \right\}. \quad (6.1.1)$$

The maximized log-likelihood is (apart from the constant factor)

$$L(\mathbf{x} \mid V) = -\frac{1}{2} \sum_{j=1}^n (\mathbf{x}_j - \bar{\mathbf{x}})' V^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}). \quad (6.1.2)$$

Under the alternative (*model A*) hypothesis of a single outlier the corresponding maximized log-likelihood is

$$L_A(\mathbf{x} \mid V) = -\frac{1}{2} \sum_{j \neq i} (\mathbf{x}_j - \bar{\mathbf{x}}') V^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}') \quad (6.1.3)$$

where  $\bar{\mathbf{x}}'$  is the sample mean of the  $(n - 1)$  observations excluding  $\mathbf{x}_i$  and  $i$  is chosen to maximize

$$L_A(\mathbf{x} | V) - L(\mathbf{x} | V).$$

Thus we are led to declare as the outlier  $\mathbf{x}_{(n)}$  that observation  $\mathbf{x}_i$  for which  $R_i(\bar{\mathbf{x}}, V)$  is a maximum, so that implicitly the observations have been ordered in terms of the reduced form of sub-ordering based on the distance measure  $R(\mathbf{x}; \bar{\mathbf{x}}, V)$ . Furthermore we will declare  $\mathbf{x}_{(n)}$  a *discordant* outlier if

$$R_{(n)}(\bar{\mathbf{x}}, V) \equiv (\mathbf{x}_{(n)} - \bar{\mathbf{x}})' V^{-1} (\mathbf{x}_{(n)} - \bar{\mathbf{x}}) = \max_{j=1, \dots, n} R_j(\bar{\mathbf{x}}, V)$$

is significantly large.

The null distribution of  $R_{(n)}(\bar{\mathbf{x}}, V)$  is not readily determined in exact form nor very tractable.

However, it has been studied by Siotani (1959) who discusses the problems associated with determining percentage points of  $R_{(n)}(\mathbf{x}_0, \Gamma)$  when  $\Gamma = V$  and  $\mathbf{x}_0$  is either  $\mathbf{0}$ ,  $\boldsymbol{\mu}$  or  $\bar{\mathbf{x}}$ . For the latter case, and of immediate relevance to us at the present stage of the discussion, he presents approximate upper 5,  $2\frac{1}{2}$ , and 1 per cent points of  $R_{(n)}(\bar{\mathbf{x}}, V)$  for  $p = 2(1)4$  and  $n = 3(1)10(2)20(5)30$ . The critical values for 5 per cent and 1 per cent tests of discordancy of a single outlier in a multivariate normal sample where  $V$  is known are reproduced as Table XXVI on page 329.

If  $\boldsymbol{\mu}$  were known, the corresponding  $R_i(\boldsymbol{\mu}, V)$  would be independent  $\chi_p^2$  variates and we would then have to relate their maximum  $R_{(n)}(\boldsymbol{\mu}, V)$  to the distribution of the maximum observation in a random sample of size  $n$  from a  $\chi_p^2$  distribution.

Gupta (1960) has considered the distribution of the order statistics from gamma samples and has tabulated percentage points for a range of sample sizes, and values of the two parameters. (Note however, that the last six lines of some of the tables in Gupta (1960) are incorrect; they are revised in the 'Errata' section of the journal *Technometrics*, 1960, 2, 523.) Suitably extracted and modified values serve for the outlier problem with  $\boldsymbol{\mu}$  and  $V$  known, and Table XXVII on page 329 presents upper 5 per cent and 1 per cent points of  $R_{(n)}(\boldsymbol{\mu}, V)$  for  $p = 2(2)10$  and  $n = 3(1)10, 25, 50, 100, 200, 500, 1000$ . Note that only even values of  $p$  ( $\geq 2$ ) are accessible from this table.

In the particular case of a bivariate sample ( $p = 2$ ),  $R_{(n)}(\boldsymbol{\mu}, V)/2$  has the distribution of the maximum of  $n$  independent exponential variates (mean 1) and its percentage points are easily determined. For a level- $\alpha$  test we would conclude that  $\mathbf{x}_{(n)}$  (the observation  $\mathbf{x}_i$  yielding  $R_{(n)}(\boldsymbol{\mu}, V)$ ) is a discordant outlier if  $R_{(n)}(\boldsymbol{\mu}, V) > \xi_\alpha$  where

$$\alpha = P\{R_{(n)}(\boldsymbol{\mu}, V) > \xi_\alpha\} = 1 - \{F(\xi_\alpha/2)\}^n \quad (6.1.4)$$

with

$$F(x) = 1 - e^{-x}.$$

Thus

$$\xi_\alpha = -2 \ln[1 - (1 - \alpha)^{1/n}] \quad (6.1.5)$$

provides an explicit value for use in the test.

An informal assessment of discordancy might be based on a graphical plot of the ordered  $R_j(\mu, V)$  [ $R_{(1)}(\mu, V), R_{(2)}(\mu, V), \dots, R_{(n)}(\mu, V)$ ] as ordinates against the expected values of the order-statistics of a random sample of size  $n$  from  $\chi^2_p$  as abscissae. If  $R_{(n)}(\mu, V)$  appears to be aberrant on this basis (lying above the expected straight line) we would adjudge the corresponding observation  $x_{(n)}$  to be a discordant outlier. With  $\mu$  unknown, and replaced by  $\bar{x}$ , or  $V$  replaced by  $S$ , the same procedure retains a measure of informal propriety and appeal.

Healy (1968) advances just such a graphical procedure for the detection of non-normality, and of outliers, for (principally) bivariate data. His basis for considering the distance measure  $R(\bar{x}; \bar{x}, S)$  is its intuitive appeal, rather than any justification in terms of a prescribed alternative model for explaining outliers or any specific test construction principle (such as the maximum likelihood ratio procedure). It is informative to reproduce a detailed example he discusses.

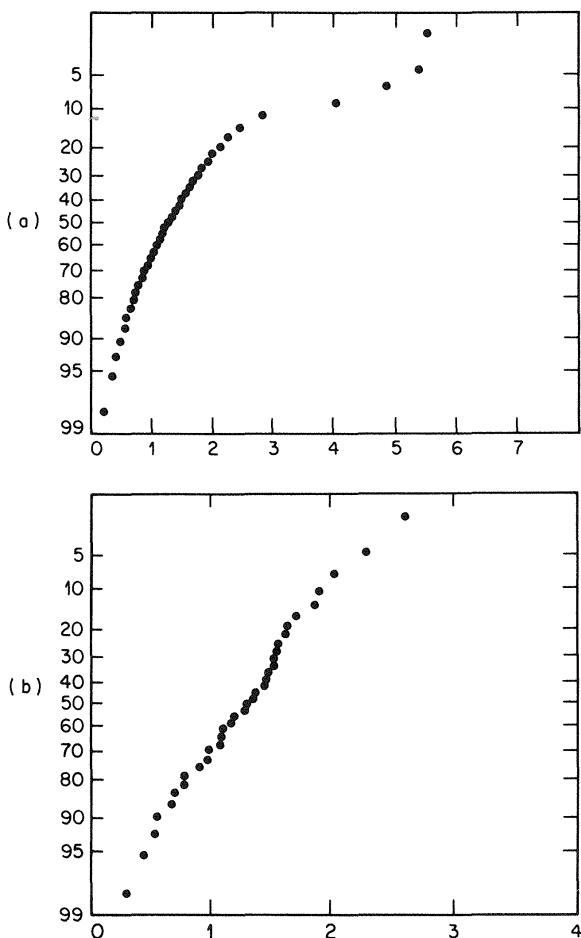
*Example 6.1. A bivariate sample of 39 observations of the logarithms of daily fat intake and serum cholesterol level for a group of hospital patients (data taken from Begg, Preston, and Healy, 1966) have values of  $R_j(\bar{x}, S)$  calculated, where  $S$  is the sample variance-covariance matrix. Ignoring the inaccuracies introduced by estimating  $\mu$  and  $V$ , we might consider plotting the ordered ‘distances’  $R_j$  against the expected values*

$$\frac{2}{n}, \frac{2}{n} + \frac{2}{n-1}, \frac{2}{n} + \frac{2}{n-1} + \frac{2}{n-2}, \dots, \frac{2}{n} + \frac{2}{n-1} \dots + \frac{2}{1}$$

(with  $n = 39$ ) of the order statistics for a sample from  $\chi^2$ . Instead, Healy introduces a further level of approximation, based on the approximate normality of  $\sqrt{\chi^2}$ , on the grounds that this simplifies the graphical process and even enhances the prospects of distinguishing outliers (in view of the reduced coefficient of variation of upper extreme sample values). Thus the square roots of the ordered distances for the sample of 39 observations are plotted on normal probability paper, with the results shown in Figure 6.1(a).

He concludes that the bivariate normal model is not unreasonable, but that there seem to be four outliers. Figure 6.1(b) is the corresponding plot with the four outliers removed. (Figures 6.1(a) and 6.1(b) are reproduced from Healy, 1968.)

We shall consider graphical procedures in more detail later (Section 6.2) but for the moment we continue with a more formal approach to the



**Figure 6.1** Normal plots of  $\sqrt{R_j(\bar{x}, S)}$  for the distribution of log daily fat intake and log serum cholesterol in a sample of 39 hospital patients. (a) Complete sample; (b) omitting four extreme values (reproduced by permission of the Royal Statistical Society)

assessment of discordancy based on the actual null distribution of  $R_{(n)}$  when  $\mu$  and  $V$  are unknown. The assumption of known  $V$  is in general unrealistic. We therefore proceed to examine the two-stage maximum likelihood ratio test for *model A* when both  $\mu$  and  $V$  are unknown (eschewing for the moment Healy's *pragmatic* simultaneous replacement of  $\mu$  by  $\bar{x}$  and  $V$  by  $S$  in the distance measure  $R(\mathbf{x}; \mu, V)$ ).

**Model A, V unknown**

With  $V$  unknown (as well as  $\mu$ ) the maximized log-likelihood under the basic model is (apart from the constant factor)

$$L(\mathbf{x}) = -\frac{n}{2} \log |A| \quad (6.1.6)$$

where  $|A|$  is the determinant of the matrix of sums of squares and cross-products of the observations about the component sample means: that is

$$A = \sum_{j=1}^n (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' \quad (6.1.7)$$

Under the *model A* alternative the maximized log-likelihood is

$$L_A(\mathbf{x}) = -\frac{n}{2} \log |A^{(i)}| \quad (6.1.8)$$

where  $A^{(i)}$  is the restricted matrix obtained on omission of  $\mathbf{x}_i$  and  $i$  is chosen to maximize

$$L_A(\mathbf{x}) - L(\mathbf{x}).$$

Thus when  $V$  is unknown it seems at first sight that quite a different principle is advanced for the declaration of an outlier  $\mathbf{x}_i$  and for the assessment of its discordancy. Here we are implicitly ordering the multivariate observations in terms of an *aggregated* form of reduced sub-ordering, based on the values of  $|A^{(i)}|$ . The  $|A^{(i)}|$  are ordered, and the observation corresponding with the smallest value of  $|A^{(i)}|$  is declared an outlier. Equivalently, if we denote

$$\mathcal{R}_i = \frac{|A^{(i)}|}{|A|}$$

the sample points are ‘ordered’ in accord with the ordered  $\mathcal{R}_i$  and the outlier is that observation corresponding with the smallest  $\mathcal{R}_i$ ,  $\mathcal{R}_{(1)}$ . If  $\mathcal{R}_{(1)}$  is significantly low in value the outlier is adjudged discordant. Thus the outlier is that observation whose removal from the sample effects the greatest reduction in the ‘internal scatter’ of the data set. But the distinction of principle for declaring an outlier in the case of unknown  $V$ , compared with the case where  $V$  is known, is less profound than might appear at first sight. Clearly we can rewrite

$$\mathcal{R}_i = \frac{\left| A - \left( \frac{n}{n-1} \right) (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \right|}{|A|} = 1 - \left( \frac{n}{n-1} \right) R_i(\bar{\mathbf{x}}, A) \quad (6.1.9)$$

and minimization of  $\mathcal{R}_i$  becomes equivalent to maximization of

$$R_i(\bar{\mathbf{x}}, A) = R_i(\bar{\mathbf{x}}, S)/(n-1). \quad (6.1.10)$$

Thus the outlier is again that observation whose distance from the body of the data set is a maximum, provided we estimate  $\mu$  and  $V$  by  $\bar{x}$  and  $S$  in the distance function. (This supports Healy's informal proposals.)

Once more the distribution of the test statistic is highly complicated. Little is known in detail of the joint distribution of the  $\mathcal{R}_j$  or more particularly of the distribution of the minimum,  $\mathcal{R}_{(1)}$ .

However, there is a deal of useful tabulated material on approximate percentage points for  $\mathcal{R}_{(1)}$ , and on the corresponding statistic for assessing the discordancy of a pair of outliers in a multivariate normal sample. The tables appear in work by Wilks (1963) which remains the most detailed applications-oriented study of outlier detection in multivariate data. We shall consider this work in some detail and apply it to a problem related to employment prospects of engineering graduates.

Concerned with testing outlying observations in a sample from a multivariate normal distribution with unknown mean vector and variance-covariance matrix, Wilks (1963) proposes an intuitively based representation of the sample in terms of the sum of squares of the volumes of all simplexes that can be found from  $p$  of the sample points augmented by the sample mean  $\bar{x}$ . He shows (Wilks, 1962) that this is just  $(p!)^{-2} |A|$ , where  $A$  is the matrix defined above. He calls  $|A|$  the *internal scatter* of the sample and suggests that a sensible criterion for the declaration of an outlier is to choose that sample member whose omission leads to the least value for the so-called *one-outlier scatter ratio*

$$\mathcal{R}_i = \frac{|A^{(i)}|}{|A|}.$$

But this is precisely the likelihood ratio criterion and corresponding test statistic. Wilks shows that the  $\mathcal{R}_i$  are identically distributed Beta variates  $\mathcal{B}\left(\frac{n-p-1}{2}, \frac{p}{2}\right)$ , with a joint distribution symmetric over  $R^n$  subject to

$$\sum \mathcal{R}_i = n\left(1 - \frac{p}{n-1}\right)$$

$$0 \leq \mathcal{R}_i \leq 1 \quad (i = 1, 2, \dots, n).$$

The joint distribution is intractable, but Wilks obtains an upper bound for the distribution function of  $\mathcal{R}_{(1)}$  (which he denotes  $r_1$ ), and hence lower bounds for the lower percentage points of  $\mathcal{R}_{(1)}$  thus enabling conservative tests of significance for a single outlier to be conducted. In comparison with exact results due to Grubbs (1950) for the case  $p = 1$ , the approximate values seem reasonable, though it must be stressed that their accuracy for  $p > 1$  has not been assessed since there is at present no yardstick (in terms of exact probabilities) for comparison.

Wilks (1963) tabulates lower bounds to the lower 10, 5,  $2\frac{1}{2}$ , and 1 per cent points of  $\mathcal{R}_{(1)}$  for  $p = 1(1)5$  and  $n = 5(1)30(5)100(100)500$ . For closer comparability with Tables XXVI and XXVII selected values from Wilks's tables have been transformed via (6.1.9) and (6.1.10) into approximate upper percentage points for  $R_{(n)}(\bar{x}, S)$ . Table XXVIII presents critical values for 5 per cent and 1 per cent tests of discordancy of a single outlier in a multivariate normal sample where  $\mu$  and  $V$  are estimated by  $\bar{x}$  and  $S$ , respectively. The table covers the ranges  $p = 2(1)5$ ,  $n = 5(1)10(2)20(5)50$ , 100, 200, 500. Note that the elements of  $S$  are unbiased sample variance and covariance estimates with divisors  $n - 1$ .

Wilks adopts a similar approach to the testing *en bloc* of 2, 3, or 4 outliers in a multivariate sample, by considering for the  $s$ -outlier case ( $s = 2, 3, 4$ ) the  $s$ -outlier scatter ratios

$$\mathcal{R}_{j_1, j_2, \dots, j_s} = \frac{|A^{(j_1, j_2, \dots, j_s)}|}{|A|}$$

where  $|A^{(j_1, j_2, \dots, j_s)}|$  is the internal scatter when  $x_{j_1}, x_{j_2}, \dots, x_{j_s}$  are omitted from the sample. Again it is the subset of observations that minimizes  $\mathcal{R}_{j_1, j_2, \dots, j_s}$  which is declared the outlying subset and their discordancy must be assessed in terms of how small is

$$r_s = \min \mathcal{R}_{j_1, j_2, \dots, j_s}.$$

For  $s = 2$ , an outlying subset of two observations, Wilks tabulates square roots of lower bounds for the lower percentage points of  $r_2$ . Extracted values are reproduced as Table XXIX on page 331 for precisely the same set of significance levels and values of  $p$  and  $n$  as are used in Table XXVIII. Thus Table XXIX gives critical values for 5 per cent and 1 per cent tests of discordancy (based on  $\sqrt{r_2}$ ) of outlier pairs in multivariate normal samples where  $\mu$  and  $V$  are unknown for  $p = 2(1)5$  and  $n = 5(1)10(2)20(5)50$ , 100, 200, 500.

*Example 6.2.* In 1974, the Southern Graduate and Student Section of the Institution of Electrical Engineers conducted a salary survey of its members, recording for each their age (to the nearest month) and their current annual salary (in £). Table 6.1 presents (by kind permission of the authors of the survey) the data for a sub-sample of 55 of the 374 returns. Figure 6.2 is a scatter diagram of the sub-sample. The observations L, M, and N appear as outliers and it is interesting to test their discordancy. For illustrative purposes we assume a bivariate normal distribution of ages and salaries, although no detailed study has been made of the underlying distribution.

Calculations show that M, L, and N yield (in decreasing order) the three

Table 6.1 Ages and salaries of a sample of 55 electrical engineers in the UK in 1974

Age (years)	Annual salary (£)	Age (years)	Annual salary (£)
27.67	2930	26.58	2600
23.42	2330	25.50	2250
24.67	2480	29.25	3600
27.92	4100	26.00	2750
26.92	2500	29.33	3500
28.92	3380	26.25	3400
26.08	2720	26.83	4500
29.92	4930	27.92	2800
29.50	3020	27.08	3610
22.42	1970	28.33	3100
23.42	1700	28.33	2900
28.00	3100	30.00	3600
23.00	1950	28.25	3600
25.17	2320	24.67	2030
26.58	2750	25.42	3520
22.75	1960	22.67	1900
25.00	2300	25.92	3230
30.00	4120	25.25	2500
23.50	3900 (L)	25.58	3020
26.58	5200 (M)	29.92	3610
28.25	3200	30.58	4200
25.33	2300	23.83	3050
25.25	2200	26.33	2760
27.92	3500	26.83	4000
29.50	3600	28.25	3100
21.17	1470 (N)		
25.67	2690		
28.42	2860		
26.00	3000		
27.42	3100		

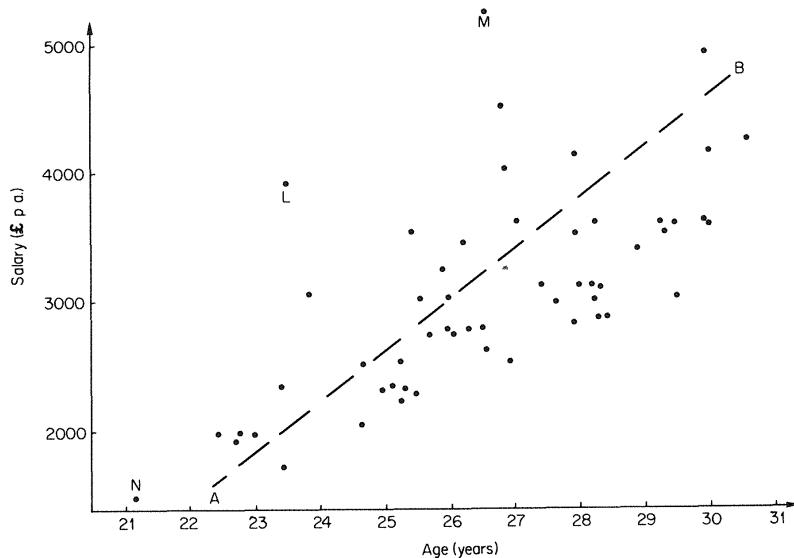
largest values of  $R_i(\bar{\mathbf{x}}, S)$ . We have

$$M: R_{(55)}(\bar{\mathbf{x}}, S) = 13.7,$$

$$L: R_{(54)}(\bar{\mathbf{x}}, S) = 9.2,$$

$$N: R_{(53)}(\bar{\mathbf{x}}, S) = 6.0.$$

From Table XXVIII we see that M is a discordant outlier. Explicit calculation (for  $n = 55$ ) from Wilks's tables gives 5 per cent and  $2\frac{1}{2}$  per cent critical values of 12.52 and 13.58 respectively. Thus M is discordant at the  $2\frac{1}{2}$  per cent level. Examination of the outlier pair (M, L) on Wilks's test gives  $\sqrt{r_2} = 0.769$  which is significant at the 5 per cent level (extrapolation in Table XXIX, or inspection of Wilks's tables, gives a 5 per cent critical value of 0.778; the  $2\frac{1}{2}$  per cent critical value is 0.767).



**Figure 6.2** Ages and salaries of electrical engineers (UK, 1974)

Of course the two tests of discordancy (for a single outlier, and for an outlier-pair) are not independent. It seems reasonable to assess M as a discordant outlier; the status of L is more questionable.

We should remark that the various tests of discordancy for multivariate normal outliers discussed in this section have the desirable property of being invariant with respect to the location and scale of the measurement basis of the observations.

For model B, with a single possible discordant value in a normal sample, Ferguson (1961a) has derived a multi-decision procedure (see Section 2.5) with certain optimal properties. See also Kudo (1957).

Suppose again that  $\mathbf{x}_1, \dots, \mathbf{x}_n$  is a random sample initially assumed to arise from  $\mathbf{N}(\mu, V)$ , with  $\mu$  and  $V$  unknown. Under the alternative, model B, hypothesis of a single outlier, we have

$$\begin{aligned} E(\mathbf{X}_j) &= \mu & (j = 1, 2, \dots, n), \\ V(\mathbf{X}_i) &= bV & (\text{some } i; b > 1), \\ V(\mathbf{X}_j) &= V & (j \neq i). \end{aligned}$$

Denoting by  $\mathcal{D}_i$  the decision to regard  $\mathbf{x}_i$  as the discordant value ( $i = 1, 2, \dots, n$ ), with  $\mathcal{D}_0$  the decision to declare no discordant value, Ferguson

considers those decision rules which satisfy four conditions:

- (a) each is invariant under the addition to  $\mathbf{X}_i$  of a constant vector;
- (b) each is invariant under the multiplication of  $\mathbf{X}_i$  by a common non-singular matrix;
- (c) the probability  $p_i(\mathcal{D}_i)$  of declaring  $\mathbf{x}_i$  the discordant value when this is true is independent of  $i$ ;
- (d) the probability of correctly declaring no discordant value is  $1 - \alpha$ , for a preassigned  $\alpha$  in  $(0, 1)$ ; that is, the procedure has size  $\alpha$ .

He seeks that decision rule which maximizes  $p_i(\mathcal{D}_i)$ . It turns out to have a familiar form: the optimum rule is to reject  $\mathbf{x}_i$  as the discordant value if  $\mathbf{x}_i$  yields the maximum value  $R_{(n)}(\bar{\mathbf{x}}, S)$  and

$$R_{(n)}(\bar{\mathbf{x}}, S) > K$$

where  $K$  is chosen to satisfy the test size condition (d).

Thus, as for *model A*, we again declare the outlier to be the observation with maximum generalized distance  $R_j(\bar{\mathbf{x}}, S)$ , and assess it as discordant if that maximum,  $R_{(n)}(\bar{\mathbf{x}}, S)$ , is sufficiently large.

Additionally, however, Ferguson demonstrates that this procedure is *uniformly best over all values of b > 1*.

Thus when  $\mu$  and  $V$  are unknown it is immaterial whether we adopt the *model A* or *model B* formulation of the alternative hypothesis describing the occurrence of a single outlier. In either case the test has the same form, and can be implemented by using the Table XXVIII on page 330.

Summarizing the status of the (identical) *model A*, and *model B*, tests of discordancy when  $\mu$  and  $V$  are unknown, we have that in, relation to *model A*, the test is the likelihood ratio test, whereas relative to *model B* it has the uniform optimality property of maximizing  $p_i(\mathcal{D}_i)$  irrespective of the value of  $b$ .

Siotani (1959) tabulates approximate percentage points for a studentized form of  $R_{(n)}(\bar{\mathbf{x}}, S)$  where  $S$  is replaced by an external unbiased estimate  $S_v$  of  $V$  having a Wishart distribution with  $v$  degrees of freedom. These are of value for an informal test of discordancy of a single outlier in a multivariate sample, where  $V$  is *not* estimated from the sample itself but by means of such an external estimate.

Table XXX on pages 332–333, extracted from Siotani (1959), presents approximate 5 per cent and 1 per cent critical values for a test of a single outlier in a bivariate normal sample ( $p = 2$ ) where  $\mu$  and  $V$  are unknown, and  $V$  is estimated by  $S_v$ . The table covers the values  $n = 3(1)14$  and  $v = 20(2)40(5)60, 100, 150, 200$ .

## 6.2 INFORMAL DETECTION OF MULTIVARIATE OUTLIERS

A host of proposals have been made for informally detecting outliers in multivariate data by quantitative or graphical methods. These cannot be

regarded as tests of discordancy; they are designed more as aids to intuition in picking out multivariate observations which are suspiciously aberrant from the bulk of the sample.

In univariate samples the ‘suspicious’ observation which is to be declared an outlier is obvious on simple inspection. It is an extreme observation in the sample. In multivariate data the extremeness concept is a nebulous one, as we have remarked above. Various forms of initial processing of the data, involving study of individual marginal components of the observations, judicious reduction of the multivariate observations to scalar quantities in the forms of distance measures or linear combinations of components, changes in the coordinate bases of the observations, and appropriate methods of graphical representation, can all help to identify or highlight a suspicious observation. If several such procedures are applied simultaneously (or individually) to a set of data they can help to overcome the difficulty caused by the absence of a natural overall ordering of the sample members. An observation which clearly stands out on one, or preferably more, processed re-representations of the sample becomes a firm candidate for identification as an outlier.

Gnanadesikan and Kettenring (1972) remark:

The consequences of having . . . [outliers] in a multivariate sample are intrinsically more complex than in the much discussed univariate case. One reason for this is that a multivariate outlier can distort not only measures of location and scale but also those of orientation (i.e. correlation). A second reason is that it is much more difficult to characterize the multivariate outlier. A single univariate outlier may be typically thought of as ‘the one which sticks out on the end’, but no such simple idea suffices in higher dimensions. A third reason is the variety of types of multivariate outliers which may arise: a vector response may be faulty because of a gross error in one of its components or because of systematic mild errors in all of its components.

The complexity of the multivariate case suggests that it would be fruitless to search for a truly omnibus outlier protection procedure. A more reasonable approach seems to be to tailor detection procedures to protect against specific types of situations, e.g., correlation distortion, thus building up an arsenal of techniques with different sensitivities. This approach recognizes that an outlier for one purpose may not necessarily be one for another purpose! However, if several analyses are to be performed on the same sample, the result of selective segregation of outliers should be more efficient and effective use of the available data.

Such methods do not (in general) lead to any formal test of discordancy; they seldom even adopt any specific assumptions about the distribution from which the sample has arisen or about the nature of an alternative (outlier generating) hypothesis. They are to be viewed as initial data screening procedures, in the spirit of the current interest in ‘data analysis’ methods for the representation and summary of large-scale sets of data.

### **6.2.1 Marginal outliers**

We should not underestimate the role to be played by the marginal samples (that is, the univariate samples of each component value in the multivariate

data) in the identification of outliers. Firstly, we know what we mean by a marginal outlier: it is an extreme member of the marginal sample. Secondly, we have facilities for testing the discordancy of such univariate outliers for a range of different basic models (and we can adopt models to explain the outliers). Clearly outliers in different marginal samples need not be independent, and conclusions about discordancy will need to reflect this. And thirdly, perhaps most important, it is quite plausible to expect outliers to be exhibited within specific components of the multivariate observations. This is particularly true when the outliers arise from gross errors of measurement or recording, where almost inevitably it will be a single component in the multivariate observation which will suffer.

The techniques described in Chapter 3 will be appropriate to the testing of marginal outliers.

### 6.2.2 Linear constraints

Another situation in which we may have easy access to the detection of outliers is where we anticipate a simple (usually linear) relationship between the components of the multivariate observation or between the expected values of the components. For example our multivariate observation might consist of proportionate measurements, such as the proportions of the total body length of a reptile corresponding with biologically distinct sections of the body, or it may be the three angles of a triangle in a geographic survey. In the first case consistency of representation demands that the proportions should have unit sum; in the second the sum of the components should, apart from errors of measurement, add to  $180^\circ$ . In either case, marked departures of the sum of the component values from their expected sum can highlight gross errors of measurement or of recording as an indication of outliers. Fellegi (1975) comments on the presence of outliers in the editing of multivariate data where just such 'pre-identified relationships' hold. He includes consideration of less specific forms of relationship where, for example, we have information on the reasonable range of values which may be taken by some ratio of marginal components.

Note that outliers identified in this way need not (indeed are unlikely to) show up merely on consideration of the marginal samples.

### 6.2.3 Graphical methods

In relation to multivariate outliers, Rohlf (1975) remarks as follows;

Despite the apparent complexity of the problem, one can still characterize outliers by the fact that they are somewhat isolated from the main cloud of points. They may not 'stick out on the end' of the distribution as univariate outliers must, but they must 'stick out' somewhere. Points which are not internal to the cloud of points (i.e. which are somewhere on the surface of the cloud) are potentially outliers. Techniques

which determine the position of a point relative to the others would seem to be useful. A second important consideration is that outliers must be separated from the other points by distinct gaps.

With this emphasis it is natural to consider different ways in which we can merely look at the data to see if they seem to contain outliers. A variety of methods employing different forms of pictorial or graphical representation have been proposed with varying degrees of sophistication in terms of the pre-processing of the data prior to its display. For obvious reasons, bivariate data is the most amenable to informative display, although it will be apparent that some of the approaches do not depend vitally on the dimensionality of the data. The basic ideas and interests in such methods of 'informal inference' applied to general problems of analysis of multivariate data (not solely the detection of outliers) are described by Gnanadesikan (1973). A variety of papers over several years by the same author, often in conjunction with others, expand and illustrate these methods; see, for example, Wilk and Gnanadesikan (1964), Gnanadesikan and Wilk (1969), and the lengthy review paper by Gnanadesikan and Kettenring (1972) on which much of the summary below is based. This paper contains a multitude of informative illustrative examples.

We shall consider some possibilities for demonstrating the presence of outliers in (predominantly) bivariate data. The most rudimentary form of representation is the scatter diagram of two, out of the  $p$ , components. Figure 6.2 shows the scatter diagram for the data of salaries and ages of electrical engineers (see page 218). Some observations do seem to 'stick out' and to be separated from others by 'distinct gaps'; notably the observations L and M previously identified as discordant outliers. But in a different respect, the observation N also seems somewhat suspicious. The observation L (and to a lesser degree M) might well have the effect of reducing the apparent correlation between age and salary in the population, whilst N leads one to assume a larger variation in ages, or salaries, than would have appeared plausible in its absence. This effect of N is heightened if we consider the projection of the observations on to the line AB through the data set as shown on Figure 6.2 (roughly the regression line of salary on age) whilst L and M now appear in no way aberrant. In contrast if we project onto the perpendicular to AB then L and M are particularly extreme, whilst N appears more reasonable.

This example embodies many of the considerations employed in designing graphical methods for exhibiting multivariate outliers. In the first place, the scatter diagram itself may throw up outliers as observations on the periphery of the 'data cloud', distinctly separated from others. The marginal samples may or may not endorse such a declaration, and yield a formal assessment of discordancy (this could be possible for N, not for L or M). The perturbation of some aggregate measure, such as the correlation coefficient, from what is anticipated may reveal the presence of outliers (likely here for L and M, not

for N). A change of coordinate basis, and re-representation of the data on the new basis, can reveal outliers not immediately apparent previously. Rotation of the axes of Figure 6.2 in the direction of AB (or of its perpendicular) can help to identify L and M, (or N) depending on which of the new axes is considered. An *appropriate* graphical representation of the *ordered* sample values (either marginally in the original data, or for particular linear combinations of the components corresponding with a transformations of axes) can dramatically augment the visual impact of outliers.

We shall consider in more detail some work which utilizes such ideas.

#### 6.2.4 Principal component analysis method

Several writers have suggested performing a preliminary principal component analysis on the data, and looking at sample values of the projection of the observations on to the principal components of different order. The example above, on electrical engineers' salaries and ages, shows how projecting the observations on the leading or secondary principal component axes (roughly AB and its perpendicular) can highlight different types of outlier. This distinction in the relative utility of the first few, and last few, principal components in outlier detection is basic to the methods described in the literature. Gnanadesikan and Kettenring (1972) discuss this in some detail, remarking how the first few principal components are sensitive to outliers inflating variances or covariances (or correlations, if the principal component analysis has been conducted in terms of the sample correlation matrix, rather than the sample covariance matrix), whilst the last few are sensitive to outliers adding spurious dimensions to the data or obscuring singularities.

Suppose we write

$$Z = LX \quad (6.2.1)$$

where  $L$  is a  $p \times p$  orthogonal matrix whose rows,  $\mathbf{l}_i'$ , are the eigenvectors of  $S$  corresponding with its eigenvalues,  $c_i$ , expressed in descending order of magnitude and  $X$  is the  $p \times n$  matrix whose  $i$ th column is the transformed observations  $\mathbf{x}_i - \bar{\mathbf{x}}$ . Then the  $\mathbf{l}_i$  are the principal component coordinates and the  $i$ th row of  $Z$ ,  $\mathbf{z}_i'$ , gives the projections on to the  $i$ th principal component coordinate of the deviations of the  $n$  original observations about  $\bar{\mathbf{x}}$ . Thus the top few, or lower few, rows of  $Z$ , provide the means of investigating the presence of outliers affecting the first few, or last few, principal components.

The construction of scatter diagrams for pairs of  $\mathbf{z}_i$  (among the first few, or last few, principal components) can graphically exhibit outliers. Additionally univariate outlier tests can be applied to individual  $\mathbf{z}_i$ ; or the ordered values in  $\mathbf{z}_i$  can usefully be plotted against an appropriate choice of plotting positions. What is 'appropriate' is not easily assessed in any exact form, especially in the absence of reliable distributional assumptions about the original data. However, if  $p$  is reasonably large, it is likely that the linear

transformations involved in the principal component analysis may lead (*via* Central Limit Theorem arguments) to the  $\mathbf{z}_i$  being samples from approximately normal distributions. In such cases *normal probability plotting* in which the  $j$ th ordered value in  $\mathbf{z}_i$  is plotted against  $\alpha_j$ , where

$$\alpha_j = E[U_{(j)}]$$

with  $U_{(j)}$  the  $j$ th order statistic of the normal distribution,  $\mathbf{N}(0, 1)$ , may well reveal outliers as extreme points in the plot lying off the linear relationship exhibited by the mass of points in the plot. Such an informal procedure has been found to be a useful aid to the identification of multivariate outliers.

To illustrate this we again consider the salary/age data for electrical engineers. Figures 6.3(a) and 6.3(b) show normal probability plots for the first and second principal components respectively. The first principal component is essentially just the salary. Figure 6.3(a) shows no marked contradiction of normality. The outliers M and N show as extreme values (L is inconspicuous) although they do not lie off the linear relationship in the manner we would expect were they discordant outliers. In the plot (Figure 6.3b) of projections onto the second principal component, L and M are distinguished as extremes (N is inconspicuous) and they do lie *below* the linear relationship, indicative of discordancy.

Added flexibility of approach is provided by basing principal component analysis on the sample correlation matrix,  $R$ , instead of on  $S$ , and also by following the proposal of Gnanadesikan and Kettenring (1972) of replacing  $R$  or  $S$  by modified *robust* estimates. The robustness aspect will be taken up in the discussion of the *accommodation* of multivariate outliers in Section 6.3.

Some modifications of approach to outlier detection by principal component analysis are suggested by Hawkins (1974) and by Fellegi (1975). Hawkins considers specifically the case where  $\mathbf{X}$  is multivariate normal, and gives some consideration to different alternative hypotheses explaining the presence of a single outlier.

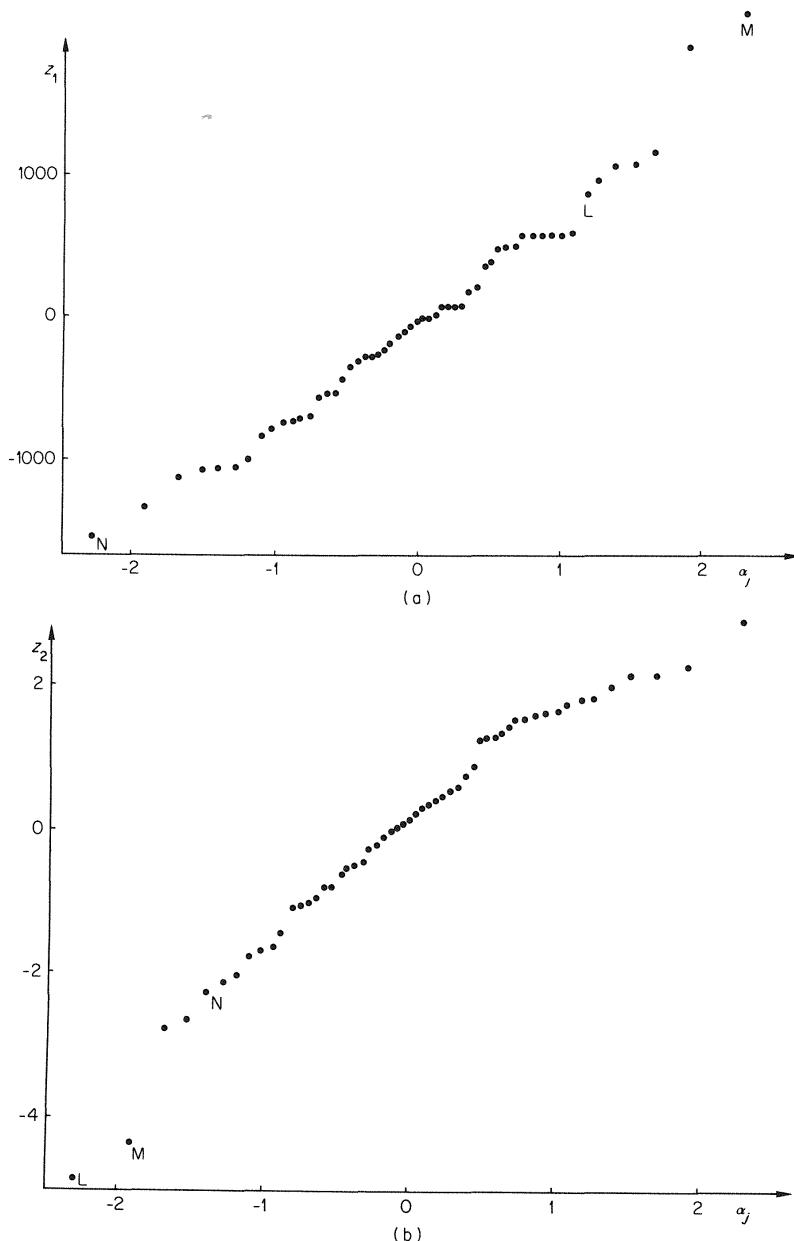
### 6.2.5 Use of generalized distances

Another way in which informal quantitative and graphical procedures may be used to exhibit outliers is to construct reduced univariate measures based on the observations  $\mathbf{x}_j$  (analogous to the distance functions more formally considered earlier). Gnanadesikan and Kettenring (1972) consider various possible measures in the classes:

$$\text{I: } (\mathbf{x}_j - \bar{\mathbf{x}})' S^b (\mathbf{x}_j - \bar{\mathbf{x}}), \quad (6.2.2)$$

$$\text{II: } (\mathbf{x}_j - \bar{\mathbf{x}})' S^b (\mathbf{x}_j - \bar{\mathbf{x}}) / [(\mathbf{x}_j - \bar{\mathbf{x}})' (\mathbf{x}_j - \bar{\mathbf{x}})]. \quad (6.2.3)$$

Particularly extreme values of such statistics, possibly demonstrated by graphical display, may reveal outliers of different types. Such measures are



**Figure 6.3** Normal probability plots for principal components of engineers' salary/age data. (a) First principal component; (b) Second principal component

of course related to the projections on the principal components and Gnanadesikan and Kettenring (1972) remark that, with class I measures, as  $b$  increases above +1 more and more emphasis is placed on the first few principal components whereas when  $b$  decreases below -1 this emphasis progressively shifts to the last few principal components (a similar effect holds for class II measures according as  $b \geq 0$ ). Extra flexibility arises by considering  $\mathbf{x}_j - \bar{\mathbf{x}}_j$  ( $j \neq j'$ ) rather than  $\mathbf{x}_j - \bar{\mathbf{x}}$  in the different measures, or  $R$  in place of  $S$ .

Let us consider some specific examples of the case I measures.

$$(b=0) \quad q_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}}) = \frac{n}{n-1} [\text{tr}(A) - \text{tr}(A^{(0)})]. \quad (6.2.4)$$

This squared Euclidean distance from  $\bar{\mathbf{x}}$  is sensitive to outliers ‘inflating the overall scale’.

$$(b=1) \quad t_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'S(\mathbf{x}_j - \bar{\mathbf{x}}) = \sum_i c_i [l_i'(\mathbf{x}_j - \bar{\mathbf{x}})]^2, \quad (6.2.5)$$

is sensitive to outliers affecting the ‘orientation and scale of the first few principal components of  $s$ ’.

$$(b=-1) \quad d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'S^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}) = \sum_i c_i^{-1} [l_i'(\mathbf{x}_j - \bar{\mathbf{x}})]^2 \quad (6.2.6)$$

is particularly useful for ‘uncovering observations which lie far afield from the general scatter of points’.

For graphical display of outliers, the ‘gamma-type probability plots’ of ordered values, with appropriately estimated shape parameter, are a useful approximate procedure. Essentially the argument is as follows. If the multivariate observation  $\mathbf{x}_j$  comes from a normal distribution, then the distance measures  $R_j(\bar{\mathbf{x}}, \Gamma)$  may be regarded as (approximately) independent observations from a gamma distribution, whatever form is taken for  $\Gamma$ . Thus if we knew, or could reasonably estimate, the parameters in the gamma distribution, a plot of the ordered  $R_j(\bar{\mathbf{x}}, \Gamma)$  against quantiles of the gamma distribution should be linear with anomalous  $R_j(\bar{\mathbf{x}}, \Gamma)$  (for example, discordant outliers) showing as extreme values lying markedly off the overall linear relationship. (See Wilk and Gnanadesikan, 1964; Wilk, Gnanadesikan, and Huyett, 1962a, 1962b, and further comments in Section 6.2.8.) If  $\mathbf{X}$  is multivariate normal then the exact marginal distribution of the  $d_j^2$  is known to be related to a Beta form with parameters  $(n-p-1)/2$  and  $p/2$ , but the  $d_j^2$  are not, of course, independent (see Section 6.1). We note again that consideration of the maximum value of  $d_j^2$  is equivalent to Wilks’s (1963) method; see also (above) the proposals of Healy (1968) for plotting the  $d_j^2$  when  $\mathbf{X}$  is approximately normal.

Rao (1964) proposes examination of the sums of squares of the lengths of the projections of individual observations on the last few ( $q$ ) principal

component coordinates for assessing the propriety of individual observations. Thus outliers may be revealed by particularly large values of

$$\sum_{i=p-q-1}^p [l_i^*(\mathbf{x}_i - \bar{\mathbf{x}})]^2$$

The suggestions of Gnanadesikan and Kettenring (1972) for informally considering residuals in least-squares fits of structured models, as a means of identifying outliers, are more appropriate to the discussion of outliers in regression models and designed experiments (Chapter 7).

### 6.2.6 Fourier-type representation

Gnanadesikan (1973) gives an interesting illustration of the potential use, for detecting outliers, of a novel means of representing multivariate observations, due to Andrews (1972). Andrews suggests that  $\mathbf{x}_j = (x_{1j}, x_{2j}, \dots, x_{pj})'$  should be represented by the function

$$f_{\mathbf{x}_j}(t) = x_{1j}/\sqrt{2} + x_{2j} \sin t + x_{3j} \cos t + x_{4j} \sin 2t + x_{5j} \cos 2t \dots \quad (6.2.7)$$

over the range  $(-\pi, \pi)$  for  $t$ . Each sample point in  $p$ -space then appears as a curve over such values of  $t$ . The idea is that this might reveal certain important qualitative features in the data.

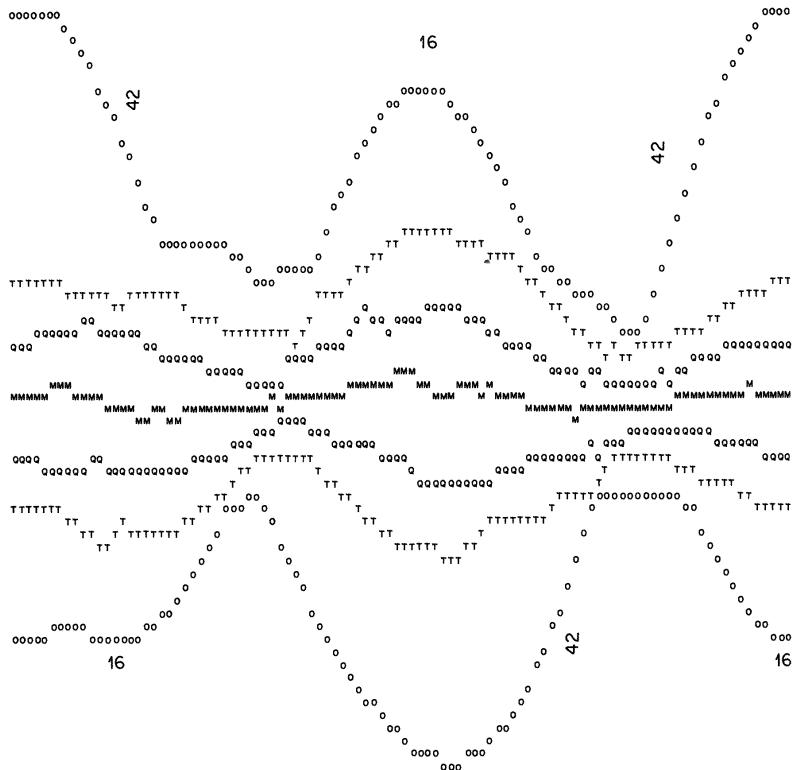
Gnanadesikan (1973) shows in an example how it might usefully distinguish outliers. He considers a quadrivariate sample of 50 observations on log-lengths and log-widths of sepals and petals for *Iris setosa* described by Fisher (1936). He chooses a grid of values of  $t$  over  $(-\pi, \pi)$ , determines  $f_{\mathbf{x}_j}(t)$  ( $j = 1, 2, \dots, 50$ ) over the grid, and at each grid value estimates certain quantiles of  $f_{\mathbf{x}_j}(t)$  from the data. The quantiles chosen are the 10, 25, 50, 75, and 90 percentiles. These are presented graphically along with any individual  $f_{\mathbf{x}_j}(t)$  values, at each  $t$ , outside the deciles. The results are shown in Figure 6.4 where we see a very clear indication of the outlying nature of observations number 16 and 42. (The median values are labelled M, quartiles Q, and extreme deciles T.)

We would not claim that other techniques, such as residuals based on a principal components analysis, would fail to exhibit these outliers. But this use of Andrews' representation may prove to have interesting possibilities for the study of outliers.

### 6.2.7 Correlation methods

We have already remarked on the way in which outliers may affect, and be revealed by, the correlation structure in the data. Some proposals for identifying multivariate outliers specifically consider this matter.

Gnanadesikan and Kettenring (1972) suggest that we examine the product-moment correlation coefficients  $r_{-j}(s, t)$  relating to the  $s$ th and  $t$ th



**Figure 6.4** Andrews' plot of the *Iris setosa* data (reproduced by permission of R. Gnanadesikan and the International Statistical Institute)

marginal samples after the omission of the single observation  $\mathbf{x}_j$ . As we vary  $j$  we can examine, for any choice of  $s$  and  $t$ , the way in which the correlation changes, substantial variations reflecting possible outliers.

Devlin, Gnanadesikan, and Kettenring (1975) make use of the *influence function* of Hampel (1974) to investigate how outliers affect correlation estimates in bivariate data ( $p = 2$ ). Their main interest is in robust estimation of correlation—see Section 6.3. But they are also concerned with the detection of outliers *per se*. They consider a multivariate distribution indexed by a parameter  $\theta$  and define in relation to an estimator  $\hat{\theta}$  the ‘sample influence function’

$$I_-(\mathbf{x}_j; \hat{\theta}) = (n - 1)(\hat{\theta} - \hat{\theta}_{-j}) \quad (j = 1, 2, \dots, n) \quad (6.2.8)$$

where  $\hat{\theta}_{-j}$  is an estimator of the same form as  $\hat{\theta}$  based on the sample omitting the observation  $\mathbf{x}_j$ . We see that  $\hat{\theta} + I_-$  is just the  $j$ th jackknife pseudo-value. As a convenient first-order approximation to the sample

influence function of  $r$ , the product-moment correlation estimate in a bivariate sample, they propose (with an obvious notation)

$$I_{-}(x_{1j}, x_{2j}; r) = (n-1)(r - r_{-j}). \quad (6.2.9)$$

$I_{-}(x_{1j}, x_{2j}; r)$  provides an estimate of the influence on  $r$  of the omission of the observation  $(x_{1j}, x_{2j})$ .

Two suggestions are made for presenting graphically how  $I_{-}(x_{1j}, x_{2j}; r)$  varies over the sample, with a view to identifying as outliers the observations which exhibit a particularly strong influence on  $r$ . The first amounts to superimposing selected (hyperbolic) contours of  $I_{-}(x_1, x_2; r)$  on the scatter diagram, thus distinguishing the outliers. Some qualitative comments are made (and illustrated) concerning the choice of which contours to plot. The second relates to the sample influence function of the Fisher transformation  $z(r) = \tanh^{-1}(r)$ :

$$I_{-}(x_{1j}, x_{2j}; z(r)) = (n-1)[z(r) - z(r_{-j})]. \quad (6.2.10)$$

For large  $n$ , the distribution of  $I_{-}(x_{1j}, x_{2j}; z(r))$  is approximately that of the product of two independent standard normal variables, and it is proposed that ordered values of  $I_{-}$  be plotted against the appropriate quantiles. The distinct  $I_{-}$  values over the sample are not seriously correlated, and a further normalizing transformation is proposed prior to the probability plotting. Again, it will be extreme values in the plot, lying away from the overall linear relationship, which indicate outliers.

### 6.2.8 A ‘gap test’ for multivariate outliers

We noted earlier the characterization of multivariate outliers suggested by Rohlf (1975): that they are separated from other observations ‘by distinct gaps’. Rohlf has used this idea to develop a *gap test* for multivariate outliers based on minimum spanning trees. Eschewing the nearest neighbour distances as measures of separation, in view of the masking effect a cluster of outliers may exert on each other, he considers instead the lengths of edges in the minimum spanning tree MST (or shortest simply connected graph) of the data set as measures of adjacency. He argues that a single isolated point will be connected to only one other point in the MST by a relatively large distance, and that *at least one* edge connection from a cluster of outliers must also be relatively large. Accordingly a gap test for outliers is proposed with the following form. Firstly, examination of the marginal samples yields estimates  $s_k$  ( $k = 1, 2, \dots, p$ ) of the standard deviations. The observations are rescaled as  $x'_{ki} = x_{ki}/s_k$  ( $k = 1, 2, \dots, p$ ;  $i = 1, 2, \dots, n$ ). Distances between  $\mathbf{x}'_i$  and  $\mathbf{x}'_j$  in the MST are calculated as

$$d_{ij} = \left\{ \sum_{k=1}^p [(x'_{ki} - x'_{kj})^2]/p \right\}^{1/2} \quad (6.2.11)$$

and in particular we denote by  $z_i$  the lengths of the  $n-1$  edges of the MST.

The  $z_i$  are now examined for homogeneity, either by means of a probability plot of their ordered values or by testing if the ratio of the square of the maximum,  $z_{(n-1)}^2$ , to the average,  $\bar{z}^2$ , of the squares is of reasonable value.

The 'gamma-type plot' of the  $z_{(i)}$  against quantiles of a gamma distribution has heuristic justification if  $\mathbf{X}$  comes from a  $p$ -variate normal distribution, on the following argument. If the components of  $\mathbf{X}$  were independent normal (each with unit variance) the inter-point squared Euclidean distances would be independently distributed as  $2\chi_p^2$ . If the components of  $\mathbf{X}$  are not independent, these distances will be dependent and may not follow too closely a gamma distribution. However, Rohlf claims that empirical investigations demonstrate that the particular subset of squared edge distances,  $z_i^2$ , do appear to have approximately independent common gamma distributions (on the assumption of homogeneity, that is, absence of discordant outliers). The relevant shape parameter will need to be estimated either (iteratively) by the maximum likelihood method, or by using the order statistics approach by Wilk, Gnanadesikan, and Huyett (1962a, 1962b). The value of the scale parameter will not need to be estimated since its value affects only the slope of the gamma plot and not its linearity, and it is therefore irrelevant to the detection of markedly anomalous values (here discordant outliers).

Wilk, Gnanadesikan, and Huyett (1962b) consider maximum likelihood estimation of the scale and shape parameters,  $\lambda$  and  $\eta$ , in the gamma distribution,  $\Gamma(\eta, \lambda)$ , based on an ordered random sample of observations,  $y_{(1)}, y_{(2)}, \dots, y_{(n)}$ . They show that  $\hat{\eta}$  satisfies

$$\frac{\Gamma'(\hat{\eta})}{\Gamma(\hat{\eta})} - \ln \hat{\eta} = \ln Q \quad (6.2.12)$$

where  $\Gamma()$  denotes the gamma function and  $Q$  is the ratio of the geometric and arithmetic means of the  $y_{(i)}$ . They present useful tabulated aids for determining  $\hat{\eta}$ . It is interesting to note that  $\eta$  may be estimated separately from  $\lambda$ , an important consideration in that the probability plotting procedure does not require  $\lambda$  to be known (it can be arbitrarily assigned).

Rohlf (1975) also remarks on a further advantage of such a means of estimation for current purposes:  $\hat{\eta}$  does not depend strongly on the larger values in the ordered sample, so that  $\hat{\eta}$  will be reasonably robust against the very outliers we are seeking to identify. In Wilk, Gnanadesikan, and Huyett (1962a) the 'gamma-type' probability plot is described in detail and useful tables of quantiles of the gamma distribution are presented.

If something nearer a formal test of discordancy for a single outlier is required, Rohlf makes the following proposals. If we knew  $\lambda$  and  $\eta$  in the (approximate) gamma distribution for which the  $z_i$  are (approximately) independent observations, then we could compare  $z_{(n-1)}/\lambda$  with the upper percentage points for the maximum observation in a sample of size  $n-1$ .

from a gamma distribution with shape parameter  $\eta$ . Equivalently

$$z_{(n-1)}^2 / [(n-1)\bar{z}^2],$$

where  $\bar{z}^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} z_i^2$ , has a Beta distribution with parameters  $\eta$  and  $(n-1)\eta$  (independent of  $\lambda$ ) and use can be made of existing results on (approximate) upper percentage points of such a Beta distribution (see Rohlff, 1975, for details). Not knowing the value of  $\eta$  it is proposed that we should relate  $z_{(n-1)}^2 / [(n-1)\bar{z}^2]$  to the approximate upper percentage points of the appropriate Beta distribution for an *estimated* value  $\hat{\eta}$ . Rohlff presents a table of upper bounds to the upper 5 per cent and 1 per cent points for  $n = 10, 20(20)100, 200$  and  $\eta = 0.1(0.1)1.0(0.5)5.0, 6.0(2.0)12.0$ .

The idea of using the MST to reflect outliers is interesting but clearly needs more detailed study and illustration. Rohlff's proposals are specifically concerned with normal data, and we have of course dealt at length in Section 6.1 with other proposals for this case.

### 6.3 ACCOMMODATION OF MULTIVARIATE OUTLIERS

We have noted the relative paucity of methods for detecting, or assessing the discordancy of, multivariate outliers. Inference techniques for the accommodation of multivariate outliers are even less in evidence. The few proposals in the literature again concentrate on a basic normal distribution or have an informal structure with intuitive, rather than theoretically justified, appeal. We shall review briefly some of the techniques available for estimating parameters (in multivariate distributions) in ways which are likely to be robust against the presence of outliers.

We have referred in Chapters 2 and 4 to the 'premium-protection' rules of Anscombe (1960a) which take the form of joint rejection/estimation procedures. For a univariate sample  $x_1, x_2, \dots, x_n$  from  $N(\mu, \sigma^2)$ , with a location-slippage alternative model  $N(\mu + a, \sigma^2)$  for the generation of at most one of the  $x_j$ , we examine the maximum absolute residual

$$\max_{j=1, 2, \dots, n} |x_j - \bar{x}|.$$

If this is sufficiently large we omit the observation  $x_i$  yielding the maximum absolute residual and estimate  $\mu$  by the sample mean of the remaining  $n-1$  observations; otherwise we estimate  $\mu$  merely by  $\bar{x}$ . There is an obvious multivariate generalization in which we consider  $R_{(n)}(\bar{x}, S)$  [or  $R_{(n)}(\bar{x}, V)$ , depending on our state of knowledge about  $V$ ] and if it is sufficiently large omit the observation  $x_i$  yielding  $R_{(n)}$  before estimating  $\mu$  from the residual sample; if  $R_{(n)}$  is not sufficiently large we estimate  $\mu$  from the total sample by means of  $\bar{x}$ . Such an approach is implicitly taken up by Golub, Guttman, and Dutter (1973) in greater detail and generality. Their

work is more general in that they consider a general normal linear model and augment the Anscombe-type rule by corresponding rules based on Winsorization and ‘semi-Winsorization’ of residuals. The greater detail is reflected in their discussion of the problems of determining the premium and protection. To ease the task of such determinations they also propose the use of orthogonalized (‘adjusted’) residuals as a basis for approximating the premium and protection. This aspect is discussed more fully in our study of the linear model situation in Section 7.3. However, since an uncorrelated error structure is assumed in the detailed discussion, the results are not immediately applicable to the case of a general multivariate normal sample.

A different approach to multivariate outlier accommodation can be based on the Bayesian analysis by Guttman (1973b). Guttman is concerned with the posterior distribution of  $\alpha$  for a basic normal model  $N(\mu, V)$ , with a mean-slippage alternative model  $N(\mu + \alpha, V)$  for at most one of the observations. Examination of the posterior distribution of  $\mu$  (and  $V$ ) is germane to the accommodation issue. Further reference to this approach to the multivariate problem, with some detailed results for the corresponding univariate case, appears in Section 8.1.

So far we have concentrated on outlier-robust estimation of the mean  $\mu$ . An alternative concern might be with outlier-robust estimation of the variance-covariance matrix  $V$ , or with derived quantities such as correlation coefficients. Gnanadesikan and Kettenring (1972) are concerned with robust estimation of multivariate location and dispersion. Whilst not preoccupied with the outlier problem some of their proposed robust estimators will *en passant* provide protection against outliers. (But we must recall our earlier discussion of the multifarious effects and manifestations of outliers in the multivariate case, influencing as they may scale, correlation structure, high- or low-order components, and so on.) Various robust estimators of  $\mu$  are reviewed (and have their performance characteristics examined by simulation in the bivariate normal case). The estimators mostly take the form of vectors of robust univariate estimators for the distinct marginal components of  $\mu$ , such as the vector of sample medians, or of trimmed means. The further prospect of using the vector of Winsorized means is not examined in detail. As far as robust estimation of dispersion parameters is concerned they consider the usual variance estimates based on trimmed or Winsorized marginal samples, or examination of the slope of appropriate probability plots. Covariances and correlations are less readily estimated robustly. One possibility is to use the relationship that, for any two random variables  $X_1$  and  $X_2$ ,

$$\text{cov}(X_1, X_2) = \frac{1}{4}[\text{var}(X_1 + X_2) - \text{var}(X_1 - X_2)] \quad (6.3.1)$$

and to obtain robust estimates of the variances from trimmed or Winsorized versions of the transformed samples

$$x_{1j} + x_{2j}, \quad x_{1j} - x_{2j} \quad (j = 1, 2, \dots, n).$$

To ensure estimated covariance matrices which are positive definite, Gnanadesikan and Kettenring suggest ranking the multivariate sample in terms of some distance measure  $R(\mathbf{x}; \mathbf{x}^*, I)$  where  $\mathbf{x}^*$  is a robust estimator of  $\mu$ , omitting a small proportion of the sample having the largest  $R(\mathbf{x}; \mathbf{x}^*, I)$  values, and computing a matrix

$$A_0 = \sum' (\mathbf{x}_i - \mathbf{x}^*)(\mathbf{x}_i - \mathbf{x}^*)' \quad (6.3.2)$$

where the summation extends over the *retained* sample members. The whole sample is then ranked in terms of  $R(\mathbf{x}; \mathbf{x}^*, A_0)$ , again a small proportion of the observations having the largest  $R(\mathbf{x}; \mathbf{x}^*, A_0)$  are omitted, and  $V$  is estimated as the appropriate multiple of the matrix of sums of squares and cross-products of the *finally retained* sample members. The procedure is intuitively appealing, but only limited empirical investigation is reported. Another ingenious method of constructing a robust estimator of  $V$ , without recourse to estimation of  $\mu$ , is also presented.

A more specific contribution to outlier-robust estimation relates to estimating the correlation coefficient in a bivariate normal distribution. This is examined by Devlin, Gnanadesikan, and Kettenring (1975). After reviewing various ad hoc estimators based on partitioning of the sample space, on transformations of Kendall's  $\tau$ , or on normal scores, they proceed to investigate (by means of an extensive simulation exercise) estimators based on (6.3.1), on trimming or on Winsorization.

## CHAPTER 7

# *Outliers in Designed Experiments, Regression, and in Time-Series*

We have concentrated so far on the examination of outliers in a single sample of observations from a common distribution, or of outlying samples in a set of independent samples from a common distribution. It has been assumed that the only departure from the null hypothesis of homogeneity of distribution arises in explanation of discordant outliers, or outlying samples, on the basis of one of the models for outlier generation. But in a great deal of statistical analysis departure from the assumption of homogeneity of distribution need have nothing to do with outliers. It is a natural, and often welcome, manifestation of the appropriateness of some linear model explaining how mean values vary with different levels of factors of classification or with different values of a set of independent variables; or it may express a time-dependent effect in the generation of the data. This applies, of course, to the whole range of designed experiments, regression situations, or time-series. Even so, a further degree of inhomogeneity may be revealed by the presence of outliers, which express ad hoc influences additional to linear model or time-series effects. Thus it is appropriate to extend our study of outliers to such more highly structured situations.

A crucial distinction must now be recognized in the occurrence of outliers. In a single univariate sample an outlier was identified subjectively as an observation which engenders 'surprise' in its value relative to the other sample members: it 'sticks out at the end' of the sample. Subjective identification is succeeded by formal processing in the sense of a test of discordancy and (perhaps) a contingent adjustment in the value prior to further processing of the data. For more structured data we would wish to retain the stimulus of 'surprise' but this concept is now far more nebulous. For example, consider the two sets of data shown in Tables 7.1 and 7.2. In Table 7.1 values  $x_i$  and  $y_i$  are respectively the loads (in lb) applied to similar structures and their resulting extensions of length (in inches). We must expect some relationship between the loads and extensions and there is no

Table 7.1 Extension of a structure under different loads

$x_i(1b)$	11.2	21.1	29.9	34.1	43.8	53.4	59.9	61.2	68.9
$y_i(\text{in})$	1.6	2.1	3.4	3.3	4.2	3.1	4.3	6.2	6.3

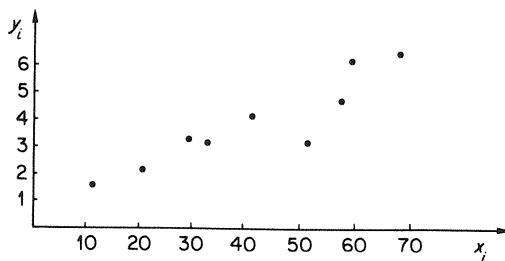
point in seeking outlying extensions of length in terms of the literal  $y_i$  values alone. If we plot the data (as in Figure 7.1) we find confirmation of the relationship. Perhaps a linear regression model is appropriate. But now an anomaly does arise; the observation (53.4, 3.1) shows up as an *outlier* in the sense that it *disturbs the general pattern* to a degree which is discomforting. An appropriate visual inspection still suffices to detect the outlier, but it is clear that in a more complicated regression situation with many independent variables this may well not be possible. Thus we now need to refine our notion of an outlier. It is not a simple extreme value, but it has a more general pattern-disrupting form. Note, however, that compensatory effects can arise for *multiple* outliers, making them less readily detected through simple pattern-disruption. We can no longer rely on direct subjective impact but may need (as in multivariate samples) to adopt an appropriate outlier detection process before we can even contemplate testing discordancy or refining inference procedures to take specific regard of outliers. (This latter consideration is, of course, the stimulus for developing outlier-robust procedures which provide protection against *possible* outliers, rather than being designed to accommodate pre-identified outliers. Trimming or Winsorization exemplify this approach for univariate samples.)

Table 7.2 (extracted from Bross, 1961; earlier presented by Daniel, 1960) presents hypothetical yields of some product at different levels of two chemicals, A and B. (The labels A and B have been interchanged for notational convenience.)

Presumably we should not be particularly surprised in such a situation if the means of the underlying distributions differed at different levels of A or B; indeed we might wish to examine the propriety of some linear model for

Table 7.2 Yields of a process at different levels of two chemicals, A and B

	Levels of B				Total
Levels of A	35	32	37	40	144
	29	29	34	36	128
	25	29	30	20	104
	19	25	25	35	104
	22	20	29	29	100
Total	130	135	155	160	580



**Figure 7.1** Relationship between load and extension for data of Table 7.1

the means. Thus again, the extreme values (19 and 40) in the data set are not the only (or even the predominant) candidates as outliers, and we must employ some basis other than the literal values of the yields for detecting any outliers.

If we consider the first row of the data (corresponding with the first level of A) the observations are uniformly the largest in their respective columns (i.e. at each level of B). It might happen, in the terms of the types of slippage test described in Chapter 5, that this constitutes a discordant outlying sub-sample of the data. But this would be no basis for suspecting the *integrity* of the data; it is just one rather specific manifestation of the type of effect we are investigating in the analysis of such a two-way experimental design. It would point to a significant influence on yield by chemical A, of a rather more specific style than arises from merely rejecting the null hypothesis of no A-effect. Such a 'discordant outlier' result is typical of the *identification* aspect of outliers described in Section 2.2 as the third of the three possible interests (as distinct from *rejection* and *accommodation*). We should wish to acknowledge the outlying sub-sample as a positive result of the overall analysis.

On the other hand the single observation 20 in the last column of the data stands out in *individual isolation* as being relatively more extreme within its column (or row) than do any other observations in the data. It appears to disrupt seriously the *overall pattern* of results where (roughly speaking) yields decrease with the levels of A and increase with the levels of B. We could formalize this impression by considering the *estimated residuals*,  $\tilde{\varepsilon}_{ij}$ , in relation to a fitted additive linear model

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (i = 1, \dots, 5; j = 1, \dots, 4)$$

where  $x_{ij}$  is the yield at levels  $i$  and  $j$  of A and B, respectively,  $\varepsilon_{ij}$  is the corresponding true residual and  $\sum_1^5 \alpha_i = \sum_1^4 \beta_j = 0$ . We have

$$\tilde{\varepsilon}_{ij} = x_{ij} - \bar{x}_i - \bar{x}_{.j} + \bar{x}_{..}$$

(on the usual dot convention for aggregation over the levels of the two factors). The table of estimated residuals appears as follows, highlighting the aberrance of the observation 20:

2	-2	-1	1
0	-1	0	1
2	5	2	-9
-4	1	-3	6
0	-3	2	1

It is this type of individual disruption of pattern that we must regard as an expression of the outlying nature of a *single* observation. We need (at least in complex experiments) to develop procedures for formal detection of such outliers and for proper assessment of their statistical significance (discordancy). Additionally, the accommodation aspect is vital. A prime aim is to examine our linear model in a designed experiment (or estimate and test parameters in a regression, or time-series, model) with as little interference as possible from isolated outlying observations.

In all of the structured models relevant to the different topics of this chapter it is appealing to examine disruption of pattern through the behaviour of *residuals*, and many of the published results on outliers approach the problem in this way. But we must recognize some shortcomings in the use of residuals. They are inevitably inter-correlated (and may even have differing variances by virtue of the assumed model, except perhaps in the null cases of zero regression parameters, no treatment effects or nil time-dependence). Any outliers affect not only their own residuals but have a carry-over influence on others. Thus their aberrance tends to be somewhat smoothed out: they hide behind the skirts of their neighbours in the data set. Extreme examples of this include two-way experiments with one factor having only two levels, where residuals arise in pairs of identical value and opposite sign, or a  $3 \times 3$  Latin square where residuals take equal values in groups of three and inter-residual correlations are either 1 or -0.5. Accordingly other principles for outlier detection, testing, or accommodation are also to be found in the literature and we shall be examining them. Often they involve non-parametric techniques, or informal graphical display procedures.

In principle the study of outliers in regression situations, or in designed experiments, can be subsumed in a wider investigation of outliers in general linear models. Much published work has this wider emphasis, and will be discussed later in the chapter. However, we shall consider first the more specific proposals which have been made for dealing with outliers in designed experiments (usually two-way) and in (principally linear) regression situations. These tend to be more applications-oriented with discussion of the implementation of particular techniques than does the work on the general linear model.

Study of outliers in time-series data is a distinct problem on which little has been published. We review the few known results in a short final section of the chapter.

## 7.1 OUTLIERS IN DESIGNED EXPERIMENTS

Basic problems in the study of outliers in data from designed experiments are that they are difficult to detect (in the respect discussed above) and that their presence influences the analysis of variance of the data set in a way which may cloak significant effects or exhibit apparent effects which, were it not for the outliers, would not arise. Most techniques for outlier detection, for testing their discordancy or for minimizing their influences in analysing the underlying linear additive model, take as a basic measure of the import of individual observations their residuals about the fitted linear model. A concern about the inter-correlation, and carry-over influences, between residuals has prompted some use of modified (outlier-robust) residuals or of measures not directly based on residuals. In all cases the aim is to exhibit, and assess, the extent to which individual observations disrupt the overall pattern anticipated in the data, by virtue of the linear model for the means which is implied by the experimental design.

Bross (1961) presents what he describes as a 'strategic appraisal' of the problems of handling outliers in 'patterned experiments'. He stresses the special difficulties of their detection: an outlier disrupts an anticipated pattern of interrelationship in the data, the pattern is itself a *non-null* representation and needs first to be characterized before it can be 'disrupted'. He develops this theme in terms of isolated departures in the values of observations relative to those of *neighbouring* observations. In describing the influence of outliers he stresses the *combined* effect of the outlier and of the analysis of variance techniques applied to the overall data set. Although Bross proposes no formal set of procedures for coping with outliers he sketches a non-parametric principle which we shall return to in Section 7.1.5.

### 7.1.1 Discordancy tests based on residuals

A natural starting point for studying discordancy tests for outliers using residuals is found in the work of Daniel (1960).

Into a set of artificial data purporting to be the outcomes of a  $4 \times 5$  factorial experiment, he introduces a substantial contamination of a single observation (the data are the same in a reordered form as the data of Table 7.2 from Bross, 1961, briefly examined above). He stresses how the outlier introduces substantial bias in the fitted value and residuals not only in its own specific location but throughout the row and column in which it appears.

To consider the implications of these results let us start with an unreplicated two-way design where the (crossed) factors A and B have  $r$  and  $c$  levels, respectively, and we suspect *no* discordant values in the resulting data. Under the usual additive linear model for the means, with normal error structure, the observations  $x_{ij}$  ( $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$ ) can be written

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (7.1.1)$$

where  $\sum \alpha_i = \sum \beta_j = 0$  and the  $\varepsilon_{ij}$  are independent  $N(0, \sigma^2)$ . The estimated means (fitted values) are

$$\tilde{\mu}_{ij} = \bar{x}_{i\cdot} + \bar{x}_{\cdot j} - \bar{x}_{..} \quad (7.1.2)$$

and the *estimated residuals* are

$$\tilde{\varepsilon}_{ij} = x_{ij} - \tilde{\mu}_{ij}. \quad (7.1.3)$$

The  $\tilde{\varepsilon}_{ij}$  will be  $N(0, \nu\sigma^2/(rc))$  where  $\nu = (r-1)(c-1)$  but they will *not* be independent. The linear constraints on the  $\tilde{\varepsilon}_{ij}$  imply correlation between  $\tilde{\varepsilon}_{ij}$  and  $\tilde{\varepsilon}_{i'j'}$  in the form

$$\rho_{ij i'j'} = \begin{cases} -1/(r-1) & i = i', j \neq j' \\ -1/(c-1) & i \neq i', j = j' \\ 1/\nu & i \neq i', j \neq j'. \end{cases} \quad (7.1.4)$$

Formalizing his numerical example Daniel considers the expected bias in the residuals due to a single outlier at position  $(i, j)$  in the two-way layout reflecting a contamination in the mean value of order  $rc$ . The expected bias will be

$$\delta_{i'j'} = \begin{cases} rc & i' = i, j' = j \\ -r & i' = i, j' \neq j \\ -c & i' \neq i, j' = j \\ 1 & i' \neq i, j' \neq j. \end{cases} \quad (7.1.5)$$

He argues that the correlation between the new (outlier affected) estimated residuals,  $\tilde{\varepsilon}_{ij}$  and their biases will be high, and proposes that the value of this correlation be used as an indication of the presence of the outlier and as a basis for testing its discordancy. The correlation can (after appropriate manipulation) be expressed in the form

$$\tilde{\rho} = \frac{rc}{\nu} \frac{\tilde{\varepsilon}_{\max}^2}{\sum \tilde{\varepsilon}_{ij}^2} \quad (7.1.6)$$

where  $\tilde{\varepsilon}_{\max}^2 = \max_{i,j} \{\tilde{\varepsilon}_{ij}^2\}$ . Thus, in spite of the carry-over effect of the outlier, the indication of its presence resides solely in the *largest* absolute value of the estimated residuals. (This remains true for other designs where the estimated residuals have *common* variance.) To assess if the outlier corresponding with this 'largest residual' is *discordant* we need to compare  $\tilde{\varepsilon}_{\max}^2$

appropriately with the true error variance  $\sigma^2$ . The residual sum of squares  $\sum \tilde{\varepsilon}_{ij}^2$  reflects the influence of the outlier and will be correspondingly inflated (relative to the situation where there is no contamination in the mean at an isolated point) to an unknown extent. Thus to assess discordancy where (typically)  $\sigma^2$  is unknown we need to replace  $\sum \tilde{\varepsilon}_{ij}^2$  by a measure of residual variability which is not influenced by the outlier. To this end it is proposed (implicitly) that we estimate residual variation by removing the outlier, treating it as a missing value and replacing it by the corresponding least-squares estimate. We are led on this argument to a test statistic for discordancy in the form

$$t^2 = \frac{\tilde{\varepsilon}_{\max}^2}{s^2} \quad (7.1.7)$$

where

$$s^2 = \left( \sum \tilde{\varepsilon}_{ij}^2 - \frac{rc}{\nu} \tilde{\varepsilon}_{\max}^2 \right) / (\nu - 1) = S_M / (\nu - 1)$$

is an estimate of  $\sigma^2$  based on the residual sum of squares  $S_M$  in an analysis of variance where the outlier is regarded as a missing observation.

Daniel's arguments concerning the null distribution of  $t^2$  are in error (as remarked in a footnote in Daniel, 1960). However, the assessment of discordancy has been taken up by others, as we shall see shortly. But the general principle is not in dispute. We see that a single outlier is detected as *the observation yielding the largest absolute value among the residuals*; its discordancy is to be assessed in terms of its null (no-outlier) distribution.

Use of the maximum absolute (so-called) studentized residual by Daniel (1960) and also, in the context of accommodation, by Anscombe (1960a) (see Section 7.1.2) was based on intuitive arguments with no overt consideration of the outlier model or of any resulting optimality of the proposed procedures. Noting its form as a natural extension of the optimum single sample procedures of Paulson (1952b), Ferguson (1961a) enquired whether the optimality properties also extended to the designed experiment situation and found this to be so. He adopted a mean-slippage model to explain a single outlier in a general linear model, developed the appropriate rather complex sampling distribution theory and proceeded to investigate the more general linear model formulation in terms of a multiple-decision approach (see Section 2.5). (For the two-way design the model is

$$x_{ij} = \mu + \alpha_i + \beta_j + \sigma a_{ij} + \varepsilon_{ij} \quad (7.1.8)$$

where  $a_{ij} \neq 0$  for precisely one pair of values  $(i, j)$ , and is zero otherwise.)

Subject to assumptions that no estimated residuals have zero variance, and no two estimated residuals have unit correlation, the outlier test based on maximum absolute studentized residual proves to be optimal in the sense of being *invariant admissible*. In the case of designs where all estimated residuals have equal variance (true of the unreplicated two-way design and

many others including all equally replicated ordinary factorial designs, Latin squares, and balanced incomplete blocks—see Anscombe, 1960a) the optimality property can be reexpressed in the terms that the outlier test is a Bayes solution in respect of a uniform prior distribution over the set of hypotheses specifying *equal* shifts in the mean for the outlier.

We must note how the work by Anscombe and by Ferguson extends the range of applicability and propriety of Daniel's proposal—to a wider set of designed experiments and to many general linear model problems. We return to this latter point in Section 7.3.

Later study of this test centres on the determination of critical values for attributing discordancy, and on placing it in the perspective of alternative proposals that have been made for examining the assumptions underlying the analysis of variance (for example, additivity, normality, and homoscedasticity, in addition to non-contamination by discordant outliers).

Let us remind ourselves of the basic nature of the Daniel test. It uses as test statistic the ratio of the maximum squared (estimated) residual to the residual sum of squares in a least-squares analysis which regards as *missing* the observation yielding the maximum squared residual. For a general model it can be expressed as

$$t^2 = \tilde{\varepsilon}_{\max}^2 / [S_M / (l - 1)] \quad (7.1.9)$$

where  $l$  ( $< n$ ) is the number of degrees of freedom associated with the residual sum of squares  $S_C$  when the model is fitted to the complete data set of  $n$  observations;  $S_M$  is the corresponding residual sum of squares when the observation yielding  $\tilde{\varepsilon}_{\max}^2$  is regarded as missing.

Related statistics have been proposed by others, as John and Prescott (1975) point out. Over 25 years ago, Quenouille (1953) suggested that we investigate an outlier in a designed experiment by considering a statistic which can be formally expressed as

$$\tau_1^2 = l(S_C - S_M) / S_C \quad (7.1.10)$$

with obvious intuitive appeal. Recognizing the fact that the missing observation is not arbitrary, but corresponds with the largest (absolute) residual, he suggested that the critical level of  $\tau_1^2$  is assessed as

$$nP(F_{1,l} > \tau_1^2) \quad (7.1.11)$$

where  $F_{1,l}$  has an  $F$ -distribution with 1 and  $l$  degrees of freedom.

In an interesting paper using a simulation method (Goldsmith and Boddy, 1973; discussed in more detail below) the authors suggested using the statistic

$$\tau_2^2 = (l - 1)(S_C - S_M) / S_M \quad (7.1.12)$$

and, supported by their simulation study, proposed assessing its critical level as

$$1.25(l - 1)P(F_{1,l-1} > \tau_2^2). \quad (7.1.13)$$

Daniel (1960) originally suggested that the critical level of  $t^2$  be assessed as

$$P(F_{l,l-1} > nt^2/l^2) \quad (7.1.14)$$

but he pointed out a flaw in his argument (with unexamined implications).

Simulation studies by John and Prescott (1975) suggest that all the three distributional proposals, (7.1.11), (7.1.13), and (7.1.14) are, in general, unsatisfactory. Incidentally, the three statistics  $t^2$ ,  $\tau_1^2$ , and  $\tau_2^2$  are all equivalent, and amount to representing the position and import of a single outlier in terms merely of the largest (absolute) residual. (This is important to recognize in relation to the work of Goldsmith and Boddy, 1973, discussed further below.) Specifically we have

$$\tau_2^2 = \frac{n}{l} t^2 = (l-1) \left[ \frac{l}{\tau_1^2} - 1 \right]^{-1}. \quad (7.1.15)$$

John and Prescott suggest that the critical level of  $\tau_2^2$  should be determined as

$$nP(F_{1,l-1} > \tau_2^2) \quad (7.1.16)$$

and show by simulation how the Quenouille and Daniel proposals are ‘far too conservative’, those of Goldsmith and Boddy too liberal, whilst their own suggestion appears to be quite accurate for a range of factorial designs with factors at two or three levels. They assess the accuracy of their simulation results by comparison with exact critical values determined by Stefansky (1972), for a range of designs:  $3^2$ ,  $4^2$ ,  $4 \times 3$ ,  $5 \times 3$ ,  $6 \times 4$  and  $8 \times 7$ .

Although the John and Prescott proposal is simple and appears reasonable in many cases, the most accurate, and useful, results to date on the null distribution of the test statistic (and hence the best prescription for application of the outlier test of discordancy) derive from the work of Stefansky (1971, 1972).

Stefansky re-expresses  $t^2$  in terms of the *maximum normed residual* (MNR). If the estimated residuals are  $\tilde{\varepsilon}_i$  ( $i = 1, 2, \dots, n$ ), the *normed residuals* are

$$z_i = \tilde{\varepsilon}_i / \sqrt{\sum_1^n \tilde{\varepsilon}_i^2} \quad (7.1.17)$$

and the MNR, denoted  $|z|_{(n)}$ , is the largest of the absolute values  $|z_i|$  ( $i = 1, 2, \dots, n$ ).

For the two-way design, the statistic (7.1.7) of Daniel can be expressed as

$$t^2 = (\nu - 1)(|z|_{(n)})^2 / [1 - n(|z|_{(n)})^2 / \nu], \quad (7.1.18)$$

a strictly increasing function of  $|z|_{(n)}$ . Thus we can conduct the test in terms of the equivalent test statistic  $|z|_{(n)}$ , or some simple function of it, provided its null distribution is known. Stefansky addresses herself to determining the null distribution.

The equivalence of  $t^2$  and  $|z|_{(n)}$  holds for designs where the residuals (in the absence of a discordant outlier) have equal variances, and the following results hold in such situations. The relationship (7.1.18) can be written, in the wider context, as

$$t^2 = (l-1)(|z|_{(n)})^2 / \{1 - n(|z|_{(n)})^2/l\} \quad (7.1.19)$$

where  $l$  is as defined below (7.1.9).

Stefansky (1971) extends an earlier observation of Pearson and Chandra Sekar (1936) that critical values of statistics based on the MNR in the single sample case could be calculated exactly from tables of the  $t$ -distribution for sufficiently large values of the statistic provided we know the largest value that can be taken by the *second largest* of the absolute values of the normed residuals. Specifically, quantities

$$F_i = n(l-1)z_i^2/(l-nz_i^2) \quad (7.1.20)$$

are considered. Then  $F_{(n)} = \max\{F_i\} = nt^2/l = \tau_2^2$ .

The method uses Bonferroni-type inequalities (see Section 5.3) to provide lower and upper bounds for the critical values of the MNR (or related quantities). Different 'orders' of inequality are available; the higher the order the sharper the bounds but the more complex the calculations. However, the crucial result is that sufficiently far out in the tail of the distribution *exact* critical values are obtained. The stage at which this facility holds depends on the values of quantities  $M_k$ : the greatest obtainable values of the  $k$ th-largest  $|z_i|$ . Stefansky (1971) shows how to determine  $M_k$  for the range of designs with homoscedastic residuals as discussed above. In Stefansky (1972) it is shown that  $P(|z|_{(n)} > z)$  can be determined *exactly* from the  $r$ th Bonferroni upper (lower) bound if  $z > M_{2r}$  ( $M_{2r+1}$ ).

It is demonstrated that for the unreplicated two-way design application of this principle to the first-order inequalities is of little practical interest since it requires  $z$  to be unreasonably large. However, the second-order inequalities yield exact critical values from about the upper 10 per cent point of the distribution, for two-way designs with up to about nine levels for each factor.

Assessment of the precise range over which exact results can be obtained depends of course on knowing  $M_2$ ,  $M_3$ , and  $M_4$  etc. For the  $r \times c$  design with one observation per cell we have

$$M_2 = [m(M-1)/2]^{\frac{1}{2}}$$

where  $m = \min(r, c)$ ,  $M = \max(r, c)$ ,

$$M_3 = [(r-1)(c-1)/(3rc-2r-2c-2)]^{\frac{1}{2}},$$

$$M_4 = 0.5.$$

Correspondingly, exact upper 1 per cent and 5 per cent points of  $|z|_{(n)}$  are obtained for  $r = 3(1)9$ ,  $c = 3(1)9$  and these are presented as Table XXXI on page 334 (extracted from Stefansky, 1972).

Multiway designs can be collapsed to two-way structure by sacrificing access to certain interactions. Apart from this prospect, Stefansky's tabulated values are restricted to two-way designs. The best prescription for factorial designs with more than two factors appears to be that of John and Prescott (1975) although we must bear in mind the limited range of their simulation studies which encompass only small designs (e.g.  $2^3$ ,  $2 \times 3^3$ ,  $2^4$ , etc.). Even in this range, the rather poor performance of their critical level proposal when main effects and first order interactions are fitted in the  $2 \times 3^2$  design sounds a warning note.

Some interesting practical examples of residual-based methods of testing outliers in designed experiments are given by Goldsmith and Boddy (1973). 23 sets of data, previously discussed in the statistical literature, are re-analysed to detect and test outliers. A 'consecutive' style of analysis is adopted, proceeding from a discordant outlier to the search for further outliers. Computer-based application of the technique is discussed in some detail in relation to any orthogonal design (admitting possible missing values). The authors express dissatisfaction with methods based on the largest residual in that its determination on the basis of fitting the model to the *whole* data set (including any outlier) will produce a test of relatively low power. Their 'alternative' proposal is to regard each observation in turn as missing and to scan the set of  $n$  residual sums of squares to see if one of them is noticeably smaller than the others, indicating an outlier. Concentrating on the minimum residual sum of squares, they employ as test statistic the quantity  $\tau_2^2$  of (7.1.12) above. However, we have noted that this is equivalent to other proposals based on the largest (absolute) residual, so that it is not clear just how a re-emphasis, or broadening of approach, is manifest in this work of Goldsmith and Boddy (1973).

*Example 7.1.* To illustrate some of the above ideas we consider a set of data from an unreplicated  $2^3$  experiment attributed by Goldsmith and Boddy (1973) to an unpublished lecture by C. Daniel in 1960. The first three columns in Table 7.3 below describe (in common notation) the eight treatment combinations, the corresponding yields, and the treatment effect totals respectively.

All the effects are of similar order of magnitude and any main-effects analysis will not show up significant treatment effects. There are no immediately obvious outliers in the data. But perhaps outliers are present, masking genuinely significant treatment effects. According to Goldsmith and Boddy (1973), Daniel argued that the largeness of all the interaction terms was in itself suspicious and might indicate one discordant outlier: specifically the yield at 'a', in view of the pattern of signs  $(-1, -1, +1, +1)$  of the interaction effects. Accordingly he estimated the outlier discrepancy as the mean of the absolute values of the interaction effects (64) and produced 'amended effects' as shown in column 4, in which the interaction terms are now low in comparison with main effects.

Table 7.3

(1) Treatment combination	(2) Yield	(3) Effect total	(4) 'Amended effects'	(5) Residuals	(6) 'New residuals'	(7) Residual Mean Square (corre- sponding yield missing)
(1)	121	1129	1065	-15.25	0.75	562.5
a	145	-93	-157	32.00	0	34.8
b	150	53	117	0.50	0.50	717.3
ab	109	-45	19	-17.25	-1.25	519.1
c	160	79	143	4.00	4.00	706.8
ac	112	-59	5	-20.75	-4.75	430.5
bc	180	67	3	10.75	-5.25	640.5
abc	152	85	21	6.00	6.00	693.5

But let us look at the residuals after fitting a main effects model to the original data. These have values as shown in column 5, with corresponding residual sum of squares, 2152.5. The residual at the treatment combination a is indeed the largest, although those at (1), ab and ac are also large. Are the residuals random, or is there contamination at the treatment combination a transmitted (via the correlation between the residuals in the same row or column of the A × B × C design) as larger residuals at (1), ab, ac, (and abc)? If we regard the yield at a as missing, its least-squares estimate is 81 (vastly different from 145, by precisely the discrepancy 64 proposed by Daniel). The new residuals are shown in column 6; the residual sum of squares is reduced from 2152.5 (on four degrees of freedom) to 104.5 (on three degrees of freedom).

The outlier test statistic value is

$$\tau_2^2 = \frac{3 \times 2048}{104.5} = 58.79$$

with critical level (on the John and Prescott proposal) 0.04. Thus at the 5 per cent level we would reject the a-yield of 145 as a discordant outlier.

A major difficulty in handling outliers in designed experiments lies in their initial detection. The residual 32.00 in column (5) hardly renders the a-yield a compelling candidate. Goldsmith and Boddy (1973) suggest that clearer initial detection may result from scanning the values of the residual mean squares when each yield in turn is regarded as a missing value. This requires a lot of calculation if the number of observations is at all large. However, for the current data, the residual mean squares (shown in column 7) really do highlight the outlier; all but the one corresponding with treatment combination a are of similar order to the original (full data) mean square 538.1; the mean square when a is missing is dramatically smaller at 34.8.

John (1978) gives two detailed practical examples of applying residual-based outlier methods to data from designed experiments. One example is a 1/3 replicate of a 3<sup>4</sup>, the other a confounded 2<sup>5</sup>.

### 7.1.2 Residual-based accommodation procedures

We discussed earlier (Chapter 4) the range of techniques available for accommodating outliers in robust analyses of univariate samples. Included in the discussion was the premium-protection approach of Anscombe (1960a), where in estimating the mean of the underlying distribution a location-slippage outlier model is employed and the mean estimated either by the overall sample mean for the sample of size  $n$ , or by the sample mean of  $(n - 1)$  observations omitting an extreme sample value if it was sufficiently large (or small) in value. The outlier is thus either rejected, or retained with full weight, in the estimation process (cf. partial retention employed in trimming, Winsorization, or more general differential weighting procedures). The procedure is assessed in terms of two criteria: the *premium* paid in terms of increase of variance (or some other measure of expected loss) of the estimator when the sample comes in fact from a homogeneous source, and the *protection* provided in terms of decrease of variance (or mean square error) when a discordant observation is present.

Anscombe (1960a) describes how the same approach can be used in designed experiments (or in analysing general linear models). Attention is restricted to situations where, in the absence of discordant values, all residuals have common variance,  $l\sigma^2/n$ , in the notation of Section 7.1.1, and where inter-residual correlation is nowhere 1 (or  $-1$ ).

Only the case where  $\sigma$  is known is considered in detail. If  $\tilde{\varepsilon}_{\max}$  is again the estimated residual having greatest absolute value it is proposed that: *if  $|\tilde{\varepsilon}_{\max}| > h\sigma$ , we reject the observation yielding  $\tilde{\varepsilon}_{\max}$ , treat it as a missing value, and estimate the unknown parameters (means) by a least-squares analysis; if  $|\tilde{\varepsilon}_{\max}| \leq h\sigma$  we retain all observations and conduct a full least-squares analysis.*

The constant  $h$  needs to be chosen to produce acceptable premium and protection guarantees. Anscombe shows that the proposed rule has a simple interpretation. If  $|\tilde{\varepsilon}_{\max}| > h\sigma$ , we merely replace  $x_m$  (the observation yielding  $\tilde{\varepsilon}_{\max}$ ) by  $x_m - n\tilde{\varepsilon}_{\max}/l$  in the estimating equations for the means. Determination of premium and of protection are less straightforward than in the single sample case. We are concerned with the effect of the rule on the variances of estimates. But we have a number of parameters of interest. One possibility is to consider the way in which application of the rule affects the determinant of the corresponding variance matrix (in the absence of discordant observations). Anscombe argues that an appropriate notion of premium is in terms of that proportional increase in  $\sigma^2$  which would increase the variance-matrix determinant by as much as the proposed rule does, in the absence of

discordant observations. He shows how, on this definition, the premium can be approximately determined.

An approximation to the *protection* provided by the rule, when one of the observations is discordant and has bias  $\beta\sigma$  in the mean, is also given by Anscombe. He reviews the numerical properties of the rule by tabulating values of the protection, of the cut-off level  $h$  and of the probability of inappropriate rejection of  $x_m$  as discordant for premium levels of 2 per cent and 1 per cent, and a range of values of  $l/n$ .

There is much that requires further investigation in this approach: the accuracy of the various approximate results, the robustness (or form of necessary modifications) of the procedure in the face of non-normality, non-additivity, etc., it is of paramount concern to develop a feel for appropriate levels of, and balances between, the premium and protection elements. Anscombe deals briefly with two other matters: multiple outliers and unknown  $\sigma$ . For the former case he suggests that the rule is applied consecutively to successive smaller samples until we reach the stage that no further rejection of observations takes place. For the (commonly encountered) situation where  $\sigma$  is unknown he gives some approximate results for the equivalent 'studentized' rule where  $\sigma^2$  is merely replaced by  $s^2$ , an estimate of  $\sigma^2$  based on  $l+l_0$  degrees of freedom ( $l_0$  corresponding with additional external or prior information about  $\sigma^2$ ). The results are not pursued to the level of useful application, and Anscombe conjectures that the rule will have 'low power' unless  $l+l_0$  is reasonably large (say, 30 or so). Some further observations are given by Anscombe and Tukey (1963) in the context of a wider discussion of the analysis of residuals in designed experiments and general linear models. See also Anscombe (1961), and the discussion of the general linear model in Section 7.3.

### 7.1.3 Graphical methods

A variety of graphical procedures have been proposed for investigating the validity of the various assumptions underlying the analysis of variance of data arising from a designed experiment. Sometimes these procedures are aimed specifically at exhibiting, or examining, outliers. More often other assumptions such as normality, additivity, or homoscedasticity are under investigation, but the performances of the relevant procedures are sensitive also to the presence of outliers, and *en passant* provide indications of outlying behaviour of data points. Frequently it is difficult to distinguish which specific departure from the underlying assumptions is manifest in an unacceptable graphical plot; sometimes the procedure is more sensitive to one departure than to another. Most of the procedures regard residuals as the natural reflection of the impropriety of the assumptions, although non-residual-based methods have also been advanced. A general study of the assumptions underlying an analysis of variance is beyond our brief, but we

will consider some of the methods with greatest relevance to the outlier issue.

Again we can conveniently commence with some work by Daniel (1959) on *half-normal plots*. Daniel considers factorial experiments of the form  $2^p$ :  $p$  factors each at two levels. He considers the *ordered* absolute values of the effect totals and remarks that in the absence of any real effects these are observations of the order statistics from a known distribution. If the error distribution is normal, the absolute effect totals will behave as independent half-normal deviates with common variance. That is, their probability density function has the form

$$f(x) = \sqrt{(2/\pi\sigma^2)} \exp[-x^2/2\sigma^2] \quad (x \geq 0). \quad (7.1.21)$$

Plotting the ordered values on appropriately constructed probability paper will produce, in the null case, observations lying close to a straight line through the origin. Departures from linearity will indicate real effects (large values lying off the straight line) or violations of the basic assumptions of the model. In particular the presence of a single outlier will similarly inflate the absolute values of all effect totals. It will show up, therefore, by the probability plot (although possibly linear apart from a few high values indicating real effects) *being not directed towards the origin* but towards a value similar to the contamination bias of the outlier. Note that we become aware of the outlier in this way; but we do not determine the offending individual observation. With more than two outliers the plot does not necessarily reveal their presence in a very dramatic form, unless the biases happen to be of the same sign and similar magnitude. We must also recognize that other aberrances such as non-normality, non-constant error variances, etc. can affect the linearity of the plot and might do so, particularly in small data sets, in a way which is indistinguishable from the manifestation of an outlier. But this problem is not restricted to this particular plotting method.

See also Birnbaum (1959) on half-normal plotting methods.

Within the context of another wide-ranging study of how to assess the validity of assumptions underlying analysis of variance, Anscombe and Tukey (1963) consider graphical display of *residuals*, including probability plots and plots against fitted values or external concomitant variables. They are at pains to emphasize the overlap of influences and indications, remarking that apparent non-normality, non-additivity, different error variances, and isolated discordant values can all show up in similar (and indistinguishable) ways.

One simple graphical presentation is obtained by plotting the ordered values of the residuals on normal probability paper. With normal error structure we expect a straight line relationship. Non-linearity will be indicative of skewness or flatness of the error distribution; outliers will be manifest in marked isolated departures at either end of the plot. Such a procedure

(termed FUNOP—full normal plot) has its limitations, of course, arising from the intercorrelation of residuals, or non-attributability of effects, for the detection of outliers. FUNOP was introduced by Tukey (1962), who also proposed a procedure entitled FUNOR-FUNOM for compressing the values of larger residuals, in contrast to reducing them to zero as is implied in the outright rejection, as outliers, of the corresponding observations.

One advantage of this simple approach is that it considers extreme values of residuals, not merely extreme *absolute* values which were emphasized in Sections 7.1.1, 7.1.2. As we shall consider later (Section 7.3), in relation to work by Andrews (1971), it can be important to consider the values of residuals relative to the design matrix, rather than merely in absolute terms.

Gnanadesikan and Kettenring (1972) also consider the self-camouflaging effect of outliers, due to their influence carrying over to other residuals and thus making their detection problematical. A solution is suggested in the use of 'modified residuals', based on outlier-robust fitted values (e.g. estimating means by medians, trimmed means, etc.) rather than full-sample least-squares estimates. It is suggested that probability plots of such ordered *modified* residuals may be more informative about outliers.

Probability plotting methods are also used to augment other methods of studying outliers in the battery of procedures described by Gentleman and Wilk (1975a, 1975b); Gentleman and Wilk (1975b) re-examine full- and half-normal plots of residuals and confirm their usefulness in detecting an outlier when a single discordant value is present in a two-way design. They demonstrate the confusion that can arise from compensating effects when there are two or more discordant outliers, and suggest that the probability plots have little value in such cases. Regarding the distribution of the residuals they show, in terms of the Shapiro and Wilk (1965)  $W$ -test for normality, that the intercorrelation of the residuals is not exhibited in apparent non-normality. Indeed, in terms of  $W$ , the residuals on the null additive model can exhibit a degree of 'super-normality'. This remains so for certain small configurations (e.g.  $2 \times 3$ ,  $3 \times 4$ ) even if a discordant value is present although, as Chen (1971) shows,  $W$  is in general sensitive to the presence of outliers. Prescott (1976) compares the effect of outliers on the Shapiro-Wilk  $W$ -test and on an entropy based test of normality, by means of *sensitivity contours*.

#### 7.1.4 Non-residual-based methods

One approach to the detection of outliers in linear models (principally two-way designs) which is not based exclusively on examining residuals about the no-outlier model is that of Gentleman and Wilk (1975a). They present a method for detecting the ' $k$  most likely outlier subset' as that set of  $k$  observations warranting attention as outliers prior to a further examination of their discordancy, or to an attempt to analyse the data in a way

which is (relatively) insensitive to their presence. The procedure consists of specifying  $k$ , determining the ' $k$  most likely outlier subset', assessing its statistical significance and, if not significant, proceeding to consider successively smaller numbers,  $k-1, k-2, \dots$  of outliers until a significant outlier subset is detected (if at all).

The method involves a deal of computational effort and needs to be conducted on a computer. It proceeds as follows.

For given  $k$  we consider all  $\binom{n}{k}$  partitions of the data set obtained on specifying particular subsets of  $k$  observations. If there are truly  $k$  discordant values present, then  $\binom{n-k}{k}$  of these subsets will *not* contain discordant values,  $\binom{n}{k} - \binom{n-k}{k}$  will do so. We suppose (in general linear model terms) that the observation vector  $\mathbf{x}$  has the form

$$\mathbf{x} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon} \quad (7.1.22)$$

where  $\mathbf{A}$  is a  $n \times q$  design matrix,  $\boldsymbol{\theta}$  a vector of  $q$  parameters,  $\boldsymbol{\varepsilon}$  the error vector and  $\boldsymbol{\delta}$  a  $n \times 1$  vector with  $n-k$  zeros and  $k$  unknown non-zero values corresponding with the  $k$  mean biases of the discordant observations.

Suppose that  $\tilde{\boldsymbol{\varepsilon}}$ ,  $\hat{\boldsymbol{\varepsilon}}$  are the estimated residuals obtained by fitting the null model  $\mathbf{x} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$  and by fitting the outlier model (7.1.22), respectively, when a particular set of  $k$  observations are under consideration. Then the difference in the sums of squares

$$Q_k = \tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}} - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \quad (7.1.23)$$

(which is inevitably non-negative) provides a measure of the effect of assuming that the  $k$  chosen observations are in fact discordant. (Note that  $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$  is also the sum of squares of residuals arising from fitting the null model to the reduced data set of  $n-k$  observations.)

If we were to evaluate the  $Q_k$  for all  $\binom{n}{k}$  partitions of the data we could examine their relative sizes as indications of the prospect of the corresponding  $k$  obsevations being outliers. The *largest*  $Q_k$  is used to detect the  *$k$  most likely outlier subset*. To determine if it is large enough not to have arisen purely by chance under the null model we would really need to know its null distribution. This distribution is unknown, and informal methods are proposed to assess the significance of the largest  $Q_k$ . These include the plotting of some large subset of the larger  $Q_k$  against 'typical values' of these obtained (presumably by simulation) under the null model, or by plotting the residuals on the outlier model (7.1.22) corresponding with the set of 'outliers' detected by the largest  $Q_k$ . The first of these leans very heavily on

the assumptions of the model and presumably will be influenced by non-normality, heteroscedasticity, and non-additivity, as well as by discordant observations.

If the ' $k$  most likely outlier subset' is not statistically significant, we go through the same exercise with  $k$  reduced to  $k - 1$ , and so on, until we detect a significant outlier subset.

The whole approach appears cumbersome, although some computational simplification is possible, e.g. for two-way designs, see Gentleman and Wilk (1975a). It is interesting in principle, but clearly needs much examination and refinement if it is to be a practical proposition. The philosophy behind starting with a specific  $k$  and then considering *smaller* values is that if we obtain significance at any stage we might argue that there is little point in considering even smaller  $k$  (even fewer outliers); the reverse policy does not have this apparent advantage. However, we must take care that 'swamping' does not occur: the phenomenon of a non-outlier being included fortuitously along with an outlying subgroup (see Fieller, 1976). This approach does help, though, to reduce the opposite phenomenon of 'masking', provided  $k$  is initially chosen sufficiently (but not unreasonably) large.

See also John (1978) and John and Draper (1978).

### 7.1.5 Non-parametric, and Bayesian, methods

Some non-parametric methods have been proposed for detecting outliers in designed experiments. We have already referred to the general proposals by Bross (1961), who seeks to give expression to the idea of outliers as disrupters of anticipated pattern in the data. He sketches out a non-parametric approach in which we detect discordant outliers in terms of pairwise inversions of the observations relative to the anticipated pattern using (*inter alia*) a *sequence sign test*. Suppose that in a two-way design we expect real effects to show up in a monotone change in the means within each row and each column. We can look for the reflection of such a relationship in the actual data: anomalous inversions in the values of successive observations within a row or a column may indicate outliers. Appropriate test statistics can be constructed in terms of accumulated numbers of inversions. But there are problems with this approach, arising from the hierarchy of models we have to consider and the intangibility of 'anticipated pattern'. We do not know at the outset what sort of pattern to expect for the means as a reflection of the additive model for the means. We do not even know if the data support an additive model, or if apparent non-additivity reflects interactions or isolated discordant values. More fundamentally, the data may be just a random sample of observations from a common basic distribution. That is to say there are no real effects. This is the null-model in terms of which we wish to conduct the analysis of variance. But if there are no real effects we have no structured pattern

against which to detect outliers; inversions, for example, will be irrelevant in this context. Of course a single-sample test of discordancy would be appropriate if there are no real effects but we have no way of knowing if effects are present or not. This uncertainty is the stimulus for studying the additive model; we want such study to be safeguarded from outliers—we find ourselves once more in a vicious circle of conflicting aims and indications.

Other non-parametric methods of a more detailed form (but with similar conceptual difficulties) have been proposed. Brown (1975) develops an approximate  $\chi^2$  test of discordancy for outliers based on the signs of the estimated residuals in the rows and columns of the data for a two-way design, and considers its extension to more complicated designs. For the two-way design he proposes the statistic

$$c^{-1} \sum_{i=1}^r R_i^2 + r^{-1} \sum_{j=1}^c C_j^2 - (rc)^{-1} T^2 \quad (7.1.24)$$

where  $R_i$  and  $C_j$  are the sums of the signs of the residuals in the  $i$ th row and  $j$ th column, respectively, and  $T$  the overall sum of the signs of the residuals. This is sensitive to the presence of outliers and will have a no-outlier distribution which is approximately a multiple  $(1 - 2/\pi)$  of  $\chi^2$  with  $r + c - 1$  degrees of freedom. Unfortunately the method does not pinpoint the outlier; also a significant result *could* arise from non-null manifestations other than a single discordant value.

The ‘extreme rank sum test for outliers’ of Thompson and Willke (1963) is not a test of discordancy for individual outliers; it is a non-parametric slippage test for outlying rows or columns in a two-way design. See Chapter 5.

A Bayesian approach to handling outliers in general linear models is presented by Box and Tiao (1968). It is not specifically directed towards designed experiments; some details are given in Section 8.1.2.

## 7.2 OUTLIERS IN REGRESSION

As with designed experiments, so a study of outliers in regression situations can be regarded as a particular case of the study of outliers in general linear models. However, there are practical advantages in examining the regression situation *per se*, and in methodological terms it represents a further step towards the general case. But the boundary becomes somewhat blurred; often regression problems are considered in the literature as illustrative examples in wider investigations and the present section thus tends to overlap with the later discussion (Section 7.3) of the general linear model.

### 7.2.1 Outliers in linear regression

The simplest case is where we have observations  $x_j$  of independent random variables  $X_j$  ( $j = 1, 2, \dots, n$ ) whose means depend linearly on

predetermined values  $u_i$  of a variable  $U$ . Thus

$$x_i = \theta_0 + \theta_1 u_i + \varepsilon_i \quad (7.2.1)$$

where the  $\varepsilon_i$  are independent with zero mean. Usually the  $\varepsilon_i$  are assumed to come from a common distribution; more specifically we might assume  $\varepsilon_i \sim N(0, \sigma^2)$ . We shall consider the detection, testing, and accommodation of outliers in this situation. If  $\theta_0$  and  $\theta_1$  are estimated by least-squares (or equivalently by maximum likelihood on the *normal* error model) as  $\tilde{\theta}_0$  and  $\tilde{\theta}_1$  we can estimate the residuals  $\varepsilon_i$  as

$$\tilde{\varepsilon}_i = x_i - \tilde{\theta}_0 - \tilde{\theta}_1 u_i \quad (7.2.2)$$

and in seeking outliers it is again sensible to examine the relative sizes of the  $\tilde{\varepsilon}_i$ .

In the discussion of designed experiments we restricted attention to designs where the estimated residuals had equal variance. Even for the simple linear regression model (7.2.1) however, we lose this simplifying feature since

$$\text{var}(\tilde{\varepsilon}_i) = \sigma^2 \left\{ \frac{n-1}{n} - (u_i - \bar{u})^2 / \sum_1^n (u_i - \bar{u})^2 \right\}. \quad (7.2.3)$$

Thus the  $\tilde{\varepsilon}_i$  are more variable the closer  $u_i$  is to  $\bar{u}$ . This ‘ballooning’ effect of the residuals (Behnken and Draper, 1972) needs to be taken into account, if it is at all marked, in examining the size of the residuals  $\tilde{\varepsilon}_i$  as a reflection of outliers. The effect is not restricted to the simple linear regression model (7.2.1).

For the general linear model

$$\mathbf{x} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad (7.2.4)$$

we have (in the full rank case)

$$\text{var}(\tilde{\boldsymbol{\varepsilon}}) = \sigma^2(I_n - \mathbf{R}) \quad (7.2.5)$$

with

$$\mathbf{R} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$$

so that we will need later to consider the implications for outlier detection of the extent to which the form of the design matrix  $A$  induces inhomogeneity of variance in the estimated residuals.

Returning to the simple model (7.2.1), one possible approach is to examine appropriately weighted estimated residuals

$$\frac{\tilde{\varepsilon}_i}{s_j} = \tilde{\varepsilon}_i / \left\{ s \sqrt{\left( \frac{n-1}{n} - (u_i - \bar{u})^2 / \sum_1^n (u_i - \bar{u})^2 \right)} \right\} \quad (7.2.6)$$

where  $s^2 = \sum \tilde{\varepsilon}_i^2 / (n-2)$  is the unbiased estimate of  $\sigma^2$ , and  $s_j^2$  is an unbiased estimate of  $\text{var}(\tilde{\varepsilon}_i)$ . Corresponding with the earlier use (Section 7.1.1) of the

maximum absolute studentized residual  $\tilde{\epsilon}_{\max}/s$ , we might now seek to detect and test the discordancy of a single outlier in terms of the statistic

$$t = \max |\tilde{\epsilon}_j/s_j|. \quad (7.2.7)$$

If  $t$  is sufficiently large, we adjudge the observation yielding  $\max |\tilde{\epsilon}_j/s_j|$  to be a discordant outlier. Note the implication of this policy. It is no longer necessarily the most extreme residual which is the prime candidate for designation as an outlier. Even a modest residual, corresponding with a small variance (7.2.3), can be promoted into an outlying role.

To conduct the test we need to know the distribution of  $t$ ; again its exact form is intractable but much work has been published on approximate forms for the distribution or its percentage points, based on Bonferroni inequalities or large scale simulation studies.

In the first category is the discussion by Srikantan (1961) of tests for outliers in the general multilinear regression model. He proposes use of a test statistic which particularizes (in the case of the simple model (7.2.1)) to  $t^2$ , where  $t$  is given by (7.2.7), and also considers the corresponding statistics  $\max\{\tilde{\epsilon}_j|\tilde{\epsilon}_j|/s_j\}$  and  $\max\{-\tilde{\epsilon}_j|\tilde{\epsilon}_j|/s_j\}$  for one-sided tests (where the outlier model contemplates slippage in a single mean). He uses the first Bonferroni inequality to derive upper bounds for the 5 per cent and 1 per cent critical values of the three test statistics, based on the  $F$ -distribution. Of particular note is his demonstration that the upper bounds are the *exact* percentage points for reasonably small samples (this predates and generalizes Stefansky, 1971 and 1972, who considers only models where the  $\tilde{\epsilon}_i$  have common variance, although Stefansky's use of higher-order inequalities extends the exact results to larger sample sizes).

Implicit in Srikantan's result about exact critical values is that the *extreme tail behaviour* of the distribution of  $t^2$  (and of the one-sided equivalents) will be independent of the configuration of values of the independent variables in the multilinear model. What constitutes the 'extreme tail' will depend, however, on this configuration which will thus determine whether the conventional significance levels (5 per cent, 1 per cent say) are included in the 'extreme tail'. Tietjen, Moore, and Beckman (1973) conduct a large-scale simulation to obtain approximate critical values for  $t$ . They reconfirm *empirically* the known relative insensitivity of the critical values to the configuration of the  $u_i$  for reasonably small  $n$ , and conclude that one can employ the tabulated critical values of the single-sample statistic  $(x_{(n)} - \bar{x})/s$  due to Grubbs (1950) (Thompson, 1935), with appropriate minor modifications to allow for its one-sided form and the discrepancy of 1 in the degrees of freedom associated with  $s^2$ . But caution is needed in three respects. The one- and two-sided distinctions imply the use of a further Bonferroni inequality so that at best the comparison provides only upper bounds for the critical values; for larger sample sizes the correspondence will deteriorate (in the light of the discussion above on the effect of the configuration of the

independent variable values); and the suggested modification to deal with the different degrees of freedom appears to be incorrectly (or at least ambiguously) described.

Prescott (1975b) takes up the insensitivity of the critical values of  $t$  to the  $u$ -values of  $t$  in another respect. He suggests that we ignore the differing variances of the  $\tilde{\epsilon}_i$  and replace  $s_i$  in (7.2.7) by  $\bar{s}$  where  $\bar{s}^2$  is the 'average variance'  $\sum_i \tilde{\epsilon}_i^2/n$  introduced by Behnken and Draper (1972). (7.2.7) then reduces to a multiple of the MNR: viz.  $n^{1/2} \max |\tilde{\epsilon}_i|/\sqrt{(\sum_i \tilde{\epsilon}_i^2)}$ , and approximate critical values are obtained (invoking Stefansky 1971, 1972) on the assumption that the estimated residuals have common variance equal to the population 'average variance'  $(n-2)\sigma^2/n$ .

Although formal distinctions exist in the principles invoked by Srikantan (1961), Tietjen, Moore, and Beckman (1973) and by Prescott (1975b) their tabulated critical values differ little for practical purposes when appropriately compared. However, perhaps the most useful tabulation of critical values for the simple linear regression model is the appropriate section of the table by Lund (1975) of critical values of the generalization of  $t$  for the general linear model. Ellenberg (1973) had investigated the joint distribution of the studentized residuals for the general case. Lund makes use of Ellenberg's results, invoking the first order Bonferroni inequality, to examine the distribution of the *maximum* individually studentized residual (7.2.7). The resulting upper bounds for the critical values depend on percentage points of  $F$ -distributions at levels not conventionally tabulated. This was the problem which faced all the previously discussed efforts to determine critical values, but Lund approaches it directly by numerically determining the specific critical  $F$ -values which are required. For the linear regression model (7.2.1) upper bounds for the 5 per cent and 1 per cent critical values of  $t$  are to be found in the second columns ( $q=2$ ) of the appropriate tables presented by Lund and reproduced as Table XXXII on pages 335–336, for selected values of  $n$  from 5 to 100. We must recall, however, that the value of  $n$  at which the entries change from being exact to being upper bounds is unknown.

Whilst space does not permit an extended discussion of other aspects of the study of outliers in simple linear regression, we must refer briefly to some further contributions. Elashoff (1972) considers the estimation of  $\theta_0$  and  $\theta_1$  in the special case of outliers generated by a particular combined mixture-slippage model. She assumes, for example, that the  $\epsilon_i$  are uncorrelated and arise from a mixed normal distribution

$$(1-\gamma)\mathbf{N}(0, \sigma^2) + \gamma\mathbf{N}(\lambda(u_i), \sigma^2)$$

with  $\gamma$  known and  $\lambda(u_i) = c(u_i - u_{(1)})^2$ , where  $c$  is constant. This is a highly specific study of the accommodation of outliers in linear regression, and is likely to be of limited applicability.

Methods based on division of the data into sub-samples, for detecting outliers and for estimating parameters in a manner which is robust against outliers, are described by Schweder (1976) and by Hinich and Talwar (1975), respectively. In the former case it is assumed that an uncontaminated sub-sample can be identified, in the latter case trimmed means of the sets of sub-sample estimated regression coefficients are utilized.

Other robust procedures for estimating or testing the regression coefficients are presented by Adichie (1967a, 1967b); they are non-parametric (based on rank tests) and provide *en passant* a degree of protection against outliers.

*Example 7.2.* Consider the load/extension data of Table 7.1. Fitting the linear regression model (7.2.1) we obtain

$$\tilde{\theta}_0 = 0.67656 \quad \tilde{\theta}_1 = 0.07565.$$

The estimated residuals,  $\tilde{\epsilon}_j$ , and studentized residuals,  $\tilde{\epsilon}_j/s_j$ , are as given in Table 7.4. Thus the observation (53.4, 3.1) stands out as an outlier. However, it is not statistically significant (the 5 per cent and 1 per cent critical values being 2.29, 2.44: see Table XXXII).

Table 7.4

$y_j$	$\tilde{\epsilon}_j$	$\tilde{\epsilon}_j/s_j$
1.6	0.076	0.123
2.1	-0.173	-0.245
3.4	0.462	0.616
3.3	0.044	0.058
4.2	0.210	0.272
3.1	-1.616	-2.137
4.9	-0.308	-0.422
6.2	0.894	1.236
6.3	0.411	0.614

### 7.2.2 Multiple regression

The immediate generalization of the simple linear regression model (7.2.1) declares that

$$x_j = \theta_0 + \theta_1 u_{1j} + \theta_2 u_{2j} \dots \theta_{q-1} u_{q-1,j} + \varepsilon_j \quad (7.2.8)$$

where  $u_{1j}, \dots, u_{q-1,j}$  ( $j = 1, 2, \dots, n$ ) are the values taken by  $q-1$  independent variables.

There is little that is special to be said about outliers in the specific context of such multilinear or polynomial regression models. Many of the formal methodological considerations of the early sections carry over in obvious ways: use of residuals, scanning by graphical procedures, implications of

non-constant variance of estimated residuals, difficulties in precise determination of critical values for test statistics, and so on.

In principle the  $u_{ij}$  could take any values and  $q$  could be of any order. Thus (7.2.8) is really just a general linear model; it encompasses simple linear regression and all designed experiments. We have thus reached a stage where it is expedient to proceed directly to the general case and we do so in the following section.

The literature contains few specific studies of multilinear or polynomial regression *per se*. One example is considered in Srikantan (1961) where  $q = 3$  and the  $u_{ij}$  ( $i = 1, 2$ ) are simple trigonometric functions. But later work on the relative insensitivity of the configuration of the values of independent variables (at least for residual-based methods of testing discordancy) implies that we do not need to be so specific and we can reasonably resort to results for the general case.

### 7.3 OUTLIERS WITH GENERAL LINEAR MODELS

Many of the general results on outlier detection, testing discordancy and accommodating outliers in the presence of a general linear model have already been indicated or illustrated by the more specific studies earlier in the chapter. However, we need to draw the threads together in re-examining the range of approaches and describing the extent and manner of generalization of the earlier results. The basic distinction between methods based exclusively on residuals, and other methods, is again evident. Most published results relate to the totally residual-based approach.

#### 7.3.1 Residual-based methods

We are concerned with situations where the observation vector  $\mathbf{x} = (x_1, x_2, \dots, x_n)'$  is represented by a basic general linear model

$$\mathbf{x} = A\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (7.3.1)$$

with  $\boldsymbol{\theta}$  a  $q \times 1$  vector of parameters,  $A$  a known  $n \times q$  matrix of coefficients (assumed here to be of full rank) and  $\boldsymbol{\epsilon}$  a  $n \times 1$  vector of residuals. We assume that  $\boldsymbol{\epsilon}$  has zero mean vector and variance matrix  $V(\boldsymbol{\epsilon}) = \sigma^2 I_n$ , where  $I_n$  is the  $n \times n$  identity matrix, so that the true residuals have common variance and are uncorrelated. (If this were not so it could be created by an appropriate orthogonal transformation.) Any distribution-theoretic results (as needed, for example, in tests of significance) will be based on the assumption that  $\boldsymbol{\epsilon}$  is multivariate normal.

In the absence of any prospect of discordant values, requiring a modification of the model (7.3.1) and possibly revealed as outliers, we have the familiar least-squares analysis of the linear model (7.3.1). The least-squares estimate of  $\boldsymbol{\theta}$  is

$$\tilde{\boldsymbol{\theta}} = (A'A)^{-1}A'\mathbf{x} \quad (7.3.2)$$

with

$$V(\tilde{\theta}) = (A'A)^{-1}\sigma^2, \quad (7.3.3)$$

and of  $\epsilon$  is

$$\tilde{\epsilon} = \mathbf{x} - A\tilde{\theta} = (I_n - R)\mathbf{x} = (I_n - R)\epsilon \quad (7.3.4)$$

where

$$R = A(A'A)^{-1}A' \quad (7.3.5)$$

and with

$$V(\tilde{\epsilon}) = (I_n - R)\hat{\sigma}^2. \quad (7.3.6)$$

The last term of (7.3.4) shows how the estimated residuals  $\tilde{\epsilon}$  relate to the unknown true residuals  $\epsilon$ , but the determination of  $\tilde{\epsilon}$  must be sought in terms of *known* quantities such as  $(I_n - R)\mathbf{x}$ . The estimated residuals  $\tilde{\epsilon}_j$  have zero means. From (7.3.6) we see that they are typically correlated and have differing variances. Explicitly we can write

$$\text{var}(\tilde{\epsilon}_j) = (1 - A'_j(A'A)^{-1}A_j)\sigma^2 = (1 - r_{jj})\sigma^2 \quad (7.3.7)$$

(say) where  $A'_j$  is the  $j$ th row of  $A$ .

The error variance  $\sigma^2$  will be unknown. An unbiased estimate is obtained as

$$\tilde{\sigma}^2 = \tilde{\epsilon}'\tilde{\epsilon}/(n - q) = \epsilon'(I_n - R)\epsilon/(n - q) \quad (7.3.8)$$

in view of the idempotency of  $(I_n - R)$ ;  $\tilde{\epsilon}'\tilde{\epsilon}$  is termed the residual sum of squares and is denoted  $S^2$ .  $V(\tilde{\epsilon})$  can now be estimated as

$$S^2(\tilde{\epsilon}) = (I_n - R)\tilde{\sigma}^2 \quad (7.3.9)$$

so that the estimated variance of  $\tilde{\epsilon}_j$  is

$$s_j^2 = (1 - r_{jj})\tilde{\sigma}^2 = (1 - r_{jj})(\tilde{\epsilon}'\tilde{\epsilon})/(n - q) = (1 - r_{jj})S^2/(n - q). \quad (7.3.10)$$

We shall have reason to consider the *studentized residuals*

$$e_j = \tilde{\epsilon}_j/s_j = \frac{\tilde{\epsilon}_j}{S} \sqrt{\frac{(n - q)}{1 - r_{jj}}}. \quad (7.3.11)$$

They have an immediate intuitive appeal in that they constitute weighted versions of the estimated residuals  $\tilde{\epsilon}_j$ , where the weights are inversely proportional to estimates of the standard deviations of the  $\tilde{\epsilon}_j$ . The variances of the  $e_j$  should thus be more or less constant (precisely so if  $S$  in (7.3.11) were replaced by  $\sigma\sqrt{(n - q)}$ ), avoiding the inconvenience of the disparate variances of the  $\tilde{\epsilon}_j$ .

If  $\epsilon$  is normally distributed the estimates  $\tilde{\theta}$ ,  $\tilde{\epsilon}$ , and  $\tilde{\sigma}^2$  are of course maximum likelihood estimators, and their distributional behaviour is well known ( $\tilde{\theta}_j$  and  $\tilde{\epsilon}_j$  are normally distributed,  $S^2$  has a  $\sigma^2\chi^2$  distribution with  $(n - q)$  degrees of freedom).

Shortly we will need to consider in some detail why the estimated residuals, and studentized residuals, are particularly relevant to the study of outliers. But first it is appropriate to review the state of knowledge about these quantities in their own right, and their use in informal methods of examining outliers. A fundamental work on the analysis of least-squares residuals is by Anscombe (1961). He proposes methods using estimated residuals to examine the assumption that the  $\varepsilon_i$  are independent and normally distributed with constant variance. The methods include a study of the regression of the estimated residuals on the corresponding fitted values, and take account of the intercorrelation of the estimated residuals. Tukey's test of non-additivity is also considered. Particular attention is given to the case where all the residuals have equal variance, typical of factorial design experiments with equal replication (see Section 7.1.1). Outliers receive only passing mention.

Srikantan (1961), however, is concerned with using estimated residuals specifically to investigate outliers. He adopts a mean-slippage alternative model for a single outlier and assumes a normal error structure. For the *labelled* slippage model (see Section 3.1) where the index of the discordant value is specified he shows that for a one-sided (two-sided) test of discordancy the corresponding *studentized* residual is the test statistic of the uniformly most powerful (unbiased) test of discordancy. Thus if the alternative hypothesis is

$$\bar{H}: E(\mathbf{X}) = A\theta + \mathbf{a} \quad (7.3.12)$$

where  $a$  is an  $n \times 1$  vector of zeros apart from one possible non-zero value in the  $i$ th position, the tests are based on  $d_i = \tilde{\varepsilon}_i/s_i$  in the following way. Suppose

$$\begin{aligned} t_i &= d_i^2 \\ u_i &= \begin{cases} d_i^2 & (d_i \geq 0) \\ 0 & (d_i < 0) \end{cases} \\ v_i &= \begin{cases} 0 & (d_i \geq 0) \\ d_i^2 & (d_i < 0). \end{cases} \end{aligned} \quad (7.3.13)$$

The uniformly most powerful (one-sided) tests of  $a_i = 0$  against  $a_i > 0$  and  $a_i < 0$  have rejection regions  $u_i > R_\alpha$  and  $v_i < R_\alpha$ , respectively, where  $R_\alpha$  has to be determined to produce a test of size  $\alpha$ . Against  $a_i \neq 0$  the two-sided test with rejection region  $t_i \geq R'_\alpha$  is uniformly most powerful unbiased (again  $R'_\alpha$  has to be chosen to yield a test of size  $\alpha$ ).

For the (more realistic) *unlabelled* slippage model, where the index  $i$  of the discordant value is unspecified, similar tests based on  $t = \max_i d_i^2$ ,  $u = \max_i u_i$ , and  $v = \max_i v_i$  are recommended, but there is no consideration of their optimality properties. We shall return to the thorny issue of the determination of the critical values (the corresponding  $R'_\alpha$  and  $R_\alpha$ ) of these tests.

Tiao and Guttman (1967) reconsider the premium-protection approach of Anscombe (1960a) in some detail for the case of a single univariate sample, depending on the extent of knowledge of the residual variance  $\sigma^2$ . (See Section 4.1.1.) A brief comment on the general linear model is interesting. Recognizing the difficulties arising from the intercorrelation of the estimated residuals, they propose using uncorrelated *modified residuals*, having the form (when  $\sigma^2$  is known)

$$\mathbf{z} = \mathbf{\epsilon} + \sigma \mathbf{A} \mathbf{P} \mathbf{u} \quad (7.3.14)$$

where  $P$  is any  $q \times q$  matrix satisfying

$$\mathbf{P}\mathbf{P}' = (\mathbf{A}'\mathbf{A})^{-1} \quad (7.3.15)$$

and  $\mathbf{u}$  is  $\mathbf{N}(0, I_q)$  independent of  $\mathbf{X}$ .

From (7.3.5) and (7.3.6) we clearly have

$$\mathbf{V}(\mathbf{z}) = [(I_n - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}') + \mathbf{A}\mathbf{P}\mathbf{P}'\mathbf{A}']\sigma^2 = I_n\sigma^2. \quad (7.3.16)$$

Such modified residuals (or *adjusted residuals*) might seem attractive in view of their independence when the error distribution is normal. Various accommodation procedures (including premium-protection rules and Winsorization) utilizing them are discussed in detail by Golub, Guttman, and Dutter (1973) for a location-slippage model for possible discordant values. They remark that for large  $n$  the correction factors in (7.3.14) arising from introducing alien independent perturbations of the residuals to "break" the correlation pattern' are small. But there is no detailed consideration of how large  $n$  needs to be, nor is there any comparison of their approach with others based on the unadjusted residuals taking proper account of their intercorrelations.

A different approach to tests of discordancy of outliers, based on residuals, is described by Andrews (1971). He considers the unit vector of *normalized residuals*  $\mathbf{r} = \tilde{\mathbf{\epsilon}} / (\tilde{\mathbf{\epsilon}}'\tilde{\mathbf{\epsilon}})^{1/2} = \tilde{\mathbf{\epsilon}} / S$  and develops tests of discordancy based on a projection of this vector onto a suitable subspace. Adopting the outlier model (7.3.12) we have

$$\mathbf{S}\mathbf{r} = \mathbf{S}'\mathbf{r}' + (I_n - \mathbf{R})\mathbf{a} \quad (7.3.17)$$

where  $(S')^2$  and  $\mathbf{r}'$  are the residual sum of squares, and vector of estimated residuals, respectively, when the basic (no-outlier) model is true, i.e.  $\mathbf{a} = \mathbf{0}$ . Thus if  $\mathbf{a} \neq \mathbf{0}$ ,  $\mathbf{r}'$  is perturbed by a vector in the direction  $(I_n - \mathbf{R})\mathbf{g}$  where  $\mathbf{g}$  is an  $n \times 1$  vector with  $n - 1$  zeros and a unit value at the position corresponding with the discordant value, as indicated by  $\mathbf{a}$ . Discordancy is revealed by  $\mathbf{r}$  being too close to  $(I_n - \mathbf{R})\mathbf{g}$  and we can assess this in terms of how small is the norm  $\|\mathbf{r}^*\|$  of the orthogonal complement  $\mathbf{r}^*$  of the projection of  $\mathbf{r}$  on  $(I_n - \mathbf{R})\mathbf{g}$ .

Of course, we will not often wish to specify the index of the potential discordant value, but will need to test the unlabelled mean-slippage model

corresponding with (7.3.12). This leads to a test which attributes discordancy to a single outlier when  $\min_j \|r_j^*\|$  is sufficiently small, where  $j$  denotes the index of the potential discordant value. Andrews develops this test in the cases where the error distribution is normal, and exponential. The approach is also extended to multiple outliers. He claims that his approach is essentially different from others which examine merely the absolute values of the residuals (or normed residuals), and he suggests that it is an improvement in view of its ability to take account of the form of  $A$ . But this facility, as expressed through the use of projection vectors, seems to reduce merely to the use of the individually studentized residuals  $e_j$  of (7.3.11) rather than of the undifferentiably weighted normed residuals. See also John and Draper (1978).

For a generalized approach to the analysis of residuals (but with only passing reference to outliers) see Cox and Snell (1968, 1971). See also Behnken and Draper (1972), Draper and Smith (1966), and Wooding (1969). Ellenberg (1973) considers in some detail the joint distribution of the studentized residuals (7.3.11) in a general linear model with normal error structure, yielding what he terms 'a standardised version of the Inverted-Student Function'. The various graphical methods described in earlier sections will also be useful for exhibiting outliers, although it will be more appropriate to plot the individually studentized residuals  $e_j$  rather than the ordinary estimated residuals  $\tilde{e}_j$ .

Summarizing the current attitude to tests of discordancy (and accommodation) of outliers in the general linear model situation we find almost total preoccupation with the maximum (positive, negative, or absolute) studentized residual as test statistic, with discordancy attributed to the observation yielding the maximum provided that maximum is sufficiently large. This is precisely the Srikantan (1961) prescription. Illustrating it for the two-sided test, we reject the basic hypothesis of (7.3.1) in favour of one which postulates a single discordant value, e.g. (7.3.12), if

$$\tau = \max_j |e_j| = \max_j |\tilde{e}_j / s_j| > h_\alpha \quad (7.3.18)$$

where for a test of size  $\alpha$  the critical value  $h_\alpha$  needs to be chosen to ensure that, under (7.3.1), (7.3.18) holds with probability  $\alpha$ . The observation which yields the maximum value of  $e_j$  is declared to be the discordant outlier.

Most discussion of such a test centres on the problem of determining the critical values  $h_\alpha$ . We assume first of all that  $\sigma^2$  is unknown and has to be estimated solely from the current data. Srikantan (1961) used the first Bonferroni inequality to determine upper bounds for the critical values of his statistics (discussed above: defined in terms of squared studentized residuals). As we saw, he was able to present results for some models with  $q$  up to 3 in value. We have also noted the special case approximations of Prescott (1975a, 1975b), and simulation results of Tietjen, Moore, and

Beckman (1973), both limited to simple linear regression. Lund (1975) presents the most useful tabulation to date. Using the results of Ellenberg (1973), and again the first Bonferroni inequality, he determines the required (unavailable) percentage points of the  $F$ -distribution *en route* to a fairly comprehensive tabulation of upper bounds to the 10, 5, and 1 per cent points of  $\tau$  under the basic model, for  $n = 5(1)10(2)20(5)50(10)100$  and  $q = 1(1)6(2)10, 15, 25$ . The 5 per cent and 1 per cent are reproduced as Table XXXII on pages 335–336.

When we have some knowledge about  $\sigma^2$ , external to the data, the discordancy test needs appropriate modification. Joshi (1972a) considers this for the normal error model where we have either an external estimate  $s_\nu^2$  of  $\sigma^2$  distributed as  $\sigma^2 \chi_\nu^2/\nu$  independent of  $\mathbf{X}$ , or where we know  $\sigma^2$  precisely. The test structure is the same as before except that the (internally) studentized residuals  $\tilde{\varepsilon}_i/s_i$  are replaced by externally studentized, pooled studentized, or standardized residuals

$$\begin{aligned} & \tilde{\varepsilon}_i / [(1 - r_{ii}) s_\nu^2]^{1/2}, \\ & (\nu + n - q)^{1/2} \tilde{\varepsilon}_i / [(1 - r_{ii})(\nu s_\nu^2 + S^2)]^{1/2}, \\ \text{and } & \tilde{\varepsilon}_i / [(1 - r_{ii}) \sigma^2]^{1/2}. \end{aligned}$$

A test of discordancy again proceeds in terms of the maximum (positive, negative, or absolute) value of the weighted residuals. For the pooled studentized residuals,  $\nu = 0$  reduces to the original test based on the  $\tilde{\varepsilon}_i/s_i$ .

Joshi's approach to the determination of the critical values of the tests again yields only upper bounds but they turn out to be intermediate between these given by the first Bonferroni inequality and the more precise (but less computationally tractable) second Bonferroni inequality. The performance of the tests is considered, and illustrated numerically for the simplest case of a single univariate normal sample. Throughout it is assumed under the alternative hypothesis that there is a single discordant value corresponding with a constant slippage of the mean; its index is unknown and is assumed to be chosen at random from the set  $(1, 2, \dots, n)$ .

In concluding this section we return to the basic question. Apart from intuitive appeal, why should we regard the individually studentized residuals as appropriate representations of the data for detecting outliers and for testing discordancy?

The basic model is (7.3.1):  $\mathbf{x} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ . The prospect of a single discordant value reflecting slippage in the mean can be expressed in terms of the set of alternative hypotheses

$$\bar{H}_j: \mathbf{x} = \mathbf{A}\boldsymbol{\theta} + \mathbf{a}_j + \boldsymbol{\varepsilon} \quad (j = 1, 2, \dots, n) \quad (7.3.19)$$

where  $\mathbf{a}_j$  is  $n \times 1$  with a value  $a$  in the  $j$ th position and zeros elsewhere. Thus  $\bar{H}_j$  declares that  $x_j$  is the discordant value, from the distribution  $\mathbf{N}(A'_j\boldsymbol{\theta} + a, \sigma^2)$  where  $A'_j$  is the  $j$ th row of  $A$ . As in our study of multivariate outliers,

for example, we might choose to detect a single outlier in terms of greatest increase in the maximized likelihood under the set of hypotheses  $\bar{H}_j$  relative to the basic model  $H: \mathbf{x} = A\boldsymbol{\theta} + \boldsymbol{\epsilon}$ .

Under  $H$  the likelihood is maximized by putting  $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$  and  $\sigma^2 = S^2/n = (\boldsymbol{\epsilon}'\boldsymbol{\epsilon})/n$ . The maximized log-likelihood is

$$-\frac{n}{2} \log \left( \frac{2\pi S^2}{n} \right) - \frac{n}{2}.$$

Under  $\bar{H}_j$ ,  $x_j$  arises from  $N(\mu, \sigma^2)$  whilst  $\mathbf{x}_{-j}$  arises (independently) from  $N(A_{-j}\boldsymbol{\theta}, \sigma^2 I_{n-1})$  where  $\mathbf{x}_{-j}$  is the set of observations excluding  $x_j$  and  $A_{-j}$  is the reduced matrix  $A$  obtained on deletion of the  $j$ th row. The likelihood is now maximized by putting  $\mu = x_j$  and

$$\begin{aligned}\boldsymbol{\theta} &= (A'_{-j} A_{-j})^{-1} A'_{-j} \mathbf{x}_{-j} \\ \sigma^2 &= S_{-j}^2/n\end{aligned}$$

where  $S_{-j}^2$  is the sum of squares of the estimated residuals when the reduced data vector  $\mathbf{x}_{-j}$  is fitted by least-squares to the model  $\mathbf{x}_{-j} = A_{-j}\boldsymbol{\theta} + \boldsymbol{\epsilon}_{-j}$ . The maximized log-likelihood now becomes

$$-\frac{n}{2} \log \left( \frac{2\pi S_{-j}^2}{n} \right) - \frac{n}{2}.$$

Thus the increase in the maximized log-likelihood is

$$\frac{n}{2} \log (S^2/S_{-j}^2)$$

and so on the above criterion (with no restriction on the value of  $a$ ) a single outlier is detected as that observation whose omission from the sample effects the greatest reduction in the residual sum of squares. Thus the outlier is the observation yielding

$$\max_j (S^2/S_{-j}^2).$$

It is to be adjudged discordant if this maximum is sufficiently large relative to its null distribution.

In fact we can show that the simple relationship holds:

$$\begin{aligned}S^2 &= S_{-j}^2 + \tilde{\epsilon}_j^2/[1 - A'_j(A'A)^{-1}A_j] \\ &= S_{-j}^2 + \tilde{\epsilon}_j^2/(1 - r_{jj}).\end{aligned}\tag{7.3.20}$$

Hence

$$\begin{aligned}S^2/S_{-j}^2 &= S^2/[S^2 - \tilde{\epsilon}_j^2/(1 - r_{jj})] \\ &= \{1 - \tilde{\epsilon}_j^2/[S^2(1 - r_{jj})]\}^{-1} \\ &= [1 - e_j^2/(n - q)]^{-1}\end{aligned}\tag{7.3.21}$$

and we see that maximization of  $S^2/S_{-j}^2$  is merely equivalent to maximization of the squares of the (individually) *studentized residuals*,  $e_j$ . So the *maximum likelihood ratio* procedure involves examination of the squares (or absolute values) of the studentized residuals, detection of the outlier as the observation yielding the maximum absolute studentized residual, and assessment of discordancy if it is statistically too large relative to the basic model. This equivalence between reduction in the residual sum of squares and the absolute values of the studentized residuals is demonstrated by Ellenberg (1973, 1976).

So the preoccupation with studentized residuals as an indication of outliers in the general linear model finds a sound foundation on the maximum likelihood ratio principle, and discordancy tests based on  $\max_j(S^2/S_{-j}^2)$  and on  $\max_j(e_j)$  are equivalent. The maximum likelihood ratio basis for these tests is exhibited by Fieller (1976).

This equivalence has important implications as we shall see in the next section.

### 7.3.2 Non-residual-based methods

There are few methods available for dealing with outliers with the general linear model which do not make use of residuals in some way. Various methods based on residual sums of squares may seem at first sight (and indeed are often claimed by their progenitors) not to involve direct study of the estimated residuals. It is sometimes advanced that this avoids a disadvantageous cloaking (or confounding) effect of an outlier arising from the fact that *all* the estimated residuals reflect the influence of the outlier. The view is expressed that there is a contradiction in trying to detect an outlier by examining estimated residuals which are 'biased by the presence' of the outlier we are trying to detect. We have examined a method due to Goldsmith and Boddy (1973) for outlier detection in factorial experiments which consists of examining the set of  $n$  residual sums of squares obtained by regarding each observation separately as a missing value. In Example 7.1 we noticed how this approach highlighted a single outlier, in a manner which seemed more dramatic than the corresponding indication from the values of the estimated residuals.

However, the results at the end of the previous section cast doubt on any advantage in the use of residual sums of squares rather than of estimated (studentized) residuals. Regarding an observation  $x_j$  as missing will have the same effect as estimating the parameters under the model  $\tilde{H}_j$  and the resulting residual sum of squares will be just  $S_{-j}^2$ . But in view of (7.3.20) the separate residual sums of squares are readily obtained merely by reducing the overall residual sum of squares  $S^2$  by the corresponding weighted squared estimated residual  $\hat{\varepsilon}_j^2/(1-r_{jj})$ . Thus three conclusions arise. The calculation is the same in both approaches, involving determination of the

estimated residuals and of  $S^2$ . The visual impact is equivalent, being the reflection of the values of weighted squared estimated residuals (it might appear heightened on the residual sum of squares approach merely because we are there looking at a reduction from  $S^2$  by the square of the estimated residual, rather than at just the absolute value of the residual). Finally, no statistical distinction exists; in particular the method is still subject to the cloaking, or confounding, influence of the outliers we are seeking to detect.

*Example 7.3. Returning to Example 7.1 we can demonstrate this equivalence. In this factorial experiment all the estimated residuals have equal variance. In fact  $r_{ij} = r = 0.5$ . Now  $S^2 = 2152.5$  and if we reduce  $S^2$  separately by the  $\tilde{\epsilon}_j^2/(1-r)$  using the values for  $\tilde{\epsilon}_j$  in column (5) we obtain (after appropriate weighting by the degrees of freedom) the residual mean squares corresponding with the entries in column (7). Alternatively, if we examine the squares of the estimated residuals we see that the outlier at 'a' shows up just as graphically as in column (7).*

Mickey (1974) and Mickey, Dunn, and Clark (1967) also proposed such examination of the separate residual sums of squares after omission of the separate residuals singly, suggesting that a large enough reduction is evidence of a discordant outlier. Their test statistic is

$$\max_j \left\{ \frac{S^2 - S_{-j}^2}{S_{-j}^2 / (n - q - 1)} \right\}$$

with attribution of discordancy if it is sufficiently large. (They seem to think it necessary, however, to conduct  $n$  separate regression analyses, rather than just one yielding  $S^2$  and the  $\tilde{\epsilon}_j$ .)

Snedecor and Cochran (1967, page 157) discuss a test of discordancy based on the maximum of the studentized differences, i.e. the test statistic is

$$\max_j \left\{ (x_j - \tilde{x}_{-j}) / [V(x_j - \tilde{x}_{-j})]^{1/2} \right\}$$

where  $\tilde{x}_{-j}$  is the least-squares estimate of  $x_j$  when  $x_j$  is regarded as a missing observation. Again the detected outlier is acclaimed discordant if the test statistic is sufficiently large.

Ellenberg (1976) demonstrates that these two methods are equivalent, and both coincide with the test based on the maximum absolute studentized residual. He employs the result (7.3.20) which he derived earlier—Ellenberg (1973).

The proposal of Gentleman and Wilk (1975a) to examine the ' $k$  most likely outlier subset' (see Section 7.1.4) is also applicable to the general linear model although it will be more computationally laborious than in the case of a two-way design examined in detail by the authors. It also possesses a corresponding link with studentized residuals.

## 7.4 OUTLIERS IN TIME-SERIES

We noted in Chapter 1 two examples of time-series exhibiting outliers. In the first example, illustrated in Figure 1.3, a realization of a non-stationary series of sales figures showed a distinct disruption of the quarterly cyclic pattern of sales, possibly reflecting an outlier. It could be that adverse trading conditions resulting from government action or fiscal policy, short-term emergency company action or even delays in returns of sales figures untypically depressed the sales figure at A but also induced a compensatory untypical sales figure at B. The second example of moisture content of tobacco (Figure 1.4) exhibited apparently *isolated* outliers at A, B, etc. which might have reflected malfunction of the recording equipment.

Notice again the complications involved in detecting an outlier. As in general linear model data, it is not necessarily an extreme value and it can be cloaked to some extent by the general structure of the process. In particular, we can experience a smoothing-out effect when we attempt to examine outliers in terms of derived quantities such as the values of estimated residuals about a fitted model. Whilst in the linear model any outlier does not tend to influence adjacent observations *per se*, merely the estimated residuals, etc., the same need not be true for time-series data in view of the correlational pattern of the basic process. It is fruitful therefore to consider the prospect of two types of outlier.

In the first case an isolated measurement, or execution, error is superimposed on an otherwise reasonable realization of the process. This will not be reflected in the values of adjacent observations and its manifestation can be dramatic and obvious. Such a 'hiccup' effect is possibly what we are noticing in the data of Figure 1.4.

Alternatively, a more inherent discordancy can arise and be reflected by the correlation structure of the process in neighbouring (usually later) observations. The data of Figure 1.3 possibly illustrate this effect. For this type of outlier there is a prospect that the realization itself conspires to conceal the outlier and the detection of outliers becomes more problematical. On the other hand any smoothing-out effected by autocorrelations can have an intrinsic role in accommodating the outlier. It is possible that its influence on parameter estimation or testing in the basic time-series model may be less acute than with independent error structure.

Huber (1972) claims that the 'hiccup' effect is rare; the more usual outlier is of the inherent type revealed more obscurely in 'bumps' and 'quakes'. These are respectively local changes in the mean and variance (requiring corresponding slippage-type alternative models) whose effect extends to influence subsequent observations. From the test-of-discordancy, and accommodation, viewpoints Huber suggests examining coefficients of skewness or kurtosis or applying a smoothing process, respectively, but offers little by way of detailed prescriptions.

Indeed there is little published work on outliers in time-series in terms either of frequency-domain or time-domain analyses and this would seem to be an important and highly challenging area for further study.

One of the few contributions to date is that of Fox (1972) who defines two types of outlier which might occur in time-series data. His type I and type II outliers are precisely the isolated independent gross execution or recording errors, independent of other observations, and the 'inherent' type of anomalous observation which influences succeeding observations, which we distinguished above. Four situations are postulated:

- (i) all outliers are of type I,
- (ii) all outliers are of type II,
- (iii) all outliers are of the same type, but we do not know which type, and
- (iv) both types of outlier are present.

How we are to assess which of these situations prevails is not considered apart from remarking that (ii) will be distinguished from (i) by the presence of the carry-over effect. Only situations (i) and (ii) are examined in some detail, on the assumption that the process is free of trend or seasonal factors; the possible effect of their removal on the examination of outliers is not discussed. Thus the methods presented have obvious limitations, compounded by an additional assumption that there is at most one outlier, but they do provide a starting point in this difficult area of study.

A test for type I outliers is developed in relation to the mean-slippage outlier model for a discrete time-series

$$x_t = u_t + \delta_{j_t} a \quad (7.4.1)$$

where  $\delta_{j_t}$  is the Kronecker delta function and the  $u_t$  satisfy an autoregressive scheme of order  $p$

$$u_t = \sum_{l=1}^p \alpha_l u_{t-l} + z_t \quad (t = p+1, \dots, n) \quad (7.4.2)$$

where the  $z_t$  are independent  $N(0, \sigma^2)$ . Thus we have a set of  $n$  observations of a discrete process, further restricted by the assumption that  $p$  is known and that  $\{u_t\}$  is a stationary process, with a superimposed discordant value at time point  $j$ . Both the cases of prescribed  $j$  and unknown  $j$  are considered and maximum likelihood ratio tests of  $H: a = 0$  against  $\bar{H}: a \neq 0$  are developed. For the latter more realistic case ( $j$  unknown) the maximum likelihood ratio statistic is equivalent to

$$\max_{j=p+1, \dots, n-p} (k_{j,n})$$

where

$$k_{j,n} = \mathbf{x}' \hat{W}^{-1} \mathbf{x} / (\mathbf{x} - \tilde{\mathbf{a}})' \tilde{W}^{-1} (\mathbf{x} - \tilde{\mathbf{a}}). \quad (7.4.3)$$

In (7.4.3)  $\tilde{\alpha}$  is  $\tilde{\alpha}(0, 0, \dots, 0, 1, 0, \dots, 0)'$  where the 1 appears in position  $j$  and  $\tilde{\alpha}$  is the maximum likelihood estimate of  $a$  under (7.4.1) and (7.4.2);  $\hat{W}^{-1}$  and  $\bar{W}^{-1}$  are the maximum likelihood estimates of  $W^{-1}$  under  $H$  and  $\bar{H}$ , respectively, where the covariance matrix of the process has the form

$$V = W\sigma^2 \quad (7.4.4)$$

which depends only on  $p$  and the auto-regressive coefficients  $\alpha_l(l = 1, 2, \dots, p)$ . Note that the elements of  $W$  have the form  $w_{t,t'} = w_{|t-t'|}$ .

The test of discordancy detects the outlier as the observation maximising the  $k_{j,n}$  and declares it discordant if the maximum value is sufficiently large. Under  $H$  we are involved in determining the distribution of the maximum of a set of  $n$  correlated  $F$ -variates. Significance levels, power calculations, and the behaviour of modified tests are all examined by Fox using simulation methods.

The outlier model used by him for a test of discordancy for a type II outlier has the form

$$x_t = \sum_{l=1}^p \alpha_l x_{t-l} + \delta_j a + z_t \quad (7.4.5)$$

with all quantities defined and limited as before. We can see here the carry-over effect of the discordant value. Again the *maximum likelihood ratio* test of  $H: a = 0$  against  $\bar{H}: a \neq 0$  is developed, and studied by simulation, for the case where  $j$  is specified. The more important case of an unspecified value for  $j$  is not pursued. Some implications of employing the wrong model ((7.4.5) instead of (7.4.1) and (7.4.2) or vice versa) are also examined by simulation.

Another approach to the study of outliers in correlated data appears in a paper by Guttman and Tiao (1978). They consider the premium-protection approach to estimating the mean of certain stationary autocorrelated discrete time processes.

## CHAPTER 8

# *Bayesian and Non-Parametric Approaches*

Passing reference has already been made to the use of Bayesian, and of non-parametric, methods in different aspects of the study of outliers. In this chapter we draw together the threads of each of these two approaches by considering in more detail the specific proposals that have been made for testing the discordancy of outliers or of coping with their presence in ‘contaminated’ data. We concentrate mainly on univariate data since most contributions are in this area.

### 8.1 BAYESIAN METHODS

In the context of the Bayesian approach a test of significance has little relevance. It is useful, for the sake of continuity of argument, to maintain a distinction between the statistical detection of outliers and their accommodation within a broader analysis of the data. The notion of a test of discordancy for assessing the import of outliers has, however, to be appropriately re-expressed.

#### **8.1.1 Bayesian ‘tests of discordancy’**

In Chapter 1 we remarked briefly on what might seem to be the somewhat anomalous role of Bayesian methods in the study of outliers. The essential nature of an outlier is found in the degree of ‘surprise’ it engenders when we examine a set of data. Early informal methods of handling outliers consisted in developing procedures for detecting and *rejecting* them as ‘foreign influences’ reflecting undesirable errors in the data collection process. Such an attitude does not really fit the Bayesian idiom with its dual regard for the *total* data set as the basic information ingredient from which conditional inferences are to be drawn and with the likelihood as the full statistical expression of the information in the data. Preliminary processing of the data for detection, and possible rejection, of outliers implies a possibly unwarranted preoccupation with a specific feature of the data with insufficient

regard to its total import. The crucial statement of the likelihood involves a commitment to a fully specified model—which was certainly not a feature of the ad hoc studies of outliers.

Again, the Bayesian approach requires an *a priori* statement about the propriety of possible models, or about possible values of the parameters in a parametric family of models. This would have to include a prior assessment of probabilities attaching to the presence and form of outliers; the assessment must be made before the data are available, and irrespective of the characteristics of any realized sample. But before we have collected our data how are we to recognize the prospect of outliers; there is nothing to surprise us? There seems to be a degree of conflict here, between a data-keyed response to anomalous observations and a data-independent incorporation of prospective outliers in the likelihood and prior probability assignment.

We have been at pains to stress throughout this book the need to advance beyond the early informal view of outliers to a recognition of the importance of adopting a specific form of outlier-generating model in any development of statistical technique for handling outliers. The inescapable modelling element in the Bayesian approach is thus welcome. Its refinement through the attribution of prior probabilities is also a potentially valuable component in outlier study—provided we really do have some tangible prior information. But to be compelled to produce a prior assessment in any circumstance might be more of an embarrassment than an aid. When all is considered, however, perhaps the major philosophical distinction that remains is found in the irrelevance of the data to the outlier-model specification, and of the sampling rule to the final inference, both of which are essential attitudes in the Bayesian approach. They are both in conflict with the view we have advanced for recognizing, interpreting, and handling outliers on a more classical approach.

The Bayesian attitude asks that we *anticipate* the possible presence of outliers and structure our data-generating model accordingly *before* we observe any data. This is a tenable standpoint, and one which some may feel ‘more objective’, though whether it is an honest expression of what happens in data analysis is another matter.

Notwithstanding the philosophical issues, many Bayesian methods for outlier study have been advanced and we shall proceed to examine some of them. Inevitably the test of discordancy does not have an immediate parallel in Bayesian terms, the final inference being in the form of a posterior distribution. We carry over the term to those situations where the major interest is in drawing conclusions about the outliers, rather than about other parameters (with the outlier merely ‘accommodated’ at a lower level of interest).

One of the earliest discussions of the Bayesian approach to outliers is that of de Finetti (1961). He is primarily concerned with exploring basic attitudes rather than developing technique. The discussion is set in the context of

some quantity  $X$  having an initial (prior) distribution which becomes modified in the light of a sample of observations  $x_1, x_2, \dots, x_n$  to yield a final (posterior) distribution for  $X$ . There is no estimation or testing problem—the total inference is expressed by the final distribution of  $X$ . Claiming that all inference problems are so represented, de Finetti argues that any reasonable approach to outlier rejection needs to be couched in such terms. He stresses that this raises a fundamental difficulty in that the final distribution (total inference) depends on all the data, an attitude which conflicts with the preliminary rejection of some observations (as outliers). He concludes that if rejection of outliers has any propriety this must hinge on the fact that any observation serving as a candidate for rejection has a ‘weak or practically negligible’ influence on the final distribution.

This viewpoint opposes much of the *rationale* for outlier processing described in the earlier chapters of this book. Apart perhaps from the *identification* of outliers as observations of special intrinsic interest, there would seem to be little point in, or basis for, *either rejecting or accommodating* outliers if their presence has negligible influence on the inferential import of the data!

In exploring his viewpoint, de Finetti insists that outlier rejection ‘could be justified only as an approximation to the exact Bayesian rule (under well-specified hypotheses), but never by empirical ad hoc reasoning’. He seeks substance in an example where interest centres on some ‘estimator’  $\hat{x}$  obtained by a Bayesian analysis in the form of a weighted average of the  $x_i$ :

$$\hat{x} = \rho_1 x_1 + \rho_2 x_2 + \dots + \rho_n x_n.$$

If the weights are complicated functions of the  $x_i$  but a simpler rule yields a good approximation to  $\hat{x}$  and also takes the form of a weighted average but with roughly equal weights for most  $x_i$  and negligible weights for the others then these latter observations ‘are outliers in regard to this method’. But de Finetti finds such a notion ‘vague and rather arbitrary’. To make it more substantial he restricts attention to an ‘estimate’ expressible as the mean of the final (posterior) distribution, where the initial (prior) distribution is uniform. To explore the possibility of realizing the above manifestation of outliers (as lowly weighted observations in a linear form of inferential interest) de Finetti examines cases where the  $x_i$  are independent, exchangeable, or partially exchangeable. Some success is achieved (for uniform prior distributions) when the error distributions (the distributions of  $X_i - X$ ) are rather simple mixtures of some common distributions.

The approach yields no direct, generally applicable, procedures for outlier rejection and the final message is that the Bayesian approach militates against such a prospect. However the author’s pessimistically expressed conclusion that the data structure and error model are crucial to what procedure should be employed seems no different to the message this book

has been presenting in relation to more traditional approaches to outlier treatment.

We have already referred (in Sections 2.3 and 4.4) to a novel approach to the modelling of outliers employed by Kale, Sinha, Veale, and others (see Kale and Sinha, 1971). Here it is assumed that  $n - k$  of the observations  $x_1, x_2, \dots, x_n$  arise from some basic population  $F$  whilst the remainder (the outliers) arise from populations  $G_1, G_2, \dots, G_k$  different from  $F$ . It is assumed that prior to taking the observations there is no way of identifying the anomalous subset of size  $k$ . Furthermore, such identification does not arise from the observed values. Instead, it is assumed that any subset of  $k$  of the  $n$  observations is equally likely to be the set of observations arising from  $G_1, G_2, \dots, G_k$ . We termed this the *exchangeable* model.

It is a moot point whether the use of the uniform distribution for the indices of the anomalous subset in this model implies that the approach is Bayesian in spirit. For a full Bayesian approach a specification of prior probabilities for the forms of the populations  $F$  and  $G_i$  ( $i = 1, 2, \dots, k$ ) would be required, and inferences would need to be expressed in terms of an appropriate posterior distribution. Kale (1974b) considers this extended prospect in the case where  $F$  and  $G_i$  ( $i = 1, 2, \dots, k$ ) are all members of the single-parameter family of exponential distributions with parameter values  $\theta$  and  $\theta_i$  ( $i = 1, 2, \dots, k$ ), respectively. He shows that with minimal restrictions on the prior distribution  $p(\theta, \theta_1, \theta_2, \dots, \theta_k)$  we obtain the same prescription for identifying the anomalous subset in terms of that set of  $k$  indices which has maximum posterior probability of corresponding with  $\theta_1, \theta_2, \dots, \theta_k$ : namely that if  $\theta_j \leq \theta$  ( $j = 1, 2, \dots, k_1$ ) and  $\theta_j \geq \theta$  ( $j = k_1 + 1, k_1 + 2, \dots, k$ ) the anomalous observations are  $x_{(1)}, x_{(2)}, \dots, x_{(k_1)}$  and  $x_{(n-k+k_1+1)}, x_{(n-k+k_1+2)}, \dots, x_{(n)}$ . This is of course the intuitively sensible conclusion.

Kale also postulates that since no serious restriction was placed on  $p(\theta, \theta_1, \theta_2, \dots, \theta_k)$  this result will hold on a more classical approach employing no specification of prior attitudes about the parameters. Indeed, Kale (1974a) has proved the corresponding result for the case  $\theta_j = \theta'$  ( $j = 1, 2, \dots, k$ ) with  $\theta' \geq \theta$  or  $\theta' \leq \theta$ .

Most of the following proposals for Bayesian analysis of outliers involve a similar form of *exchangeable* model to account for the presence of outliers, but take the discussion further by considering the choice of  $k$ , other forms of distribution  $F$  and  $G_i$  (although usually limited to  $G_i \equiv G$ ;  $i = 1, 2, \dots, k$ ) and the estimation of parameters in the face of outliers. These include the work of Dempster and Rosner (1971), Guttman (1973b) on identification and discordancy; and Box and Tiao (1968), Sinha (1972, 1973b) on accommodation.

The first example we consider of a more detailed Bayesian approach to detecting and testing outliers is embodied in the proposals of Dempster and Rosner (1971). They describe their approach as 'semi-Bayesian' in that the detection and informal assessment of discordancy of a set of  $k$  outliers (for

prescribed  $k$ ) is in the Bayesian mould, but the choice of  $k$  does not proceed from any prior probability distribution of possible values for  $k$ . Instead, the choice of  $k$  involves a combination of the fixed- $k$  analyses and classical significance test ideas. This latter aspect will not be considered. We concentrate on the fundamental Bayesian aspects, suggesting minor extensions of interpretation where the authors are not too specific in their proposals.

Suppose that  $x_1, x_2, \dots, x_n$  are independent observations from normal distributions with common unknown variance  $\sigma^2$ . If no outliers are present the distributions all have zero mean. Alternatively, a location-shift model prevails with  $k$  of the means different from zero. A Bayesian analysis for a prescribed value of  $k$  proceeds as follows.

For given  $k$ , there are  $\binom{n}{k}$  subsets of observations which are candidates for assessment as sets of  $k$  outliers. Assuming each subset to be equally likely, *a priori*, to fulfil this role the posterior probability  $\pi(I)$  that subset  $I$  is the outlier subset is proportional to

$$\left[ \sum_{i \in I} x_i^2 \right]^{-n/2}$$

on the assumption of ‘relatively innocuous’ uniform prior distributions for  $\log \sigma$  and for the unknown location parameters. (But we are advised to recall some of the anomalies that can arise from multi-parameter uniform prior distributions; see Dawid, Stone, and Zidek, 1973.)

Clearly  $\pi(I)$  is maximized when  $I$  consists of the  $k$  observations with largest absolute values. Such would have been the observations singled out as outlying in a more traditional data-oriented approach where the detection stage proceeds intuitively in terms of the degree of ‘surprise’. Here, no such pre-detection is admitted—indeed it is ruled out by the adopted uniform prior distribution of  $I$ .

It is proposed that the (marginal) posterior probability that  $x_i$  is an outlier is measured by

$$p_i = \sum_{I \ni i} \pi(I).$$

The question of whether or not the detected set of  $k$  outliers is discordant presumably hinges on how large is

$$\pi_k = \max_{I \in \mathcal{J}} \pi(I) \quad (8.1.1)$$

where  $\mathcal{J}$  is the set of all subsets  $I$  of size  $k$ . We might decide that this needs to exceed, say, 0.95 before we attribute discordancy to the outlier subset.

The question of choice of  $k$  is crucial. Dempster and Rosner discuss some fundamental obstacles to a full Bayesian analysis of this matter. Instead, they propose that we consider the sequence of maximized posterior probabilities  $\pi_k$  for  $k = 1, 2, 3, \dots$  and suggest that we choose  $k$  to yield  $\pi_k$

'large enough to provide reasonable assurance that the  $k$  most discrepant data points are outliers', coupling this rather general prescription with informal aids involving significance testing concepts. An alternative might be to choose  $k$  to maximize  $\pi_k$  and to conclude that the  $k$  detected outliers are discordant if  $\max_k \pi_k$  is sufficiently large. But clearly any such proposal would need a careful study of its implications.

For prescribed  $k$ , Dempster and Rosner suggest estimators of  $\sigma^2$  and of the anomalous means  $\mu_i$ , which robustly accommodate the set of  $k$  outliers. In the former case they propose

$$\tilde{\sigma}_k^2 = \sum_{I \in \mathcal{S}} \pi(I) S_I^2 \quad (8.1.2)$$

where

$$S_I^2 = \frac{1}{n-k} \sum_{i \notin I} x_i^2. \quad (8.1.3)$$

For the  $\mu_i$  the estimators are

$$\tilde{\mu}_i = \sum_{I \in \mathcal{S}} \pi(I) \tilde{\mu}_i(I) \quad (8.1.4)$$

where

$$\tilde{\mu}_i(I) = \begin{cases} x_i & \text{if } i \in I \\ 0 & \text{otherwise.} \end{cases} \quad (8.1.5)$$

To illustrate their proposals, Dempster and Rosner reconsider the data discussed in Daniel (1959) in his work on half-normal plots. (See Sections 2.2 and 7.1.3.) It is interesting to reproduce some of the results. The Daniel data (the 31 contrasts in a  $2^5$  experiment) arranged in ascending order of magnitude appear thus:

0.0000	0.0281	-0.0561	-0.0842	-0.0982	0.1263	0.1684
0.1964	0.2245	-0.2526	0.2947	-0.3087	0.3929	0.4069
0.4209	0.4350	0.4630	-0.4771	0.5472	0.6595	0.7437
-0.7437	-0.7577	-0.8138	-0.8138	-0.8980	1.080	-1.305
2.147	-2.666	-3.143				

Taking the data at face value as observations from  $\mathbf{N}(\mu_i, \sigma^2)$  (i.e. ignoring their structured origin in the  $2^5$  experiment) Dempster and Rosner seek outliers using the ideas above. They tabulate for  $k = 1(1)5$  the  $\tilde{\mu}_1$  and  $p_i$  and  $\tilde{\sigma}_k$  and  $\pi_k$ . The results are reproduced as Table 8.1 and provide rather compelling evidence for the three observations with largest absolute value being discordant outliers. We notice in particular how  $\pi_k$  builds up to 0.9247 at  $k=3$ , dropping to 0.3311 at  $k=4$ , and the maintenance of anomalously high  $\mu_i$  and  $p_i$  for the last three observations for  $k=3$  and  $k=4$ .

Table 8.1 Bayesian outlier analysis of the Daniel data (reproduced by permission of Academic Press)

$x_i$	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
	$\tilde{\mu}_i$	$p_i$								
0.0000	0.0000	0.0019	0.0000	0.0015	0.0000	0.0016	0.0000	0.0127	0.0000	0.0160
0.0281	0.0001	0.0019	0.0000	0.0015	0.0000	0.0016	0.0004	0.0127	0.0004	0.0160
0.0561	0.0001	0.0019	0.0001	0.0015	0.0001	0.0016	0.0007	0.0128	0.0009	0.0161
0.0842	0.0002	0.0019	0.0001	0.0015	0.0001	0.0016	0.0011	0.0129	0.0014	0.0162
0.0982	0.0002	0.0019	0.0001	0.0015	0.0002	0.0016	0.0013	0.0129	0.0016	0.0163
0.1263	0.0002	0.0019	0.0002	0.0015	0.0002	0.0016	0.0016	0.0131	0.0021	0.0165
0.1684	0.0003	0.0019	0.0002	0.0015	0.0003	0.0016	0.0022	0.0133	0.0029	0.0169
0.1964	0.0004	0.0019	0.0003	0.0015	0.0003	0.0017	0.0027	0.0136	0.0034	0.0173
0.2245	0.0004	0.0019	0.0003	0.0015	0.0004	0.0017	0.0031	0.0139	0.0040	0.0177
0.2526	0.0005	0.0019	0.0004	0.0015	0.0004	0.0017	0.0036	0.0142	0.0046	0.0183
0.2947	0.0006	0.0020	0.0005	0.0015	0.0005	0.0018	0.0044	0.0148	0.0057	0.0192
0.3087	0.0006	0.0020	0.0005	0.0016	0.0005	0.0018	0.0046	0.0150	0.0060	0.0195
0.3929	0.0008	0.0020	0.0006	0.0016	0.0007	0.0019	0.0065	0.0166	0.0087	0.0222
0.4069	0.0008	0.0020	0.0007	0.0016	0.0008	0.0019	0.0069	0.0170	0.0092	0.0227
0.4209	0.0009	0.0020	0.0007	0.0017	0.0008	0.0019	0.0073	0.0173	0.0098	0.0233
0.4350	0.0009	0.0021	0.0007	0.0017	0.0009	0.0020	0.0077	0.0177	0.0104	0.0239
0.4630	0.0010	0.0021	0.0008	0.0017	0.0009	0.0020	0.0086	0.0185	0.0117	0.0252
0.4771	0.0010	0.0021	0.0008	0.0017	0.0010	0.0021	0.0090	0.0189	0.0124	0.0260
0.5472	0.0012	0.0022	0.0010	0.0018	0.0012	0.0022	0.0118	0.0215	0.0166	0.0304
0.6595	0.0015	0.0023	0.0013	0.0020	0.0017	0.0026	0.0181	0.0275	0.0270	0.0410
0.7437	0.0018	0.0025	0.0016	0.0022	0.0022	0.0030	0.0253	0.0341	0.0398	0.0535
0.7437	0.0018	0.0025	0.0016	0.0022	0.0022	0.0030	0.0253	0.0341	0.0398	0.0535
0.7577	0.0019	0.0025	0.0017	0.0022	0.0023	0.0031	0.0268	0.0354	0.0425	0.0561
0.8138	0.0021	0.0026	0.0019	0.0024	0.0028	0.0034	0.0340	0.0417	0.0559	0.0687
0.8138	0.0021	0.0026	0.0019	0.0024	0.0028	0.0034	0.0340	0.0417	0.0559	0.0687
0.8980	0.0025	0.0028	0.0024	0.0027	0.0037	0.0041	0.0492	0.0547	0.0865	0.0963
1.0804	0.0037	0.0034	0.0038	0.0035	0.0068	0.0063	0.1193	0.1104	0.4076	0.3773
1.3049	0.0059	0.0046	0.0071	0.0054	0.0163	0.0125	0.4414	0.3382	1.0523	0.8064
2.1468	0.0507	0.0236	0.1482	0.0690	1.9957	0.9296	2.1321	0.9932	2.1443	0.9989
2.6659	0.3029	0.1136	2.3677	0.8882	2.6537	0.9954	2.6650	0.9997	2.6658	1.0000
3.1430	2.5128	0.7995	3.1055	0.9881	3.1419	0.9996	3.1429	1.0000	3.1430	1.0000
$\bar{\sigma}_k$	0.8650		0.7102		0.5856		0.5572		0.5221	
$\pi_k$	0.7995		0.8762		0.9247		0.3311		0.1825	

A somewhat similar, if more specific, application of Bayesian methods to the ‘detection of spuriousity’ is described by Guttman (1973b). Adopting a slippage-type alternative hypothesis to describe the occurrence of outliers, Guttman produces a procedure for determining whether or not a ‘spurious observation’ has occurred in the data. This interest in identification and discordancy can be contrasted with the work of Box and Tiao (1968, see also Section 8.1.2) who employ a somewhat similar model to investigate the accommodation issue: the way in which estimates of basic parameters are influenced by the presence of outliers.

Guttman concentrates on a set of independent normal observations  $x_1, x_2, \dots, x_n$  arising, in the absence of ‘spuriousity’, from a common normal distribution,  $N(\mu, \sigma^2)$ . Under the alternative model, one observation comes

from  $\mathbf{N}(\mu + a, \sigma^2)$ . This location-shift model is a special case of *model A* of Ferguson (1961a); we shall see that Box and Tiao (1968) deal with the dispersion-shift analogue (*model B*). It is assumed that any of the observations is equally likely to be the *one* that is spurious (see Section 2.3 for a discussion of the restriction to one, or at least very few, possible spurious observations) and Guttman offers a ‘succinct description’ of his model in terms of the likelihood

$$P\{\mu, a, \sigma^2 | \mathbf{x}\} = \frac{1}{n} (2\pi\sigma^2)^{-n/2} \sum_{j=1}^n \left\{ \exp \left[ -\frac{1}{2\sigma^2} (x_j - \mu - a)^2 \right] \times \exp \left[ -\frac{1}{2\sigma^2} \sum_{i \neq j} (x_i - \mu)^2 \right] \right\}. \quad (8.1.6)$$

Note that this is *not* equivalent to a mixture-type model where each observation has some small probability of arising from  $\mathbf{N}(\mu + a, \sigma^2)$  or a larger complementary probability of arising from  $\mathbf{N}(\mu, \sigma^2)$ .

Adopting a non-informative prior distribution for  $\mu$  and  $\sigma^2$  with density proportional to  $\sigma^{-2}$  the posterior distribution of  $(\mu, \sigma^2, a)$  is obtained. It has probability density function proportional to

$$\sigma^{-(n+2)} \sum_{j=1}^n \exp \left[ -\frac{1}{2\sigma^2} \left\{ A_{-j} + \frac{n+1}{n} \left[ a - \frac{n}{n-1} (x_j - \bar{x}) \right]^2 + n \left( \mu - \bar{x} + \frac{a}{n} \right)^2 \right\} \right] \quad (8.1.7)$$

where  $A_{-j} = \sum_{i \neq j} (x_i - \bar{x}_{-j})^2$  with  $\bar{x}_{-j} = (n-1)^{-1} \sum_{i \neq j} x_i$ . Integrating out  $\mu$  and  $\sigma^2$  the posterior distribution of  $a$  is obtained. It has probability density function proportional to

$$\sum_{j=1}^n \left\{ A_{-j} + \frac{n-1}{n} \left[ a - \frac{n}{n-1} (x_j - \bar{x}) \right]^2 \right\}^{-(n-1)/2} \quad (8.1.8)$$

which can be regarded as a weighted combination of densities of the Student’s *t* type.

This latter characteristic enables the posterior mean and variance of  $a$  to be obtained as

$$\mu_a = \frac{n}{n-1} \sum_{j=1}^n c_j (x_j - \bar{x}) \quad (8.1.9)$$

$$\sigma_a^2 = \left( \frac{n}{n-1} \right)^2 \left\{ \sum_{j=1}^n c_j (x_j - \bar{x})^2 - \left[ \sum_{j=1}^n c_j (x_j - \bar{x}) \right]^2 + \frac{(n-1)}{n(n-4)} \sum c_j A_{-j} \right\} \quad (8.1.10)$$

where

$$c_j = (A_{-j})^{-(n-2)/2} / \sum_{j=1}^n (A_{-j})^{-(n-2)/2}. \quad (8.1.11)$$

The attribution of spuriousity to a member of the sample is approached in terms of the posterior distribution of  $a$ , and in particular of the values of  $\mu_a$  and  $\sigma_a^2$ . If the weights  $c_i$  are roughly equal (to  $n^{-1}$ ) we have little evidence of spuriousity; a rough argument is given to support the attribution of spuriousity to an observation  $x_i$  whose weight  $c_i$  exceeds

$$\frac{1}{n} + \frac{2}{n} \sqrt{\left(\frac{n-1}{n+1}\right)} \quad (\text{for } n \geq 3).$$

Alternatively, Bayesian confidence (credibility) intervals for the parameter  $a$  yield criteria for ascribing spuriousity expressed in terms of the distribution function of the  $t$ -distribution with  $(n-2)$  degrees of freedom.

Guttman illustrates his recommendations by reference to some sets of simulated data.

The approach can be immediately extended to multivariate data  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  arising from  $\mathbf{N}(\boldsymbol{\mu}, V)$  if there is no spurious observation, with spuriousity manifest in a single observation from  $\mathbf{N}(\boldsymbol{\mu} + \mathbf{a}, V)$ . Again it is through the posterior distribution of  $\mathbf{a}$  that we seek to detect a discordant outlier. The detection criteria again revolve around the values taken by a set of weights attached, in the posterior distribution of  $\mathbf{a}$ , to the separate observations. It is interesting to note that the implicit concept of extremeness used to detect the outlier is again expressible in terms of the distance metric, or scatter-ratio, discussed in Chapter 6.

### 8.1.2 Bayesian accommodation of outliers

We have maintained a distinction between two basic attitudes in the study of outliers: identification and rejection on the one hand, accommodation on the other. This distinction also appears in the use of Bayesian methods and we find several contributions of the accommodation type where methods of estimation or testing of parameters in a model are proposed which are robust against the presence of outliers.

A major contribution in this category is the work of Box and Tiao (1968) who consider a Bayesian analysis of the linear model when outliers may be present in the data. They particularize their results to a linear model where the error terms arise as independent observations from normal distributions with zero mean but where slippage in the variance may have occurred for a limited number of observations (a more structured form of the Ferguson, 1961a, *model B*).

We start by considering proposals for the general linear model where the observation vector  $\mathbf{x}$  has the form

$$\mathbf{x} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{8.1.12}$$

with  $\boldsymbol{\theta}$  a  $p \times 1$  vector of parameters,  $\mathbf{A}$  an  $n \times p$  design matrix and  $\boldsymbol{\varepsilon}$  a  $p \times 1$  vector of independent random errors. It is supposed that the individual

errors may have arisen from one or other of two distributions: a basic distribution  $f(\varepsilon | \xi_1)$  or an alternative (outlier generating) distribution  $g(\varepsilon | \xi_2)$ . Interest centres on drawing inferences about  $\theta$ , with the parameter sets  $\xi_1$  and  $\xi_2$  regarded as nuisance parameters. Attribution of the individual errors  $\varepsilon_i$  to  $f(\cdot)$  or to  $g(\cdot)$  is not triggered by the corresponding observed  $x_i$ . Indeed the structure of the model (8.1.12) may render intuitive detection of outliers impossible (see Chapter 7). Instead events  $a_{(k)}$  are defined under which a specific  $k$  of the  $\varepsilon_i$  come from  $g(\cdot)$ , the remainder from  $f(\cdot)$ , and inferences employ the corresponding likelihood which is made up of  $2^n$  components  $P(a_{(k)} | \theta, \xi_1, \xi_2)$  corresponding with all possible  $a_{(k)}$ . A general theory is developed leading to a formal expression for the posterior distribution of  $\theta$  based on general prior distributions  $\{p^{(k)}\}$  for  $\{a_{(k)}\}$  and  $p(\theta, \xi_1, \xi_2)$  for  $(\theta, \xi_1, \xi_2)$ .

Tangible expression is given to this in terms of the above normal-error model with possible scale-shift. Thus the errors  $\varepsilon_i$  arise either from  $N(0, \sigma^2)$  or from  $N(0, b\sigma^2)$ . A particular case is studied where each  $\varepsilon_i$  arises with probability  $(1 - \lambda)$  from  $N(0, \sigma^2)$  or with probability  $\lambda$  from  $N(0, b\sigma^2)$ . It is assumed that  $b$  is prescribed (presumably  $b > 1$  to make sense of the notion of outliers) and that  $(\theta, \log \sigma)$  is independent, uniform, *a priori*. The posterior distribution of  $\theta$  is exhibited in the form of a  $p$ -dimensional multivariate  $t$ -distribution. Marginal distributions of the components  $\theta_i$  are also derived.

It is interesting to note that the particular explanation adopted for the way in which the errors arise from  $N(0, \sigma^2)$  or  $N(0, b\sigma^2)$  implies a mixture-type (rather than slippage-type) model for outlier generation. Box and Tiao point out that a modified approach making formal recognition of the ‘mixing’ leads to the same results as were obtained under their wider formulation where the likelihood consists of contributions from each of the  $2^n$  configurations of error source.

The method is illustrated for estimation of a single mean,  $\mu$ . We have  $x_1, x_2, \dots, x_n$  as independent observations each arising with probability  $(1 - \lambda)$  from  $N(\mu, \sigma^2)$  or with probability  $\lambda$  from  $N(\mu, b\sigma^2)$ , with  $\lambda$  and  $b$  prescribed and  $\sigma^2$  unknown. The posterior distribution of  $\mu$  (adopting uniform, independent, prior distributions for  $\mu$  and  $\log \sigma$ ) turns out to have the form

$$\pi(\mu | \mathbf{x}) = \sum_{(k)} w_{(k)} \frac{(n - \phi k)^{\frac{1}{2}}}{s_{(k)}} f_{n-1} \left[ \frac{\mu - \tilde{\mu}_{(k)}}{s_{(k)} / \sqrt{(n - \phi k)}} \right]. \quad (8.1.13)$$

The summation in (8.1.13) ranges over all events  $a_{(k)}$ , and the weights  $w_{(k)}$  are proportional to

$$\left( \frac{\lambda}{1 - \lambda} \right)^k b^{-k/2} \left( \frac{n}{n - k} \right)^{\frac{1}{2}} \left( \frac{s_{(k)}^2}{s^2} \right)^{-\frac{1}{2}(n-1)}, \quad (8.1.14)$$

with  $\bar{x}$ ,  $s^2$  the sample mean and variance, respectively,  $\bar{x}_{(k)}$  the mean of those  $x_i$  attributed to  $N(0, b\sigma^2)$  under  $a_{(k)}$  and

$$\left. \begin{aligned} \phi &= 1 - b^{-1} \\ \tilde{\mu}_{(k)} &= \bar{x} - \frac{k\phi}{n - k\phi} (\bar{x}_{(k)} - \bar{x}) \\ (n-1)s_{(k)}^2 &= (n-1)s^2 - \phi \left\{ \sum' (x_i - \bar{x})^2 + \frac{\phi k^2}{(n-\phi k)} (\bar{x}_{(k)} - \bar{x})^2 \right\} \end{aligned} \right\} \quad (8.1.15)$$

where  $\sum'$  implies summation over all  $x_i$  attributed to  $N(0, b\sigma^2)$  under  $a_{(k)}$ . The function  $f_{n-1}(\cdot)$  is the probability density function of the  $t$ -distribution, so that the posterior distribution of  $\mu$  is a weighted average of  $2^n$  scaled  $t$ -distributions with  $n-1$  degrees of freedom.

Determination of (8.1.13) is most tedious. Proposals are made by Box and Tiao for easing the load which make the exercise feasible at least for moderate  $n$  (up to 20 or so) and small  $\lambda$ . The method is illustrated on a classical set of data due to Darwin on heights of plants quoted by Fisher (1960, page 37) and examined from an alternative viewpoint to the present one by Box and Tiao (1962). There they attempted to accommodate two lower outliers by using a broader model than that previously employed. In the current context the posterior distribution of the mean  $\mu$ , allowing for possible discordant outliers, is exhibited in relation to extreme alternatives that there are no outliers or that the two outliers are discordant and genuinely arise from the alternative model  $N(\mu, 25\sigma^2)$ . A value for  $\lambda$  of 0.05 is arbitrarily employed, although efforts are made to study the sensitivity of the analysis to the choice of values of  $\lambda$  and  $b$ . Within the limited study it appears that the posterior mean and standard deviation of  $\mu$  are far more sensitive to the value of  $\lambda$  than to the value of  $b$ .

In passing we should recall the use of Bayesian methods in the proof of optimality properties of slippage procedures based on the Paulson type of multiple decision approach (see Chapter 5 for some details).

The *exchangeable* model for outliers has been used by Kale, Sinha, and Veale in developing classical methods for estimating or testing the mean of an exponential distribution, where outliers may be present in the data. (Section 4.4 presented some details of this work.) In the same applications context of life-testing and reliability, with exponentially distributed lifetimes, Sinha (1972, 1973b) has considered corresponding Bayesian methods.

Sinha (1972) considers  $n$  independent observations  $x_1, x_2, \dots, x_n$  where all but one ( $x_i$ ) arise from an exponential distribution with p.d.f.

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta} \quad (\theta > 0) \quad (8.1.16)$$

whilst  $x_i$  arises from an exponential distribution with p.d.f.  $f(x, \theta/\eta)$  where  $0 < \eta \leq 1$ . The index  $i$  is assumed, *a priori*, to be equally likely to take any of the values  $1, 2, \dots, n$ . If (8.1.16) is the basic lifetime distribution whilst  $f(x, \theta/\eta)$  is an inconvenient intrusion representing perhaps an unidentified alien component in the sample, a quantity of basic interest in reliability is the survivor function

$$R_\theta(\tau) = P(X \geq \tau) = e^{-\tau/\theta}$$

and we might wish to estimate this free from serious influence of  $x_i$ . In the absence of contamination in the data (i.e. if  $\hat{\eta} = 1$ ) a desirable estimator is

$$\tilde{R}(\tau) = \begin{cases} [1 - \tau/(n\bar{x})]^{n-1} & (\tau \leq n\bar{x}), \\ 0 & (\text{otherwise}). \end{cases} \quad (8.1.17)$$

$\tilde{R}(\tau)$  is the uniform minimum variance unbiased estimator of  $R(\tau)$ . Sinha examines the variance  $\text{var}[\tilde{R}(\tau)]$ . Clearly this depends on both  $\tau$  and  $\theta$  but this joint dependence takes a simple form in that  $\text{var}[\tilde{R}(\tau)]$  is a function of the ratio  $\tau/\theta$ , and we will denote it  $V(\tau/\theta)$ .

One aspect of the influence of an outlier on estimation of  $\tilde{R}(\tau)$  is the effect of  $x_i$  on  $E[\tilde{R}(\tau)]$  and on the mean square error of  $\tilde{R}(\tau)$ . Both of these are also functions of  $\tau/\theta$ , and of  $\eta$ . We will denote them  $\mu_\eta(\tau/\theta)$  and  $MSE_\eta(\tau/\theta)$ . Their explicit forms are intractable, but Sinha derives lower and upper bounds for each of them.

An alternative approach to investigating  $\mu_\eta(\tau/\theta)$  and  $MSE_\eta(\tau/\theta)$  for fixed  $\eta$  is to determine the distribution of the basic statistic  $n\bar{x}/\theta$  arising from some prescribed prior probability distribution for  $\eta$  and thence to set bounds on the mean and  $MSE$  of  $\tilde{R}(\tau)$ . No prior distribution is assigned to  $\theta$ ; the approach is accordingly termed ‘semi-Bayesian’. For convenience a prior Beta distribution is adopted for  $\eta$ , with p.d.f.

$$p(\eta) \propto \eta^{p-1}(1-\eta)^{q-1} \quad (p, q > 0). \quad (8.1.18)$$

Denoting the posterior mean and  $MSE$  of  $\tilde{R}(\tau)$  by  $\mu_{p,q}(\tau/\theta)$  and  $MSE_{p,q}(\tau/\theta)$  respectively, Sinha shows that

$$\frac{p}{p+q} e^{-\tau/\theta} {}_2F_2(1, q, n, p+q+1, \tau/\theta) \leq \mu_{pq}(\tau/\theta) \leq e^{-\tau/\theta} \quad (8.1.19)$$

and

$$\begin{aligned} \left( \frac{p}{p+q} \right) {}_2F_2(1, q, n, p+q+1, \tau/\theta) k(\tau/\theta) - e^{-2\tau/\theta} &\leq MSE_{p,q}(\tau/\theta) \\ &\leq k(\tau/\theta) - \frac{2p}{p+q} e^{-2\tau/\theta} {}_2F_2(1, q, n, p+q+1, \tau/\theta) + e^{-2\tau/\theta} \end{aligned} \quad (8.1.20)$$

where  ${}_2F_2( )$  is the hypergeometric function, and

$$k(\tau/\theta) = \{(\tau/\theta)^n e^{-\tau/\theta} / \Gamma(n)\} J \quad (8.1.21)$$

with

$$J = \int_0^\infty [e^{-u\tau/\theta} u^{2n-2}/(1+u)^{n-1}] du. \quad (8.1.22)$$

In principle these results give some indication of how the sampling behaviour of  $\tilde{R}(\tau)$  is affected by the presence of a single outlier, but their complexity militates against any simple interpretation of this influence, the prior probability structure for  $\eta$  is arbitrary and there seems no good reason why we should maintain  $\tilde{R}(\tau)$  as an estimator when an outlier is present. Indeed Kale and Sinha (1971) proposed the use of  $s = \sum_1^{n-1} x_{(j)} + x_{(n-1)}$  instead of  $n\bar{x}$ , when a single spurious observation is present, since this is most likely to correspond with  $x_{(n)}$ . Accordingly

$$\tilde{R}^*(\tau) = \begin{cases} (1 - \tau/s)^{n-1} & (\tau \leq s) \\ 0 & (\text{otherwise}) \end{cases} \quad (8.1.23)$$

has some appeal as an estimator of  $R(\tau)$ , but no Bayesian analysis of  $\tilde{R}^*(\tau)$  is offered.

Sinha (1973b) considers a fuller Bayesian treatment employing again the prior distribution (8.1.18) for  $\eta$ , and three possible families of prior distributions for  $\theta$  (independent of  $\eta$ ). He derives the Bayes estimators of  $\eta$ , of  $\theta$  (the mean life-time), and of the survivor function  $R(\tau)$ . The forms are again highly complicated, and specific to the chosen prior structures. No simple qualitative interpretation of the influence of the outlier is offered, nor does it seem feasible.

An alternative basic exponential model with p.d.f.

$$g(x, \mu) = \exp[-(x - \mu)] \quad (x > \mu) \quad (8.1.24)$$

where the outlier arises from an exponential distribution with p.d.f.  $g(x, \mu + \delta)$  ( $x > \mu + \delta, \delta > 0$ ) is also considered in Sinha (1972) and Sinha (1973b).

Finally in this brief review of some Bayesian and 'semi-Bayesian' methods for studying outliers we must mention a proposal by Lingappaiah (1976) for estimating the shape parameter in a different wide-ranging family of distributions (including the Weibull and gamma) where several outliers may be present arising from different members of the same family of distributions. The basic model has p.d.f.

$$f(x) = bx^{\alpha-1}\beta^{\alpha/b} \exp(-\beta x^b)/\Gamma(\alpha/b) \quad (x > 0). \quad (8.1.25)$$

In a sample of size  $n$  we contemplate the prospect that  $k$  of the  $n$  observations arise from (8.1.25) with  $\beta$  replaced by  $\theta_i\beta$  ( $i = 1, 2, \dots, k$ ;  $0 < \theta_i \leq 1$ ).

Adopting an exponential prior distribution for  $\beta$  and Beta prior distributions for the  $\theta_i$ , and assuming for fixed  $k$  that the set of outliers is equally likely to be any set of the  $k (< n)$  observations, the posterior distribution of  $\beta$  is obtained for fixed  $(\alpha, b)$ . The Bayes estimator of  $\beta$  is also derived.

Particular cases are derived for  $k = 1$  and where (8.1.25) reduces to a Weibull, gamma, or exponential distribution.

This contribution by Lingappaiah exemplifies the impracticality of much of the Bayesian contribution to the study of outliers in the literature to date! Notwithstanding the fundamental obstacles confronting a Bayesian approach in this branch of statistics, some offerings have been made at the Bayesian altar. But they are far from acceptable in terms of the arbitrary mix of Bayesian and sampling theoretic components, uniform distributions of the outliers over the set of observations denying the principle of 'surprise' in the identification of outliers, expedient (unjustified and undiscussed) choice of prior distributions for an arbitrary subset of the basic parameters in the model, and formal and unmanageable results with little interpretation or application. Not all such unsatisfactory components are present at the same time, but some exist in almost all the Bayesian contributions and we must conclude that very much remains to be done to achieve a convincing advance in the Bayesian study of outliers.

## 8.2 NON-PARAMETRIC METHODS

Non-parametric procedures for identifying outliers, testing their discordancy or rendering them uninfluential in a statistical analysis of the bulk of the data, have been presented in a variety of contexts. The original approach to the slippage problem (Mosteller, 1948) was non-parametric and generated a flood of refinements or modifications. Chapter 5 has discussed these in some detail.

Non-parametric methods for the analysis of data arising from designed experiments extends the slippage problem in a more structured form; looking for effects of different factors in the designed experiment can be interpreted (to a degree) as identifying extreme, or outlying, sub-samples with special intrinsic interest. The outlying sub-samples are not an inconvenience, they are the very manifestations we seek in order to express the statistical import of the data. But in designed experiments it may happen that individual observations, rather than sub-samples corresponding with certain factor levels, are anomalous. Such outliers have a different role; they serve to cloud the treatment effects we are investigating and need either to be rejected on a sound statistical basis or accommodated with minimal import in a robust analysis of the data from the standpoint of principal interest. Of course, such outliers may not be immediately apparent on simple inspection. They are extreme only relative to some peer group, e.g. the members of the sub-sample corresponding with a certain factor level.

Thus we are seeking to express anomalies of pattern, or of residuals, in the data.

Chapter 7 has considered this problem in some detail, and its corresponding form in regression or time-series analyses. Many appropriate procedures are non-parametric in form, based on signs or ranks or inversions. An interesting paper on the basic philosophy of identifying outliers in designed experiments, in the sense of disruptors of overall pattern, is by Bross (1961). It employs the notion of inversion in the data as a principal basis for identifying outliers.

Non-parametric proposals for the more fundamental outlier problem—identification and discordancy testing in a single unstructured univariate sample—are less in evidence.

Some contributions have been made by Walsh [1950 (and correction, 1953), 1959, 1965]; the accommodation issue is briefly considered by Walsh and Kelleher (1973). We conclude this chapter by considering some of Walsh's proposals, and by posing a few questions about the value of non-parametric and distribution-free methods in the context of outliers.

Some non-parametric tests of outliers are discussed and described by Walsh (1965, in the second volume of his book on non-parametric statistics), including the proposals in Walsh (1950). For one of the proposed tests the basic model assumes that the data consist of independent observations from common symmetric distributions with median  $\phi$ ; the alternative model postulates upward slippage in location of a prescribed number,  $k$ , of the distributions. The test statistic employs the values of the  $k$  extremes  $x_{(n-k+1)}, \dots, x_{(n)}$ . It has a complicated form, is laborious to operate, has unknown but bounded significance level, and is useful only for  $k > 4$  (at least four upper outliers).

The test operates as follows.  $T$  is some integer less than or equal to  $k$ , and  $\{u_i\}$  and  $\{v_i\}$  are chosen sequences of  $T$  numbers monotone increasing in  $i$  with  $u_T = k$ . We reject the basic model  $H_0$  if

$$S(\alpha) = \min_{1 \leq i \leq T} [x_{(n+1-u_i)} - x_{(v_i)}] - 2x_{(w(\alpha))} > 0 \quad (8.2.1)$$

where

$$\alpha = P \left\{ \min_{1 \leq i \leq T} [x_{(n+1-u_i)} - x_{(v_i)}] > 2\phi \mid H_0 \right\} \quad (8.2.2)$$

and  $w(\alpha)$  is the smallest integer such that

$$P\{x_{(w(\alpha))} < \phi \mid H_0\} \leq \alpha. \quad (8.2.3)$$

The significance level of the test is bounded above by  $2\alpha$ , and approaches  $\alpha$  for  $n$  sufficiently large provided certain additional assumptions about the distributions of the  $x_{(i)}$  are satisfied. The restrictions of large  $n$ , at least 4 outliers, and computational complexity severely limit the usefulness of the test.

Obvious modifications produce tests for  $k$  lower outliers or two-sided tests.

Another type of test for large samples is proposed by Walsh (1959). The distributions need not now be symmetric; the alternative model is again of the location slippage type. The asymptotic forms of the moments, and distributional behaviour, of order statistics are invoked to produce tests for  $k$  upper outliers (or for  $k$  lower, or a corresponding two-sided test). The one-sided tests for  $k$  upper outliers have rejection criterion of the form

$$x_{(n+1-k)} - [1 + A_n(\alpha)]x_{(n-k)} + A_n(\alpha)x_{(n+1-s)} > 0 \quad (8.2.4)$$

where  $s$  is the largest integer less than  $k + \sqrt{2n}$  and  $A_n(\alpha) > 0$  depends on  $n$  and on an upper bound  $\alpha$  for the significance level of the test. The test statistic is chosen as a particular case of the more general form

$$S = x_{(n+1-k)} - (1 + A)x_{(r)} + Ax_{(n+1-s)}$$

to meet certain requirements about the test behaviour. Given  $k$  and  $n$ ,  $r = n - k$  effects an approximate large-sample minimization of  $\text{var}(S)$  under the basic model which postulates homogeneity of distribution for the sample observations. Subject to this,  $s$  and  $A$  are determined by the additional assumptions that  $A > 0$  and that

$$E(S) = K\sqrt{\text{var}(S)}[1 + o(1)] \quad (8.2.5)$$

under the basic model, for prescribed  $K$ . Chebyshev's inequality implies

$$P(S < 0) \approx 1/K^2 \quad (8.2.6)$$

so that  $1/K^2$  is an estimate of the significance level of the test. Specifically, we are lead to the prescription

$$s = k + [\sqrt{2n}] \quad (8.2.7)$$

and

$$A = \frac{1 + K\sqrt{([\sqrt{2n}] - K^2)/([\sqrt{2n}] - 1)}}{[\sqrt{2n}] - K^2 - 1}. \quad (8.2.8)$$

The test is considered applicable for large enough  $n$ : namely where

$$\sqrt{2n} > K^2 + 1. \quad (8.2.9)$$

A fairly detailed discussion of the form and properties of this test is given by Walsh (1959).

Such a test inevitably suffers from a variety of difficulties of conception and application. Determination of its precise form is again rather tedious: the structure (8.2.5) and rough probabilistic assessment provided by Chebyshev's inequality must seriously limit any assessment of test properties and applicability; the need to specify  $K$  (and  $k$ ) introduces an unreasonable degree of arbitrariness; limitation to very large  $n$  (as is often implied) takes

us away from the most contentious and important area of application of outlier tests. Finally, we need to consider the problem of the likely robustness or power of non-parametric tests in the particular context of outlier studies. We return to this shortly.

One further contribution of non-parametric methods for outliers appears in the work of Walsh and Kelleher (1973). They consider the unbiased estimation of the mean and variance of a continuous distribution from which a set of  $n$  independent observations is purported to arise, but where there is the possibility of some upper and lower outliers. The numbers of upper and lower outliers are prescribed, and small in relation to  $n$ . The assumptions in the work are similar to those employed in Walsh (1959).

It is a feature of non-parametric procedures that we adopt a minimum of distributional assumptions about the data-generating mechanism. Such procedures can have relatively low power in comparison with procedures specifically geared to a particular detailed parametric model. If such a model can be justified we would often wish to employ methods tailored to it. Otherwise, non-parametric methods are appealing in their lack of commitment to a model—their ubiquity or robustness. But this type of appeal has an element of delusion in the outlier context. Outliers are ‘atypical’ observations—they impress themselves on us by appearing to be unrepresentative of the overall sample data. With no knowledge (or assumptions) about the general distributional structure of the data-generation process we have no grounds for ‘surprise’, nothing ‘typical’ against which to ascribe ‘atypicality’. The *macrolepidoptera* light-trap data at the end of Section 1.2 illustrate this well. It is only by considering (at least informally) the way in which the data might reasonably have been generated that we have any basis for examining the possibility of discordant outliers. There must always be the possibility of a homogeneous explanation of the values in any sample.

In its most extreme form a non-parametric approach makes no assumptions about the basic data-generating mechanism. At this level it seems a contradiction to seek to investigate outliers. Broad specifications such as symmetry of the basic distribution, or of location-slippage explanations of outliers, raise some prospect for outlier study but seem bound to be highly speculative in their conclusions. If such specifications are as much as we dare contemplate, then perhaps we have no alternative but to accept the highly limited assessment of outliers yielded by a non-parametric approach. But more than in almost any other area of statistical enquiry, the study of outliers hinges on as precise a model formulation as is feasible. To deliberately abandon the model, by seeking non-parametric (or distribution-free) methods in some broad aim of robustness, smacks of throwing out the bathwater before the baby has even been immersed.

## CHAPTER 9

### *Perspective*

We have covered a lot of ground in this study of outliers. We hope that in the process we have both armed the experimental scientist with a range of useful techniques and provided some food for thought for the professional statistician in clarifying basic principles and indicating some of the methodological gaps.

The outlier problem seems to be arousing more interest today than it has ever done, in spite of its long history. The pages of many statistical journals frequently contain new contributions and the reader who wants to keep abreast of the subject would find it profitable to keep a regular eye on journals such as *Technometrics*, the *Journal of the American Statistical Association*, and *Applied Statistics*.

Current research activity seems to centre largely on informal methods for outliers in highly structured situations (with an emphasis on computer application and graphical display) and on refinements of understanding of familiar univariate single-sample procedures. On both fronts useful results abound, but much remains to be done. Some of the directions for the study of univariate single-sample procedures are clear. We need to know much more about performance criteria of tests of discordancy: this embodies the need for a greater understanding of distributional theory elements. Many of the outlier-generating models warrant further investigation; the interactions between, and respective relevances of, the different models have not been appropriately examined. The whole question of block and consecutive testing procedures needs a thorough investigation such as it has never yet received. These important practical and conceptual matters are also likely to involve some challenging mathematical work.

But when all is said and done, the major problem in outlier study remains the one that faced the very earliest workers in the subject—*what is an outlier?* We have taken the view that the stimulus lies in the subjective concept of surprise engendered by one, or a few, observations in a set of data: that this surprise initiates an investigation of the statistical propriety (or influence) of the detected outliers. We have noted that surprise is not always

immediate; for example, in multivariate or highly structured data an explicit detection process may be required to reveal the outliers (the 'surprising' observations). However, the chain of operations remains the same. We find an observation surprising, we then proceed to investigate it from the statistical viewpoint, and conclude by rejecting it, welcoming it, or accommodating it. If accommodation is the aim, the analysis proceeds a further stage, to a treatment of the overall data set with allowance for the presence of the outlier.

Other viewpoints exist, however. We must face the fact that such a sequence of operations was, and is, not universally accepted as reasonable. From the Bayesian viewpoint, there are philosophical obstacles to the pre-processing of selected observations from a larger data set (though less serious objections to the subjectivity of the outlier detection process, at least if formalized). The statistician adopting a more classical approach may find the subjective element not to his taste. Insofar as he is prepared to countenance the rejection, for example, of individual members of a sample, he may require this to be done 'objectively' by a routine procedure carried out regularly and indiscriminately on every similar set of data that arises.

Concerning such fundamental objections, it is indeed not clear that the implications of the subjective detection of outliers are, or even can be, appropriately measured and reflected in the fuller statistical analysis of a set of data. Of course, proper concern for the construction of the outlier model can go a long way to resolve this problem.

Another difficulty which we have to keep firmly in view is that of *masking*. Its manifestation in single univariate samples is straightforward. But in more complex situations it needs to be viewed in a much wider context. In a designed experiment it remains true that one outlier may not be declared discordant because of the masking effect of another. In another respect the very presence of outliers may be masked by particularly strong real effects or by breakdowns in the conventional assumptions, such as normality, homoscedasticity, additivity, and so on. In reverse we may falsely attribute idiosyncratic facets of the behaviour of the data to such breakdowns, whilst in reality they truly reflect the presence of individual discordant values.

Multivariate or highly structured data further highlight the 'subjectivity' versus 'objectivity' argument. We have remarked above on the more nebulous nature of 'surprise' in such situations. Explicit detection procedures are needed for outliers here, and almost inevitably data analysis methods will be computerized, perhaps augmented by graphical display. The modern preoccupation with thoughtless do-it-yourself computer packages has a spill-over effect. We will not be surprised at pleas for outlier packages which relieve the analyst of any responsibility. But computers are not easily taught to be surprised. The concept is a human one. It can only to a limited extent be translated into a mechanized form. On the extreme viewpoint that the element of surprise must be central to the study of outliers, so, it would

follow, must the individual continue to shoulder the major burden of outlier detection through his personal and regular intervention in any data-screening procedure. Of course we can, and should, react to summarized or graphically presented data from the computer to assist in the detection of outliers. Of course we can build into the computer formal procedures for testing discordancy. But we do this to a large extent at the sacrifice of the subjective 'surprise' stimulus. It is too much to expect to be able to teach the computer just what it is that would engender the surprise in you (or me) necessary as the precursor of a test of discordancy.

The very fact that we cannot formalize the 'surprise' is only part of the problem. Even on an 'objective' approach to outlier processing the computer may have its disadvantages to be set against its indispensability as a digester and presenter of large-scale data. In a highly structured situation such as a designed experiment the notion of an outlier remains so primitive (perhaps an extreme residual, perhaps some sort of unexpected break in pattern) that any total replacement of situation-specific analysis by depersonalized routine computer processing could inhibit the development of clearer understanding of outliers in such areas.

What of the future? There are fashions in statistics as in all things. Outliers were out of fashion for long enough. They exist as part of the experimentalist's reality; they are part of the analyst's inescapable responsibility. Surely the professional statistician cannot reasonably withhold his contribution to outlier methodology on the grounds that he is not too sure how to define an outlier?

## APPENDIX

### *Statistical Tables*

Table I Critical values for 5% and 1% tests of discordancy for an upper outlier in a gamma sample, using the ratio  $x_{(n)}/\sum x_i$  as test statistic. This table is reproduced, with permission from McGraw-Hill Book Company, from Eisenhart, Hastay, and Wallis (1947), Tables 15.1 and 15.2, with appropriate change of notation

5% critical values  
Ga1(Ea1)

$n \backslash r$	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	8	18	72	$\infty$
2	0.9985	0.9750	0.9392	0.9057	0.8772	0.8534	0.8332	0.8159	0.8010	0.7880	0.7341	0.6602	0.5813	0.5000	
3	0.9669	0.8709	0.7977	0.7457	0.7071	0.6771	0.6530	0.6333	0.6167	0.6025	0.5466	0.4748	0.4031	0.3333	
4	0.9065	0.7679	0.6841	0.6287	0.5895	0.5598	0.5365	0.5175	0.5017	0.4884	0.4366	0.3720	0.3093	0.2500	
5	0.8412	0.6838	0.5981	0.5441	0.5065	0.4783	0.4564	0.4387	0.4241	0.4118	0.3645	0.3066	0.2513	0.2000	
6	0.7808	0.6161	0.5321	0.4803	0.4447	0.4184	0.3980	0.3817	0.3682	0.3568	0.3135	0.2612	0.2119	0.1667	
7	0.7271	0.5612	0.4800	0.4307	0.3974	0.3726	0.3535	0.3384	0.3259	0.3154	0.2756	0.2278	0.1833	0.1429	
8	0.6798	0.5157	0.4377	0.3910	0.3595	0.3362	0.3185	0.3043	0.2926	0.2829	0.2462	0.2022	0.1616	0.1250	
9	0.6385	0.4775	0.4027	0.3584	0.3286	0.3067	0.2901	0.2768	0.2659	0.2568	0.2226	0.1820	0.1446	0.1111	
10	0.6020	0.4450	0.3733	0.3311	0.3029	0.2823	0.2666	0.2541	0.2439	0.2353	0.2032	0.1655	0.1308	0.1000	
12	0.5410	0.3924	0.3264	0.2880	0.2624	0.2439	0.2299	0.2187	0.2098	0.2020	0.1737	0.1403	0.1100	0.0833	
15	0.4709	0.3346	0.2758	0.2419	0.2195	0.2034	0.1911	0.1815	0.1736	0.1671	0.1429	0.1144	0.0889	0.0667	
20	0.3894	0.2705	0.2205	0.1921	0.1735	0.1602	0.1501	0.1422	0.1357	0.1303	0.1108	0.0879	0.0675	0.0500	
24	0.3434	0.2354	0.1907	0.1656	0.1493	0.1374	0.1286	0.1216	0.1160	0.1113	0.0942	0.0743	0.0567	0.0417	
30	0.2929	0.1980	0.1593	0.1377	0.1237	0.1137	0.1061	0.1002	0.0958	0.0921	0.0771	0.0604	0.0457	0.0333	
40	0.2370	0.1576	0.1259	0.1082	0.0968	0.0887	0.0827	0.0780	0.0745	0.0713	0.0595	0.0462	0.0347	0.0250	
60	0.1737	0.1131	0.0895	0.0765	0.0682	0.0623	0.0583	0.0552	0.0520	0.0497	0.0411	0.0316	0.0234	0.0167	
120	0.0998	0.0632	0.0495	0.0419	0.0371	0.0337	0.0312	0.0292	0.0279	0.0266	0.0218	0.0165	0.0120	0.0083	
$\infty$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

*1% critical values*

<i>n</i>	<i>r</i>	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	8	18	72	$\infty$
2	0.9999	0.9950	0.9794	0.9586	0.9373	0.9172	0.8988	0.8823	0.8674	0.8539	0.7949	0.7067	0.6062	0.5000	
3	0.9933	0.9423	0.8831	0.8335	0.7933	0.7606	0.7335	0.7107	0.6912	0.6743	0.6059	0.5153	0.4230	0.3333	
4	0.9676	0.8643	0.7814	0.7212	0.6761	0.6410	0.6129	0.5897	0.5702	0.5536	0.4884	0.4057	0.3251	0.2500	
5	0.9279	0.7885	0.6957	0.6329	0.5875	0.5531	0.5259	0.5037	0.4854	0.4697	0.4094	0.3351	0.2644	0.2000	
6	0.8828	0.7218	0.6258	0.5635	0.5195	0.4866	0.4608	0.4401	0.4229	0.4084	0.3529	0.2858	0.2229	0.1667	
7	0.8376	0.6644	0.5685	0.5080	0.4659	0.4347	0.4105	0.3911	0.3751	0.3616	0.3105	0.2494	0.1929	0.1429	
8	0.7945	0.6152	0.5209	0.4627	0.4226	0.3932	0.3704	0.3522	0.3373	0.3248	0.2779	0.2214	0.1700	0.1250	
9	0.7544	0.5727	0.4810	0.4251	0.3870	0.3592	0.3378	0.3207	0.3067	0.2950	0.2514	0.1992	0.1521	0.1111	
10	0.7175	0.5358	0.4469	0.3934	0.3572	0.3308	0.3106	0.2945	0.2813	0.2704	0.2297	0.1811	0.1376	0.1000	
12	0.6528	0.4751	0.3919	0.3428	0.3099	0.2861	0.2680	0.2535	0.2419	0.2320	0.1961	0.1535	0.1157	0.0833	
15	0.5747	0.4069	0.3317	0.2882	0.2593	0.2386	0.2228	0.2104	0.2002	0.1918	0.1612	0.1251	0.0934	0.0667	
20	0.4799	0.3297	0.2654	0.2288	0.2048	0.1877	0.1748	0.1646	0.1567	0.1501	0.1248	0.0960	0.0709	0.0500	
24	0.4247	0.2871	0.2295	0.1970	0.1759	0.1608	0.1495	0.1406	0.1338	0.1283	0.1060	0.0810	0.0595	0.0417	
30	0.3632	0.2412	0.1913	0.1635	0.1454	0.1327	0.1232	0.1157	0.1100	0.1054	0.0867	0.0658	0.0480	0.0333	
40	0.2940	0.1915	0.1508	0.1281	0.1135	0.1033	0.0957	0.0898	0.0853	0.0816	0.0668	0.0503	0.0363	0.0250	
60	0.2151	0.1371	0.1069	0.0902	0.0796	0.0722	0.0668	0.0625	0.0594	0.0567	0.0461	0.0344	0.0245	0.0167	
120	0.1225	0.0759	0.0585	0.0489	0.0429	0.0387	0.0357	0.0334	0.0316	0.0302	0.0242	0.0178	0.0125	0.0083	
$\infty$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

*n* = number of observations.

*r* = shape parameter of the gamma distribution (*r* = 1 for exponential distribution).

Table II Critical values for 5% and 1% tests of discordancy for a lower outlier in an exponential sample, using  $x_{(1)}/\sum x_i$  as test statistic. Values of the statistic *lower* than the critical value are significant

<b>Ea3</b>		
<i>n</i>	5%	1%
3	0.00844	0.00167
4	0.00424	$^a 0.0^3 836$
5	0.00255	$0.0^3 502$
6	0.00170	$0.0^3 335$
7	0.00122	$0.0^3 239$
8	$0.0^3 913$	$0.0^3 179$
9	$0.0^3 710$	$0.0^3 140$
10	$0.0^3 568$	$0.0^3 112$
12	$0.0^3 388$	$0.0^4 761$
14	$0.0^3 281$	$0.0^4 552$
16	$0.0^3 213$	$0.0^4 419$
18	$0.0^3 167$	$0.0^4 328$
20	$0.0^3 135$	$0.0^4 264$
30	$0.0^4 589$	$0.0^4 116$
40	$0.0^4 329$	$0.0^5 644$
50	$0.0^4 209$	$0.0^5 410$
100	$0.0^5 518$	$0.0^5 102$

*n* = number of observations.

Table III Critical values for 5% and 1% Dixon-type tests of discordancy for an upper outlier in an exponential sample, using  $\frac{x_{(n)} - x_{(n-1)}}{x_{(n)}}$  or  $\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$  as test statistic

Ea2, E2			
For testing		For testing	
$T_{\text{Ea2}} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)}}$	$n$	$T_{\text{E2}} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$	$n$
2		3	0.974
3		4	0.894
4		5	0.830
5		6	0.782
6		7	0.746
7		8	0.717
8		9	0.694
9		10	0.675
10		11	0.658
11		12	0.644
12		13	0.631
13		14	0.620
14		15	0.610
15		16	0.601
16		17	0.593
17		18	0.586
18		19	0.579
19		20	0.573
20		21	0.567

$n$  = number of observations.

Table IV Critical values for 5% and 1% tests of discordancy for a lower and upper outlier-pair in a gamma sample, using the ratio  $x_{(n)}/x_{(1)}$  as test statistic. This table is reproduced, with permission of the Biometrika Trustees, from Pearson and Hartley (1966), Table 31, after changing the notation where appropriate

*Upper 5% points*

		Ga7(Ea7)										
		Ga7(Ea7)										
		Ga7(Ea7)										
<i>n</i>	<i>r</i>	2	3	4	5	6	7	8	9	10	11	12
1	39.0	87.5	142	202	266	333	403	475	550	626	704	
1.5	15.4	27.8	39.2	50.7	62.0	72.9	83.5	93.9	104	114	124	
2	9.60	15.5	20.6	25.2	29.5	33.6	37.5	41.1	44.6	48.0	51.4	
2.5	7.15	10.8	13.7	16.3	18.7	20.8	22.9	24.7	26.5	28.2	29.9	
3	5.82	8.38	10.4	12.1	13.7	15.0	16.3	17.5	18.6	19.7	20.7	
3.5	4.99	6.94	8.44	9.70	10.8	11.8	12.7	13.5	14.3	15.1	15.8	
4	4.43	6.00	7.18	8.12	9.03	9.78	10.5	11.1	11.7	12.2	12.7	
4.5	4.03	5.34	6.31	7.11	7.80	8.41	8.95	9.45	9.91	10.3	10.7	
5	3.72	4.85	5.67	6.34	6.92	7.42	7.87	8.28	8.66	9.01	9.34	
6	3.28	4.16	4.79	5.30	5.72	6.09	6.42	6.72	7.00	7.25	7.48	
7.5	2.86	3.54	4.01	4.37	4.68	4.95	5.19	5.40	5.59	5.77	5.93	
10	2.46	2.95	3.29	3.54	3.76	3.94	4.10	4.24	4.37	4.49	4.59	
15	2.07	2.40	2.61	2.78	2.91	3.02	3.12	3.21	3.29	3.36	3.39	
30	1.67	1.85	1.96	2.04	2.11	2.17	2.22	2.26	2.30	2.33	2.36	
$\infty$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	

Upper 1% points

$r \backslash n$	2	3	4	5	6	7	8	9	10	11	12
1	199	448	729	1036	1362	1705	2063	2432	2813	3204	3605
1.5	47.5	85	120	151	184	21(6)	24(9)	28(1)	31(0)	33(7)	36(1)
2	23.2	37	49	59	69	79	89	97	106	113	120
2.5	14.9	22	28	33	38	42	46	50	54	57	60
3	11.1	15.5	19.1	22	25	27	30	32	34	36	37
3.5	8.89	12.1	14.5	16.5	18.4	20	22	23	24	26	27
4	7.50	9.9	11.7	13.2	14.5	15.8	16.9	17.9	18.9	19.8	21
4.5	6.54	8.5	9.9	11.1	12.1	13.1	13.9	14.7	15.3	16.0	16.6
5	5.85	7.4	8.6	9.6	10.4	11.1	11.8	12.4	12.9	13.4	13.9
6	4.91	6.1	6.9	7.6	8.2	8.7	9.1	9.5	9.9	10.2	10.6
7.5	4.07	4.9	5.5	6.0	6.4	6.7	7.1	7.3	7.5	7.8	8.0
10	3.32	3.8	4.3	4.6	4.9	5.1	5.3	5.5	5.6	5.8	5.9
15	2.63	3.0	3.3	3.4	3.6	3.7	3.8	3.9	4.0	4.1	4.2
30	1.96	2.2	2.3	2.4	2.4	2.5	2.5	2.6	2.7	2.7	2.7
$\infty$	1.00	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Values in the column  $n = 2$  and in the rows  $r = 1$  and  $\infty$  are exact. Elsewhere the third digit may be in error by a few units for the 5% points and several units for the 1% points. The third digit figures in parentheses for  $r = 1.5$  are the most uncertain.

$n$  = number of observations.

$r$  = shape parameter of the gamma distribution.

Table V Critical values for 5% and 1% Dixon-type tests of discordancy for a lower outlier in an exponential sample, using  $[x_{(2)} - x_{(1)}]/[x_{(n)} - x_{(1)}]$  as test statistic

E4

<i>n</i>	5%	1%
3	0.905	0.980
4	0.618	0.808
5	0.429	0.618
6	0.316	0.479
7	0.246	0.381
8	0.198	0.312
9	0.165	0.262
10	0.140	0.224
11	0.121	0.194
12	0.106	0.171
13	0.094	0.152
14	0.085	0.136
15	0.077	0.124
16	0.070	0.113
17	0.064	0.103
18	0.059	0.095
19	0.055	0.088
20	0.051	0.082

*n* = number of observations.

Table VI Critical values for 5% and 1% tests for the presence of an undefined number of discordant values in an exponential sample, using Shapiro and Wilks' 'W-Exponential' statistic

E12

<i>n</i>	Lower 1%	Lower 5%	Upper 5%	Upper 1%
3	0.254	0.270	0.993	0.9997
4	0.130	0.160	0.858	0.968
5	0.0905	0.119	0.668	0.860
6	0.0665	0.0956	0.509	0.678
7	0.0591	0.0810	0.416	0.571
8	0.0512	0.0710	0.350	0.485
9	0.0442	0.0633	0.300	0.401
10	0.0404	0.0568	0.253	0.339
12	0.0358	0.0494	0.202	0.272
14	0.0317	0.0428	0.165	0.213
16	0.0280	0.0374	0.136	0.177
18	0.0250	0.0332	0.116	0.148
20	0.0227	0.0302	0.100	0.129
30	0.0164	0.0213	0.0593	0.0719
40	0.0131	0.0164	0.0414	0.0499
50	0.0111	0.0137	0.0317	0.0360
60	0.0095	0.0117	0.0252	0.0291
70	0.0084	0.0103	0.0209	0.0241
80	0.0075	0.0091	0.0177	0.0205
90	0.0069	0.0082	0.0156	0.0176
100	0.0063	0.0074	0.0139	0.0153

*n* = number of observations.

Table VII Critical values for 5% and 1% tests of discordancy for a single outlier in a normal sample, using the deviation from the sample mean or population mean, studentized or standardized, as test statistic

<i>n</i>	Table VIIa N1		Table VIIb N2		Table VIIc $N\mu_1$		Table VIId $N\mu_2$	
	5%	1%	5%	1%	5%	1%	5%	1%
3	1.15	1.15	1.15	1.15	1.68	1.72	1.70	1.73
4	1.46	1.49	1.48	1.50	1.85	1.95	1.90	1.97
5	1.67	1.75	1.71	1.76	1.98	2.12	2.05	2.15
6	1.82	1.94	1.89	1.97	2.07	2.25	2.16	2.30
7	1.94	2.10	2.02	2.14	2.15	2.36	2.26	2.42
8	2.03	2.22	2.13	2.28	2.21	2.45	2.33	2.52
9	2.11	2.32	2.21	2.38	2.26	2.52	2.40	2.61
10	2.18	2.41	2.29	2.48	2.31	2.59	2.45	2.68
12	2.29	2.55	2.41	2.63	2.40	2.70	2.55	2.80
14	2.37	2.66			2.47	2.79		
15	2.41	2.71	2.55	2.81	2.50	2.83	2.66	2.94
16	2.44	2.75			2.53	2.86		
18	2.50	2.82			2.58	2.91		
20	2.56	2.88	2.71	3.00	2.62	2.95	2.79	3.10
30	2.74	3.10			2.80	3.17	2.96	3.30
40	2.87	3.24			2.92	3.30	3.08	3.43
50	2.96	3.34			2.98	3.39		
60	3.03	3.41					3.23	3.59
100	3.21	3.60			3.23	3.61		
120	3.27	3.66					3.46	3.83

Table VII (Continued)  
 Table VIIe N $\sigma$ 1      Table VIIf N $\sigma$ 2      Table VIIg N $\mu\sigma$ 1

<i>n</i>	5%	1%	5%	1%	5%	1%
3	1.74	2.22	1.93	2.39	2.12	2.71
4	1.94	2.43	2.15	2.61	2.23	2.81
5	2.08	2.57	2.29	2.76	2.32	2.88
6	2.18	2.68	2.40	2.87	2.39	2.93
7	2.27	2.76	2.48	2.95	2.44	2.98
8	2.33	2.83	2.55	3.02	2.49	3.02
9	2.39	2.88	2.60	3.07	2.53	3.06
10	2.44	2.93	2.65	3.12	2.57	3.09
12	2.52	3.01			2.63	3.14
14	2.59	3.07			2.68	3.19
15	2.62	3.10	2.82	3.29	2.71	3.21
16	2.64	3.12			2.73	3.23
18	2.69	3.17			2.77	3.26
20	2.73	3.21	2.94	3.39	2.80	3.29
30	2.88	3.38	3.08	3.53	2.93	3.40
40	2.99	3.44	3.18	3.62	3.02	3.48
50	3.05	3.53			3.08	3.54
60			3.30	3.73	3.14	3.59
100	3.27	3.67			3.28	3.72
120					3.33	3.76
200					3.47	3.89
500					3.71	4.11
1000					3.88	4.26

*n* = number of observations.

Table VIIa Critical values for 5% and 1% tests of discordancy for a single outlier in a normal sample, using the externally studentized deviation from the mean as test statistic

		Nv1									
		5% critical values									
$n$	$\nu$	5	6	8	10	15	20	30	40	60	$\infty$
3	2.37	2.24	2.09	2.01	1.91	1.87	1.82	1.80	1.78	1.74	
4	2.71	2.55	2.37	2.27	2.15	2.10	2.04	2.02	1.99	1.94	
5	2.95	2.78	2.57	2.46	2.32	2.26	2.20	2.17	2.14	2.08	
6	3.15	2.95	2.72	2.60	2.45	2.38	2.31	2.28	2.25	2.18	
7	3.30	3.09	2.85	2.72	2.55	2.47	2.40	2.37	2.33	2.27	
8	3.43	3.21	2.95	2.81	2.64	2.56	2.48	2.44	2.41	2.33	
9	3.54	3.31	3.04	2.89	2.71	2.63	2.54	2.50	2.47	2.39	
10	3.64	3.39	3.12	2.96	2.77	2.68	2.60	2.56	2.52	2.44	
12	3.80	3.54	3.25	3.08	2.88	2.78	2.69	2.65	2.61	2.52	

1% critical values

$n$	$\nu$	5	6	8	10	15	20	30	40	60	$\infty$
3	3.65	3.32	2.96	2.78	2.57	2.47	2.38	2.34	2.29	2.22	
4	4.11	3.72	3.31	3.10	2.84	2.73	2.62	2.57	2.52	2.43	
5	4.45	4.02	3.56	3.32	3.03	2.91	2.79	2.73	2.68	2.57	
6	4.70	4.24	3.74	3.48	3.17	3.04	2.91	2.85	2.79	2.68	
7	4.93	4.43	3.89	3.62	3.29	3.14	3.01	2.94	2.88	2.76	
8	5.11	4.58	4.02	3.73	3.38	3.23	3.08	3.02	2.95	2.83	
9	5.26	4.71	4.13	3.82	3.46	3.30	3.15	3.08	3.01	2.88	
10	5.39	4.82	4.22	3.90	3.53	3.37	3.21	3.13	3.06	2.93	
12	5.62	5.01	4.38	4.04	3.65	3.47	3.30	3.22	3.15	3.01	

$n$  = number of observations.

$\nu$  = degrees of freedom of independent estimate of  $\sigma^2$ .

Table VIIIB Critical values for 5% and 1% tests of discordancy for a single outlier in a normal sample, using the greatest externally studentized deviation from the mean as test statistic

		Nv3								
5% critical values		6	8	10	15	20	30	40	60	$\infty$
$n \setminus \nu$										
3		2.6	2.4	2.3	2.2	2.1	2.0	2.0	2.0	1.9
4			2.7	2.6	2.4	2.3	2.3	2.2	2.2	2.1
5				2.9	2.8	2.6	2.5	2.4	2.4	2.3
6					2.9	2.7	2.6	2.5	2.5	2.4
8						2.9	2.8	2.7	2.7	2.6
10							3.1	3.0	2.9	2.7
15								3.3	3.2	2.8
20									3.3	2.9
30									3.4	3.2
40										3.4
60										3.5

		1% critical values								
1% critical values		8	10	15	20	30	40	60	$\infty$	
$n \setminus \nu$										
3			3.3	3.1	2.8	2.7	2.6	2.5	2.5	2.4
4				3.7	3.4	3.1	3.0	2.9	2.8	2.7
5					4.0	3.7	3.3	3.2	2.9	2.8
6						3.9	3.5	3.3	3.2	2.9
8							4.1	3.7	3.5	3.2
10								4.3	3.8	3.6
15									4.1	3.9
20										4.0
30										
40										
60										

$n$  = number of observations.

$\nu$  = degrees of freedom of independent estimate of  $\sigma^2$ .

Table VIIIC Critical values for 5% and 1% tests of discordancy for a single outlier in a normal sample, using the externally and internally studentized deviation from the mean as test statistic

Nv2

5% critical values

$n \backslash \nu$	1	2	3	4	6	12	50
3	1.37	1.48	1.55	1.59	1.63	1.68	1.72
4	1.60	1.68	1.73	1.77	1.81	1.87	1.92
5	1.76	1.82	1.87	1.90	1.94	2.00	2.06
6	1.89	1.94	1.97	2.00	2.04	2.09	2.16
7	1.99	2.03	2.06	2.08	2.11	2.17	2.24
8	2.07	2.10	2.13	2.15	2.18	2.23	2.30
10	2.20	2.23	2.24	2.26	2.29	2.33	2.40
12	2.31	2.32	2.34	2.35	2.37	2.41	2.48
15	2.42	2.44	2.45	2.46	2.47	2.51	2.58
20	2.57	2.58	2.58	2.59	2.60	2.63	2.68

1% critical values

$n \backslash \nu$	1	2	3	4	6	12	50
3	1.40	1.58	1.70	1.79	1.90	2.04	2.17
4	1.69	1.82	1.92	1.99	2.09	2.22	2.30
5	1.90	2.00	2.08	2.14	2.23	2.36	2.51
6	2.06	2.14	2.21	2.26	2.33	2.46	2.61
7	2.19	2.25	2.31	2.35	2.42	2.53	2.69
8	2.29	2.35	2.40	2.43	2.49	2.60	2.75
10	2.46	2.50	2.54	2.57	2.61	2.70	2.85
12	2.59	2.62	2.65	2.67	2.70	2.79	2.92
15	2.73	2.75	2.77	2.79	2.82	2.88	3.01
20	2.90	2.91	2.93	2.94	2.96	3.01	3.12

$n$  = number of observations.

$\nu$  = degrees of freedom of independent estimate of  $\sigma^2$ .

Table VIIId Critical values for 5% and 1% tests of discordancy for a single outlier in a normal sample, using the greatest externally and internally studentized deviation from the mean as test statistic

		Nv4						
5% critical values		1	2	3	4	6	12	50
$n \backslash \nu$								
3		1.39	1.54	1.63	1.69	1.76	1.8	1.9
4		1.65	1.76	1.83	1.88	1.95	2.03	2.1
5		1.83	1.92	1.97	2.02	2.08	2.16	2.2
6		1.98	2.04	2.09	2.12	2.18	2.26	2.35
7		2.09	2.14	2.18	2.21	2.26	2.34	2.43
8		2.18	2.22	2.26	2.29	2.33	2.40	2.49
10		2.33	2.36	2.38	2.40	2.44	2.50	2.59
12		2.44	2.46	2.48	2.50	2.53	2.58	2.67
15		2.57	2.58	2.60	2.61	2.63	2.68	2.77
20		2.72	2.73	2.74	2.75	2.77	2.80	2.87

1% critical values								
$n \backslash \nu$		1	2	3	4	6	12	50
3		1.41	1.60	1.74	1.84	1.97	2.15	2.3
4		1.70	1.86	1.97	2.06	2.18	2.35	2.53
5		1.93	2.05	2.14	2.21	2.32	2.48	2.67
6		2.10	2.20	2.28	2.34	2.43	2.58	2.77
7		2.24	2.32	2.39	2.44	2.52	2.66	2.85
8		2.36	2.42	2.48	2.53	2.60	2.73	2.92
10		2.54	2.59	2.63	2.67	2.73	2.84	3.01
12		2.68	2.71	2.75	2.78	2.82	2.92	3.09
15		2.84	2.86	2.89	2.91	2.94	3.02	3.18
20		3.01	3.03	3.05	3.06	3.09	3.15	3.28

$n$  = number of observations.

$\nu$  = degrees of freedom of independent estimate of  $\sigma^2$ .

Table IXa,b,c,d,e,f,g,h Critical values for 5% and 1% tests of discordancy for  $k$  upper outliers in a normal sample (part of Table IXe is reproduced by permission of the American Statistical Association and the American Society for Quality Control)

Table IXa  
N3

Table IXb  
N4

$n$	$k = 2$			$k = 3$			$k = 4$			$k = 2$			$k = 3$			$k = 4$		
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
5	2.10	2.16					0.019	0.0034										
6	2.41	2.50					0.056	0.019										
7	2.66	2.79	2.97	3.08			0.102	0.044										
8	2.87	3.02	3.29	3.42			0.148	0.075										
9	3.04	3.22	3.58	3.73	3.82	3.98	0.191	0.108	0.099	0.048	0.045	0.018						
10	3.18	3.40	3.82	4.00	4.17	4.34	0.231	0.141	0.129	0.070	0.070	0.032						
12	3.44	3.70	4.24	4.44	4.72	4.92	0.300	0.204	0.196	0.120	0.125	0.070						
14	3.66	3.92	4.57	4.83	5.20	5.42	0.357	0.261	0.250	0.172	0.174	0.113						
16	3.83	4.10	4.85	5.14	5.60	5.85	0.405	0.310	0.300	0.219	0.219	0.151						
18	3.96	4.25	5.08	5.38	5.91	6.20	0.446	0.353	0.337	0.260	0.259	0.192						
20	4.11	4.41	5.30	5.60	6.22	6.54	0.480	0.391	0.377	0.300	0.299	0.231						
30	4.56	4.92	6.03	6.41	7.26	7.64	0.601	0.527	0.506	0.434	0.434	0.369						
40	4.84	5.29	6.49	6.98	7.93	8.38	0.672	0.610	0.588	0.522	0.523	0.460						
50	5.06	5.51	6.82	7.34	8.38	8.88	0.720	0.667	0.646	0.592	0.588	0.531						
100	5.62	6.06	7.77	8.27	9.71	10.3	0.833	0.802										

Table IXc  
N $\mu$ 3

Table IXd  
N $\mu$ 4

n	k = 2			k = 3			k = 4			k = 2	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	
4	2.68	2.77							0.034	0.0072	
5	2.90	3.05							0.086	0.027	
6	3.10	3.27	3.85	4.04					0.128	0.052	
7	3.27	3.47	4.07	4.29					0.169	0.091	
8	3.41	3.63	4.30	4.52	4.98	5.25			0.215	0.125	
9	3.53	3.77	4.50	4.72	5.23	5.52			0.255	0.157	
10	3.63	3.89	4.66	4.90	5.44	5.76			0.286	0.186	
12	3.83	4.12	4.94	5.24	5.84	6.17			0.349	0.244	
14	3.98	4.30	5.18	5.53	6.18	6.55			0.400	0.298	
16	4.10	4.45	5.39	5.75	6.47	6.84			0.448	0.346	
18	4.22	4.59	5.57	5.95	6.72	7.10			0.484	0.386	
20	4.34	4.70	5.75	6.13	6.94	7.35			0.510	0.416	
30	4.71	5.13	6.32	6.78	7.73	8.17			0.616	0.543	
40	4.96	5.44	6.74	7.26	8.32	8.83			0.683	0.618	
50	5.12	5.65	6.99	7.52	8.65	9.19			0.729	0.670	
100	5.68	6.13	7.87	8.38	9.88	10.5			0.836	0.808	

Table IXe  
N $\sigma$ 3

Table IXf  
N $\sigma$ 4

n	k = 2			k = 3			k = 4			k = 2	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	
4	2.39	2.93							0.0012	0.00004	
5	2.81	3.38							0.037	0.0077	
6	3.10	3.69	3.37	3.97					0.162	0.049	
7	3.33	3.93	3.81	4.45					0.351	0.137	
8	3.51	4.12	4.13	4.80	4.35	5.02			0.585	0.268	
9	3.66	4.28	4.45	5.14	4.83	5.61			0.869	0.438	
10	3.79	4.41	4.66	5.40	5.16	5.93			1.22	0.631	
12	4.00	4.63	5.03	5.81	5.74	6.56			2.03	1.21	
14	4.17	4.79	5.31	6.09	6.18	7.01			2.94	1.95	
16	4.31	4.93	5.60	6.34	6.56	7.38			4.03	2.77	
18	4.43	5.05	5.80	6.53	6.86	7.67			5.11	3.69	
20	4.53	5.14	5.96	6.66	7.10	7.90			6.21	4.42	
30	4.89	5.52	6.57	7.31	7.97	8.83			12.4	9.8	
40	5.13	5.76	6.95	7.67	8.55	9.36			19.5	16.2	
50	5.28	5.94	7.18	7.94	8.89	9.74			26.8	22.9	
100	5.77	6.32	8.02	8.73	10.1	10.9			67.7	60.0	

Table IXg  
 $N_{\mu\sigma 3}$

Table IXh  
 $N_{\mu\sigma 4}$

n	k = 2			k = 3			k = 4			k = 2	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	
4	3.27	4.14							0.071	0.015	
5	3.44	4.26							0.232	0.067	
6	3.65	4.46	4.48	5.47					0.450	0.166	
7	3.82	4.58	4.72	5.69					0.683	0.335	
8	3.95	4.70	4.91	5.91	5.57	6.72			1.05	0.539	
9	4.04	4.78	5.10	6.08	5.88	6.99			1.41	0.741	
10	4.12	4.87	5.25	6.23	6.12	7.23			1.82	0.984	
12	4.28	5.00	5.54	6.51	6.54	7.63			2.66	1.57	
14	4.42	5.14	5.81	6.70	6.89	7.94			3.63	2.40	
16	4.54	5.25	6.02	6.87	7.18	8.21			4.73	3.24	
18	4.64	5.35	6.18	7.03	7.40	8.43			5.89	4.24	
20	4.73	5.44	6.30	7.16	7.60	8.63			7.12	5.28	
30	5.01	5.68	6.77	7.61	8.31	9.26			13.3	10.7	
40	5.24	5.85	7.15	7.90	8.81	9.73			20.4	16.9	
50	5.36	6.03	7.34	8.17	9.12	10.1			27.9	23.8	
100	5.81	6.39	8.11	8.83	10.2	11.1			68.6	61.4	

n = number of observations.

k = number of outliers.

Table X Critical values for 5% and 1% tests of discordancy for a lower and upper outlier-pair in a normal sample, using as statistic the ratio of the reduced sum of squares to either the total sum of squares or the population variance

n	Table Xa N5		Table Xb $N_{\mu 5}$		Table Xc $N_{\sigma 5}$		Table Xd $N_{\mu\sigma 5}$	
	5%	1%	5%	1%	5%	1%	5%	1%
4	0.00044	0.00001	0.019	0.0030	0.00009	0.00002	0.038	0.0062
5	0.011	0.002	0.049	0.017	0.025	0.0049	0.131	0.042
6	0.044	0.014	0.092	0.040	0.123	0.036	0.308	0.126
7	0.078	0.033	0.127	0.062	0.265	0.100	0.511	0.224
8	0.120	0.060	0.169	0.093	0.492	0.239	0.801	0.400
9	0.159	0.093	0.205	0.127	0.727	0.370	1.09	0.583
10	0.195	0.122	0.239	0.158	1.01	0.562	1.46	0.839
12	0.266	0.181	0.299	0.214	1.80	1.09	2.24	1.43
14	0.320	0.236	0.352	0.265	2.66	1.75	3.15	2.11
16	0.369	0.288	0.393	0.308	3.62	2.54	4.20	2.88
18	0.411	0.325	0.432	0.347	4.69	3.35	5.22	3.74
20	0.448	0.363	0.468	0.381	5.78	4.16	6.43	4.71
30	0.571	0.509	0.581	0.517	11.7	9.45	12.5	10.2
40	0.644	0.584	0.651	0.591	18.8	15.6	19.4	16.3
50	0.699	0.648	0.703	0.654	26.0	22.2	26.8	22.8
100	0.821	0.794	0.823	0.796	66.6	59.2	67.4	60.0

n = number of observations.

Table XIa,b Critical values for 5% and 1% tests of discordancy for a lower and upper outlier-pair in a normal sample, using the studentized range (XIa) or the standardized range (XIb) as test statistic

Table XIa  
N<sub>6</sub>

Table XIb  
N<sub>σ6</sub> (N<sub>μσ6</sub>)

<i>n</i>	5%	1%	5%	1%
3	2.00	2.00	3.31	4.12
4	2.43	2.45	3.63	4.40
5	2.75	2.80	3.86	4.60
6	3.01	3.10	4.03	4.76
7	3.22	3.34	4.17	4.88
8	3.40	3.54	4.29	4.99
9	3.55	3.72	4.39	5.08
10	3.69	3.87	4.47	5.16
12	3.91	4.13	4.62	5.29
14	4.09	4.34	4.74	5.40
16	4.24	4.52	4.85	5.49
18	4.37	4.67	4.93	5.57
20	4.49	4.80	5.01	5.65
30	4.89	5.26	5.30	5.91
40	5.16	5.56	5.50	6.09
50	5.35	5.77	5.65	6.23
60	5.51	5.94	5.76	6.34
100	5.90	6.36	6.08	6.64
200	6.39	6.84		
500	6.94	7.42		
1000	7.33	7.80		

*n* = number of observations.

Table XIc Critical values for 5% and 1% tests of discordancy for a lower and upper outlier-pair in a normal sample, using the externally studentized range as test statistic

Nv6

*5% critical values*

<i>n</i>	<i>v</i>	1	2	3	4	5	6	8	10	12	15	20	30	60
3	27.0	8.33	5.91	5.04	4.60	4.34	4.04	3.88	3.77	3.67	3.58	3.49	3.40	
4	32.8	9.80	6.82	5.76	5.22	4.90	4.53	4.33	4.20	4.08	3.96	3.85	3.74	
5	37.1	10.9	7.50	6.29	5.67	5.30	4.89	4.65	4.51	4.37	4.23	4.10	3.98	
6	40.4	11.7	8.04	6.71	6.03	5.63	5.17	4.91	4.75	4.59	4.45	4.30	4.16	
7	43.1	12.4	8.48	7.05	6.33	5.90	5.40	5.12	4.95	4.78	4.62	4.46	4.31	
8	45.4	13.0	8.85	7.35	6.58	6.12	5.60	5.30	5.12	4.94	4.77	4.60	4.44	
9	47.4	13.5	9.18	7.60	6.80	6.32	5.77	5.46	5.27	5.08	4.90	4.72	4.55	
10	49.1	14.0	9.46	7.83	6.99	6.49	5.92	5.60	5.39	5.20	5.01	4.82	4.65	
12	52.0	14.8	9.95	8.21	7.32	6.79	6.18	5.83	5.61	5.40	5.20	5.00	4.81	
14	54.3	15.4	10.3	8.52	7.60	7.03	6.39	6.03	5.80	5.57	5.36	5.15	4.94	
16	56.3	15.9	10.7	8.79	7.83	7.24	6.57	6.19	5.95	5.72	5.49	5.27	5.06	
18	58.0	16.4	11.0	9.03	8.03	7.43	6.73	6.34	6.09	5.85	5.61	5.38	5.15	
20	59.6	16.8	11.2	9.23	8.21	7.59	6.87	6.47	6.21	5.96	5.71	5.47	5.24	

*I % critical values*

<i>n</i>	<i>v</i>	1	2	3	4	5	6	8	10	12	15	20	30	60
3	135.0	19.0	10.6	8.12	6.98	6.33	5.64	5.27	5.05	4.84	4.64	4.45	4.28	
4	164.3	22.3	12.2	9.17	7.80	7.03	6.20	5.77	5.50	5.25	5.02	4.80	4.59	
5	185.6	24.7	13.3	9.96	8.42	7.56	6.62	6.14	5.84	5.56	5.29	5.05	4.82	
6	202.2	26.6	14.2	10.6	8.91	7.97	6.96	6.43	6.10	5.80	5.51	5.24	4.99	
7	215.8	28.2	15.0	11.1	9.32	8.32	7.24	6.67	6.32	5.99	5.69	5.40	5.13	
8	227.2	29.5	15.6	11.5	9.67	8.61	7.47	6.87	6.51	6.16	5.84	5.54	5.25	
9	237.0	30.7	16.2	11.9	9.97	8.87	7.68	7.05	6.67	6.31	5.97	5.65	5.36	
10	245.6	31.7	16.7	12.3	10.2	9.10	7.86	7.21	6.81	6.44	6.09	5.76	5.45	
12	260.0	33.4	17.5	12.8	10.7	9.48	8.18	7.49	7.06	6.66	6.28	5.93	5.60	
14	271.8	34.8	18.2	13.3	11.1	9.81	8.44	7.71	7.26	6.84	6.45	6.08	5.73	
16	281.8	36.0	18.8	13.7	11.4	10.1	8.66	7.91	7.44	7.00	6.59	6.20	5.84	
18	290.4	37.0	19.3	14.1	11.7	10.3	8.85	8.08	7.59	7.14	6.71	6.31	5.93	
20	298.0	37.9	19.8	14.4	11.9	10.5	9.03	8.23	7.73	7.26	6.82	6.41	6.01	

*n* = number of observations.

*v* = degrees of freedom of independent estimate of  $\sigma^2$ . The critical values for  $v = \infty$  are given by Table XIb.

Table XII Critical values for 5% and 1% tests of discordancy for  $k$  lower outliers and  $k$  upper outliers in a normal sample, using the standardized  $(k-1)$ th quasi-range as test statistic

$N\sigma^9$  ( $N\mu\sigma^9$ )

$n$	$k = 2$		$k = 3$		$k = 4$	
	5%	1%	5%	1%	5%	1%
4	1.58	2.17				
5	2.05	2.61				
6	2.35	2.89	1.11	1.56		
7	2.58	3.10	1.51	1.95		
8	2.75	3.26	1.78	2.21	0.859	1.23
9	2.90	3.40	1.99	2.41	1.20	1.56
10	3.03	3.52	2.15	2.57	1.44	1.80
12	3.23	3.71	2.42	2.82	1.79	2.14
14	3.39	3.86	2.62	3.01	2.04	2.38
16	3.53	3.98	2.79	3.17	2.24	2.57
18	3.64	4.09	2.93	3.30	2.40	2.73
20	3.74	4.18	3.05	3.41	2.54	2.86
30	4.11	4.52	3.47	3.81	3.02	3.32
40	4.35	4.75	3.74	4.07	3.32	3.61
50	4.53	4.92	3.94	4.26	3.54	3.82
60	4.67	5.05	4.10	4.41	3.71	3.98
70	4.79	5.16	4.23	4.54	3.85	4.12
80	4.88	5.26	4.34	4.64	3.97	4.23
90	4.97	5.34	4.43	4.73	4.07	4.33
100	5.05	5.41	4.52	4.81	4.16	4.41

$n$  = number of observations.

$2k$  = number of outliers ( $k$  upper and  $k$  lower).

Table XIIIa,b,c,d,e,f,g Critical values for 5% and 1% Dixon-type tests of discordancy for one or more outliers in a normal sample

Table XIIIa N7 ( $N\mu 7$ )		Table XIIIb N8 ( $N\mu 8$ )		Table XIIIc N9 ( $N\mu 9$ )		Table XIIId N10 ( $N\mu 10$ )		
<i>n</i>	5%	1%	5%	1%	5%	1%	5%	1%
3	0.941	0.988						
4	0.765	0.889	0.831	0.922	0.955	0.991		
5	0.642	0.780	0.717	0.831	0.807	0.916	0.960	0.992
6	0.560	0.698	0.621	0.737	0.689	0.805	0.824	0.925
7	0.507	0.637	0.570	0.694	0.610	0.740	0.712	0.836
8	0.468	0.590	0.524	0.638	0.554	0.683	0.632	0.760
9	0.437	0.555	0.492	0.594	0.512	0.635	0.580	0.701
10	0.412	0.527	0.464	0.564	0.477	0.597	0.537	0.655
12	0.376	0.482	0.429	0.520	0.428	0.541	0.473	0.590
14	0.349	0.450	0.397	0.485	0.395	0.502	0.432	0.542
16	0.329	0.426	0.376	0.461	0.369	0.472	0.401	0.508
18	0.313	0.407	0.354	0.438	0.349	0.449	0.377	0.480
20	0.300	0.391	0.340	0.417	0.334	0.430	0.358	0.458
25	0.277	0.362	0.316	0.386	0.304	0.394	0.324	0.417
30	0.260	0.341	0.300	0.368	0.283	0.369	0.301	0.389

Table XIIIe N11 ( $N\mu 11$ )		Table XIIIf N12 ( $N\mu 12$ )		Table XIIIg N13 ( $N\mu 13$ )		
<i>n</i>	5%	1%	5%	1%	5%	1%
4	0.967	0.992				
5	0.845	0.929	0.976	0.995		
6	0.736	0.836	0.872	0.951	0.983	0.995
7	0.661	0.778	0.780	0.885	0.881	0.945
8	0.607	0.710	0.710	0.829	0.803	0.890
9	0.565	0.667	0.657	0.776	0.737	0.840
10	0.531	0.632	0.612	0.726	0.682	0.791
12	0.481	0.579	0.546	0.642	0.600	0.704
14	0.445	0.538	0.501	0.593	0.546	0.641
16	0.418	0.508	0.467	0.557	0.507	0.595
18	0.397	0.484	0.440	0.529	0.475	0.561
20	0.372	0.464	0.419	0.506	0.450	0.535
25	0.343	0.428	0.382	0.464	0.406	0.489
30	0.322	0.402	0.355	0.433	0.376	0.457

*n* = number of observations.

Table XIIIh,i Critical values for 5% and 1% tests of discordancy for one, or two, upper outliers in a normal sample, using  $[x_{(n)} - x_{(n-1)}]/\sigma$  (XIIIh) or  $[x_{(n-1)} - x_{(n-2)}]/\sigma$  (XIIIi) as test statistic

n	Table XIIIh N $\sigma$ 7 (N $\mu\sigma$ 7)		Table XIIIi N $\sigma$ 8 (N $\mu\sigma$ 8)	
	5%	1%	5%	1%
3	2.17	2.90	2.17	2.90
10	1.46	2.03	0.96	1.38
20	1.28	1.80	0.79	1.14
30	1.20	1.70	0.73	1.05
40	1.14	1.63	0.68	1.00
60	1.08	1.56	0.63	0.93
80	1.04	1.50	0.61	0.90
100	1.02	1.47	0.58	0.86
200	0.95	1.38	0.54	0.81
500	0.87	1.28	0.48	0.73
1000	0.83	1.22	0.45	0.67

n = number of observations.

Table XIVa,b,c Critical values for 5% and 1% tests of discordancy for one or more outliers in a normal sample, using as test statistic the sample skewness (XIVa), the sample kurtosis (XIVb), or the sample kurtosis based on deviations from the population mean (XIVc)

n	Table XIVa N14		Table XIVb N15	
	5%	1%	5%	1%
5	1.0	1.3	2.9	3.1
10	0.9	1.3	3.9	4.8
15	0.8	1.2	4.1	5.1
20	0.8	1.1	4.1	5.2
25	0.71	1.06	4.0	5.0
30	0.66	0.99		
40	0.59	0.87		
50	0.53	0.79	3.99	4.88
60	0.49	0.72		
70	0.46	0.67		
75			3.87	4.59
80	0.43	0.63		
90	0.41	0.60		
100	0.39	0.57	3.77	4.39
200	0.28	0.40	3.57	3.98
500	0.18	0.26	3.37	3.60
1000	0.13	0.18	3.26	3.41

Table XIVc  
N $\mu$ 6

<i>n</i>	5%	1%
3	2.8	3.0
4	3.3	3.7
5	3.6	4.3
6	3.8	4.7
7	3.9	4.9
8	4.0	5.1
9	4.0	5.3
10	4.1	5.4
12	4.2	5.5
14	4.2	5.5
16	4.2	5.5
18	4.2	5.5
20	4.2	5.4
30	4.2	5.2
40	4.1	5.0
50	4.0	4.9
100	3.8	4.3

*n* = number of observations.

Table XV Critical values for 5% and 1% tests of discordancy for  $k$  outliers in a normal sample, using Tietjen and Moore's  $E_k$ -statistic

N16

<i>n</i>	$k = 2$		$k = 3$		$k = 4$	
	5%	1%	5%	1%	5%	1%
5	0.010	0.002				
6	0.034	0.012				
7	0.065	0.028	0.016	0.006		
8	0.099	0.050	0.034	0.014		
9	0.137	0.078	0.057	0.026	0.021	0.009
10	0.172	0.101	0.083	0.043	0.037	0.018
12	0.234	0.159	0.133	0.083	0.073	0.042
14	0.293	0.207	0.179	0.123	0.112	0.072
16	0.340	0.263	0.227	0.166	0.153	0.107
18	0.382	0.306	0.267	0.206	0.187	0.141
20	0.416	0.339	0.302	0.236	0.221	0.170
30	0.549	0.482	0.443	0.386	0.364	0.308
40	0.629	0.574	0.534	0.480	0.458	0.408
50	0.684	0.636	0.599	0.550	0.529	0.482

*n* = number of observations.

*k* = number of outliers.

Table XVIa Critical values for 5% and 1% tests for the presence of an undefined number of discordant values in a normal sample, using Shapiro and Wilks' W-statistic

N17

<i>n</i>	5%	1%
3	0.767	0.753
4	0.748	0.687
5	0.762	0.686
6	0.788	0.713
7	0.803	0.730
8	0.818	0.749
9	0.829	0.764
10	0.842	0.781
12	0.859	0.805
14	0.874	0.825
16	0.887	0.844
18	0.897	0.858
20	0.905	0.868
25	0.918	0.888
30	0.927	0.900
35	0.934	0.910
40	0.940	0.919
45	0.945	0.926
50	0.947	0.930

*n* = number of observations.

Table XVIb Values of the constants  $a_{n,n-i+1}$  required for calculating Shapiro and Wilks'  $W$ -statistic  $T_{N17} = L^2/S^2$ , where

$$L = \sum_{i=1}^{[n/2]} a_{n,n-i+1} [x_{(n-i+1)} - x_{(i)}] \quad \text{and} \quad S^2 = \sum_{i=1}^n (x_i - \bar{x})^2$$

$n \backslash i$	1	2	3	4	5	6	7	8	9	10
3	0.7071									
4	0.6872	0.1677								
5	0.6646	0.2413								
6	0.6431	0.2806	0.0875							
7	0.6233	0.3031	0.1401							
8	0.6052	0.3164	0.1743	0.0561						
9	0.5888	0.3244	0.1976	0.0947						
10	0.5739	0.3291	0.2141	0.1224	0.0399					
12	0.5475	0.3325	0.2347	0.1586	0.0922	0.0303				
14	0.5251	0.3318	0.2460	0.1802	0.1240	0.0727	0.0240			
16	0.5056	0.3290	0.2521	0.1939	0.1447	0.1005	0.0593	0.0196		
18	0.4886	0.3253	0.2553	0.2027	0.1587	0.1197	0.0837	0.0496	0.0163	
20	0.4734	0.3211	0.2565	0.2085	0.1686	0.1334	0.1013	0.0711	0.0422	0.0140
25	0.4450	0.3069	0.2543	0.2148	0.1822	0.1539	0.1283	0.1046	0.0823	0.0610
30	0.4254	0.2944	0.2487	0.2148	0.1870	0.1630	0.1415	0.1219	0.1036	0.0862
35	0.4096	0.2834	0.2427	0.2127	0.1883	0.1673	0.1487	0.1317	0.1160	0.1013
40	0.3964	0.2737	0.2368	0.2098	0.1878	0.1691	0.1526	0.1376	0.1237	0.1108
45	0.3850	0.2651	0.2313	0.2065	0.1865	0.1695	0.1545	0.1410	0.1286	0.1170
50	0.3751	0.2574	0.2260	0.2032	0.1847	0.1691	0.1554	0.1430	0.1317	0.1212

$n \backslash i$	11	12	13	14	15	16	17	18	19	20
25	0.0403	0.0200								
30	0.0697	0.0537	0.0381	0.0227	0.0076					
35	0.0873	0.0739	0.0610	0.0484	0.0361	0.0239	0.0119			
40	0.0986	0.0870	0.0759	0.0651	0.0546	0.0444	0.0343	0.0244	0.0146	0.0049
45	0.1062	0.0959	0.0860	0.0765	0.0673	0.0584	0.0497	0.0412	0.0328	0.0245
50	0.1113	0.1020	0.0932	0.0846	0.0764	0.0685	0.0608	0.0532	0.0459	0.0386

$n \backslash i$	21	22	23	24	25
45	0.0163	0.0081			
50	0.0314	0.0244	0.0174	0.0104	0.0035

$n$  = number of observations.

Table XVIIa Critical values for 5% and 1% tests of discordancy for an upper outlier  $x_{(n)}$  in a Poisson sample, using  $x_{(n)}$  conditional on  $\sum x_j$  as test statistic

P1

5% critical values

$n$	$\sum x_j - x_{(n)}$	0	1	2	3	4	5	6	8	10	12	14	16	18	20	22	24
3	4	7	9	11	13	15	16	20	24	27	31	34	38	41	44	48	
4	4	6	8	9	11	13	14	17	21	24	27	30	33	36	39	42	
5	3	5	7	9	10	12	13	16	19	22	25	28	30	33	36	39	
6	3	5	7	8	10	11	12	15	18	21	24	26	29	32	34	37	
8	3	5	6	7	9	10	11	14	17	19	22	25	27	30	32	35	
10	3	4	6	7	8	10	11	14	16	19	21	24	26	28	31	33	
12	3	4	6	7	8	10	11	13	16	18	21	23	25	28	30	32	
16	3	4	5	7	8	9	10	13	15	17	20	22	24	27	29	31	
20	2	4	5	6	8	9	10	12	15	17	19	22	24	26	28	31	
25	2	4	5	6	7	9	10	12	14	17	19	21	23	26	28	30	
50	2	4	5	6	7	8	9	12	14	16	18	20	22	25	27	29	
100	2	3	5	6	7	8	9	11	13	15	18	20	22	24	26	28	

The outlier  $x_{(n)}$  is significant at 5% if  $\sum x_j$  is greater than or equal to the tabulated value in the column corresponding to the observed value of  $\sum x_j - x_{(n)}$ .

1% critical values

$n$	$\sum x_j - x_{(n)}$	0	1	2	3	4	5	6	8	10	12	14	16	18	20	22	24
3	6	8	11	13	15	17	19	23	26	30	34	37	41	44	48	51	
4	5	7	9	11	13	14	16	19	23	26	29	32	35	38	41	45	
5	4	6	8	10	11	13	15	18	21	24	27	30	32	35	38	41	
6	4	6	8	9	11	12	14	17	20	22	25	28	31	33	36	39	
8	4	5	7	8	10	11	13	15	18	21	23	26	29	31	34	36	
10	3	5	7	8	9	11	12	15	17	20	22	25	27	30	32	35	
12	3	5	6	8	9	10	12	14	17	19	22	24	27	29	31	34	
16	3	5	6	7	9	10	11	14	16	18	21	23	26	28	30	32	
20	3	4	6	7	8	10	11	13	16	18	20	23	25	27	29	32	
25	3	4	6	7	8	9	11	13	15	18	20	22	24	27	29	31	
50	3	4	5	6	8	9	10	12	14	17	19	21	23	25	27	30	
100	2	4	5	6	7	8	9	12	14	16	18	20	22	24	27	29	

The outlier  $x_{(n)}$  is significant at 1% if  $\sum x_j$  is greater than or equal to the tabulated value in the column corresponding to the observed value of  $\sum x_j - x_{(n)}$ .

$n$  = number of observations.

$x_{(n)}$  = greatest observation.

$\sum x_j$  = sum of observations.

Table XVIIb Critical values for 5% and 1% tests of discordancy for a lower outlier  $x_{(1)}$  in a Poisson sample, using  $x_{(1)}$  conditional on  $\sum x_i$  as test statistic

P2

*5% critical values*

$n \backslash x_{(1)}$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3	11	16	21	25	29	33	37	41	45	49	53	57	61	64	68	72
4	16	23	30	36	42	47	53	58	64	69	74	80	85	90	95	
5	21	31	39	47	55	62	69	76	83	89	96					
6	27	38	49	58	68	76	85	93								

The outlier  $x_{(1)}$  is significant at 5% if  $\sum x_i$  is greater than or equal to the tabulated value in the column corresponding to the observed value of  $x_{(1)}$ .

*1% critical values*

$n \backslash x_{(1)}$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3	15	20	26	30	35	39	44	48	52	56	60	64	68	72	76	80
4	21	30	37	43	50	56	62	67	73	79	84	90	95	100		
5	28	39	48	57	65	72	80	87	95							
6	35	48	60	70	80	89	98									

The outlier  $x_{(1)}$  is significant at 1% if  $\sum x_i$  is greater than or equal to the tabulated value in the column corresponding to the observed value of  $x_{(1)}$ .

$n$  = number of observations.

$x_{(1)}$  = smallest observation.

$\sum x_i$  = sum of observations.

Table XVIIIA Critical values for 5% and 1% tests of discordancy for an upper outlier-pair  $x_{(n-1)}, x_{(n)}$  in a Poisson sample, using  $x_{(n-1)} + x_{(n)}$  conditional on  $\sum x_j$  as test statistic

P3

*5% critical values*

$\sum x_j - x_{(n-1)} - x_{(n)}$		0	1	2	3	4	5	6	8	10	12	14
n		-										
4	7	11	14	17	20	23	25	30	35	40	45	
5	6	9	12	14	16	19	21	25	29	34	38	
6	6	8	11	13	15	17	19	23	26	30	34	
8	5	7	9	11	13	15	16	20	23	26	29	
10	5	7	9	10	12	14	15	18	21	24	27	
12	4	6	8	10	11	13	14	17	20	23	26	
16	4	6	8	9	11	12	13	16	19	22	24	
20	4	6	7	9	10	12	13	16	18	21	23	

The outlier-pair  $x_{(n-1)}, x_{(n)}$  is significant at 5% if  $\sum x_j$  is greater than or equal to the tabulated value in the column corresponding to the observed value of  $\sum x_j - x_{(n-1)} - x_{(n)}$ .

*1% critical values*

$\sum x_j - x_{(n-1)} - x_{(n)}$		0	1	2	3	4	5	6	8	10	12	14
n		-										
4	10	14	17	20	23	26	29	34	39	45	50	
5	8	11	14	16	19	21	24	28	32	37	41	
6	7	10	12	15	17	19	21	25	29	33	36	
8	6	8	11	13	14	16	18	21	25	28	31	
10	6	8	10	12	13	15	17	20	23	26	29	
12	5	7	9	11	12	14	16	19	22	25	27	
16	5	7	8	10	12	13	14	17	20	23	26	
20	5	7	8	10	11	12	14	17	19	21	24	

The outlier-pair  $x_{(n-1)}, x_{(n)}$  is significant at 1% if  $\sum x_j$  is greater than or equal to the tabulated value in the column corresponding to the observed value of  $\sum x_j - x_{(n-1)} - x_{(n)}$ .

$n$  = number of observations.

$x_{(n)}$  = greatest observation.

$x_{(n-1)}$  = second greatest observation.

$\sum x_j$  = sum of observations.

Table XVIIIb Critical values for 5% and 1% tests of discordancy for a lower outlier-pair  $x_{(1)}, x_{(2)}$  in a Poisson sample, using  $x_{(1)} + x_{(2)}$  conditional on  $\sum x_i$  as test statistic

*5% critical values*

P4

		P4								
		0	1	2	3	4	5	6	7	8
$n$	$x_{(1)} + x_{(2)}$									
	4	7	11	14	17	20	23	25	28	30
5	11	16	20	24	27	31	34	38	41	
6	15	20	26	30	35	39	44	48	52	
8	22	31	38	45	51	57	63	69	75	
10	31	42	51	60	68	76	84	91	99	

The outlier-pair  $x_{(1)}, x_{(2)}$  is significant at 5% if  $\sum x_i$  is greater than or equal to the tabulated value in the column corresponding to the observed value of  $x_{(1)} + x_{(2)}$ .

*1% critical values*

		P4								
		0	1	2	3	4	5	6	7	8
$n$	$x_{(1)} + x_{(2)}$									
	4	10	14	17	20	23	26	29	31	34
5	14	19	23	28	32	35	39	43	46	
6	19	25	30	35	40	45	50	54	58	
8	28	37	45	52	59	65	71	78	84	
10	38	50	60	69	78	86	94			

The outlier-pair  $x_{(1)}, x_{(2)}$  is significant at 1% if  $\sum x_i$  is greater than or equal to the tabulated value in the column corresponding to the observed value of  $x_{(1)} + x_{(2)}$ .

$n$  = number of observations.

$x_{(1)}$  = smallest observation.

$x_{(2)}$  = next smallest observation.

$\sum x_i$  = sum of observations.

Table XIX Critical values for 5% and 1% tests of discordancy for an upper outlier  $x_{(n)}$  in a binomial sample, using  $x_{(n)}$  conditional on  $\sum x_i$  as test statistic

B1,B2

5% critical values

$x_{(n)} = m$

$n \backslash m$	3	4	5	6	7	8	9	10
$n$	3	4	5	6	7	8	9	10
3	3	5	7	10	12	15	18	21
4	3	6	9	12	16	20	23	27
5	4	7	11	15	19	24	28	33
6	4	8	12	17	22	28	33	38
7	4	9	14	19	25	31	37	44
8	5	10	15	22	28	35	42	50
9	5	11	17	24	31	39	47	55
10	6	11	18	26	34	43	51	60

An outlier  $x_{(n)}$  equal to  $m$  is judged discordant at level 5% if  $\sum x_i$  is less than or equal to the tabulated value in the column corresponding to  $m$ .

$x_{(n)} = m - 1$

$n \backslash m$	3	4	5	6	7	8	9	10
$n$	—	—	4	7	9	12	14	17
3	—	—	4	7	9	12	14	17
4	—	3	5	8	11	15	18	22
5	—	3	6	10	14	18	22	26
6	—	4	7	11	16	20	25	31
7	—	4	8	12	18	23	29	35
8	—	4	9	14	20	26	32	39
9	—	5	9	15	21	28	36	43
10	—	5	10	16	23	31	39	47

An outlier  $x_{(n)}$  equal to  $m - 1$  is judged discordant at level 5% if  $\sum x_i$  is less than or equal to the tabulated value in the column corresponding to  $m$ .

Table XIX (Continued)

 $x_{(n)} = m - 2$ 

$n \backslash m$	3	4	5	6	7	8	9	10
n	—	—	—	4	6	9	11	14
3	—	—	—	4	6	9	11	14
4	—	—	3	5	8	11	14	17
5	—	—	3	6	9	13	17	21
6	—	—	3	7	11	15	19	24
7	—	—	4	7	12	17	22	27
8	—	—	4	8	13	19	25	31
9	—	—	4	9	14	20	27	34
10	—	—	5	9	15	22	29	37

An outlier  $x_{(n)}$  equal to  $m - 2$  is judged discordant at level 5% if  $\sum x_i$  is less than or equal to the tabulated value in the column corresponding to  $m$ .

### 1% critical values

 $x_{(n)} = m$ 

$n \backslash m$	3	4	5	6	7	8	9	10
n	—	4	6	8	10	13	16	19
3	—	4	6	8	10	13	16	19
4	—	4	7	10	13	17	20	24
5	—	5	8	12	16	20	24	29
6	3	6	9	14	18	23	27	33
7	3	6	11	15	21	26	31	38
8	3	7	12	17	23	29	36	43
9	3	7	13	19	25	33	40	47
10	4	8	14	20	28	35	44	52

An outlier  $x_{(n)}$  equal to  $m$  is judged discordant at level 1% if  $\sum x_i$  is less than or equal to the tabulated value in the column corresponding to  $m$ .

$x_{(n)} = m - 1$ 

Table XIX (Continued)

$n \backslash m$	3	4	5	6	7	8	9	10
$n$	—	—	—	5	8	10	12	15
3	—	—	—	5	8	10	12	15
4	—	—	4	6	9	12	15	19
5	—	—	5	7	11	14	18	22
6	—	—	5	8	12	17	21	26
7	—	3	6	9	14	19	24	29
8	—	3	6	10	15	21	27	33
9	—	3	7	11	17	23	29	36
10	—	3	7	12	18	25	32	40

An outlier  $x_{(n)}$  equal to  $m - 1$  is judged discordant at level 1% if  $\sum x_i$  is less than or equal to the tabulated value in the column corresponding to  $m$ .

 $x_{(n)} = m - 2$ 

$n \backslash m$	3	4	5	6	7	8	9	10
$n$	—	—	—	—	5	7	9	12
3	—	—	—	—	5	7	9	12
4	—	—	—	4	6	9	12	15
5	—	—	—	4	7	10	14	17
6	—	—	—	5	8	12	16	20
7	—	—	—	5	9	13	18	22
8	—	—	3	6	10	14	19	25
9	—	—	3	6	10	16	21	28
10	—	—	3	7	11	17	23	30

An outlier  $x_{(n)}$  equal to  $m - 2$  is judged discordant at level 1% if  $\sum x_i$  is less than or equal to the tabulated value in the column corresponding to  $m$ .

$n$  = number of observations.

$x_{(n)}$  = greatest observation.

$\sum x_i$  = sum of observations.

$m$  = parameter of binomial distribution.

Table XX Optimal choice of the number,  $m^*$ , of lower ordered sample values out of  $n$  used in estimating the scale  $\sigma$  of an exponential distribution, and associated relative efficiency  $e_{m^*}$ , when one observation has slipped in scale to  $\sigma/h$  (reproduced by permission of the author and the American Statistical Association)

n \ h	0.05		0.10		0.15		0.20		0.25	
	$m^*$	$e_{m^*}$								
2	1	88.59	1	21.96	1	9.66	1	5.38	1	3.43
3	1	74.71	2	17.58	2	7.88	2	4.49	2	2.93
4	2	67.10	2	15.97	2	6.83	3	3.95	3	2.63
5	3	60.42	3	14.51	3	6.29	4	3.57	4	2.41
6	4	54.84	4	13.27	4	5.81	4	3.29	5	2.25
7	5	50.19	5	12.22	5	5.40	5	3.10	6	2.12
8	6	46.28	6	11.33	6	5.05	6	2.93	7	2.01
9	7	42.96	7	10.57	7	4.75	7	2.79	8	1.93
10	8	40.10	8	9.91	8	4.49	8	2.66	9	1.85
15	12	30.73	13	7.65	13	3.59	13	2.22	13	1.62
20	17	25.06	17	6.36	18	3.06	18	1.96	18	1.49
30	27	18.45	27	4.87	27	2.46	28	1.68	28	1.34
40	36	14.77	37	4.03	37	2.14	38	1.52	38	1.25
50	46	12.39	47	3.49	47	1.93	48	1.42	48	1.20

n \ h	0.30		0.35		0.40		0.45		0.50	
	$m^*$	$e_{m^*}$								
2	1	2.39	1	1.79	1	1.41	1	1.16	1	1.00
3	2	2.11	2	1.62	2	1.32	2	1.13	2	1.00
4	3	1.93	3	1.52	3	1.27	3	1.11	3	1.00
5	4	1.80	4	1.45	4	1.23	4	1.09	4	1.00
6	5	1.70	5	1.39	5	1.20	5	1.08	6	1.00
7	6	1.63	6	1.35	6	1.17	6	1.07	7	1.00
8	7	1.56	7	1.31	7	1.16	7	1.06	8	1.00
9	8	1.51	8	1.28	8	1.14	8	1.05	9	1.00
10	9	1.47	9	1.26	9	1.13	9	1.05	10	1.00
15	14	1.33	14	1.17	14	1.08	14	1.03	14	1.00
20	19	1.25	19	1.13	19	1.06	19	1.02	20	1.00
30	29	1.17	29	1.08	29	1.04	29	1.01	30	1.00
40	39	1.12	39	1.06	39	1.03	39	1.01	40	1.00
50	49	1.10	49	1.05	49	1.02	49	1.00	50	1.00

Table XXI Critical values for the Mosteller non-parametric slip-page test ( $\frac{5\%}{1\%}$ )

$m \backslash n$	2	3	4	5	6
2	(—)	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
3	( $\frac{4}{2}$ )	( $\frac{4}{3}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
4	( $\frac{5}{2}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
5	( $\frac{5}{3}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
6	( $\frac{6}{2}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
7	( $\frac{5}{2}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
8	( $\frac{5}{2}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
9	( $\frac{5}{2}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
10	( $\frac{5}{2}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
15	( $\frac{5}{2}$ )	( $\frac{4}{2}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
20	( $\frac{5}{2}$ )	( $\frac{4}{2}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
25	( $\frac{5}{2}$ )	( $\frac{4}{2}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
100	( $\frac{6}{2}$ )	( $\frac{4}{2}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )
$\infty$	( $\frac{6}{2}$ )	( $\frac{4}{2}$ )	( $\frac{4}{2}$ )	( $\frac{3}{2}$ )	( $\frac{3}{2}$ )

$n$  = number of samples;  $m$  = sample size.

Table XXII Critical values of  $T_{\max}$  for the Doornbos and Prins rank test of slippage  
 $\begin{pmatrix} 5\% \\ 1\% \end{pmatrix}$

$n \backslash m$	2	3	4	5	6
3	( <u>—</u> )	( <u>24</u> )	( <u>32</u> )	( <u>40</u> )	( <u>49</u> ) 51
4	( <u>26</u> )	( <u>39</u> ) 42	( <u>53</u> ) 56	( <u>67</u> ) 71	( <u>81</u> ) 86
5	( <u>38</u> ) 40	( <u>58</u> ) 62	( <u>79</u> ) 84	( <u>99</u> ) 106	( <u>120</u> ) 128
6	( <u>52</u> ) 55	( <u>80</u> ) 86	( <u>109</u> ) 116	( <u>139</u> ) 150	( <u>168</u> ) 181
7	( <u>69</u> ) 73	( <u>106</u> ) 113	( <u>145</u> ) 155	( <u>183</u> ) 197	( <u>222</u> ) 238
8	( <u>88</u> ) 93	( <u>136</u> ) 145	( <u>184</u> ) 198	( <u>234</u> ) 250	( <u>283</u> ) 303

$n$  = number of samples;  $m$  = sample size.

Table XXIII Critical values of  $M(1, n)$  for the Conover non-parametric slippage test  
 $\binom{5\%}{1\%}$  (reproduced by permission of the American Statistical Association)

$n \backslash m$	2	3	4	5	6	8	10	12	14	16	18	20
4	(4) (—)	(—) (—)	(—) (—)	(—) (—)	(—) (—)	(—) (—)	(—) (—)	(—) (—)	(—) (—)	(—) (—)	(—) (—)	(—) (—)
5	(4) (5)	(5) (—)	(5) (—)	(5) (—)	(5) (—)	(5) (—)	(5) (—)	(—) (—)	(—) (—)	(—) (—)	(—) (—)	(—) (—)
6	(5) (6)	(5) (6)	(5) (6)	(6) (6)	(6) (—)	(6) (—)	(6) (—)	(6) (—)	(6) (—)	(6) (—)	(6) (—)	(6) (—)
7	(5) (6)	(5) (7)	(6) (7)	(7) (7)	(7) (7)							
8	(5) (6)	(5) (7)	(6) (7)	(6) (7)	(6) (7)	(6) (8)	(6) (8)	(7) (8)	(7) (8)	(7) (8)	(7) (8)	(7) (8)
9	(5) (6)	(6) (7)	(6) (7)	(6) (8)	(7) (8)	(7) (8)	(7) (8)	(7) (8)	(7) (8)	(7) (8)	(7) (8)	(7) (8)
10	(5) (7)	(6) (7)	(6) (7)	(6) (8)	(6) (8)	(7) (8)	(7) (8)	(7) (8)	(7) (8)	(7) (8)	(7) (8)	(7) (9)
12	(5) (7)	(6) (8)	(6) (8)	(7) (8)	(7) (8)	(7) (9)	(7) (9)	(7) (9)	(7) (9)	(8) (9)	(8) (9)	(8) (9)
14	(5) (7)	(6) (8)	(6) (8)	(7) (8)	(7) (8)	(7) (9)	(7) (9)	(7) (9)	(8) (9)	(8) (9)	(8) (9)	(8) (9)
16	(5) (7)	(6) (8)	(6) (8)	(7) (9)	(7) (9)	(7) (9)	(8) (9)	(8) (9)	(8) (10)	(8) (10)	(8) (10)	(8) (10)
18	(5) (7)	(6) (8)	(7) (9)	(7) (9)	(7) (9)	(8) (9)	(8) (10)	(8) (10)	(8) (10)	(8) (10)	(8) (10)	(8) (10)
20	(5) (7)	(6) (8)	(7) (9)	(7) (9)	(7) (9)	(8) (10)	(8) (10)	(8) (10)	(8) (10)	(8) (10)	(8) (10)	(8) (10)
25	(6) (7)	(6) (8)	(7) (9)	(7) (9)	(7) (9)	(8) (10)	(8) (10)	(8) (10)	(8) (10)	(8) (10)	(9) (10)	(9) (11)
30	(6) (8)	(6) (8)	(7) (9)	(7) (9)	(7) (9)	(8) (10)	(8) (10)	(8) (10)	(8) (10)	(9) (11)	(9) (11)	(9) (11)
35	(6) (8)	(6) (8)	(7) (9)	(7) (9)	(7) (10)	(8) (10)	(8) (10)	(8) (10)	(9) (11)	(9) (11)	(9) (11)	(9) (11)
40	(6) (8)	(6) (9)	(7) (9)	(7) (9)	(8) (10)	(8) (10)	(8) (10)	(8) (11)	(9) (11)	(9) (11)	(9) (11)	(9) (11)
$\infty$	(6) (8)	(7) (9)	(7) (10)	(8) (10)	(8) (10)	(8) (11)	(9) (11)	(9) (12)	(9) (12)	(10) (12)	(10) (12)	(10) (12)

$n$  = number of samples;  $m$  = sample size.

Table XXIV Approximate 5% critical values for downward slippage of a single variance in a set of normal samples (reproduced by permission of R. Doornbos)

<i>m</i>	<i>n</i>	2	3	4	5	6	7	8	9	10	12	15	20
2	0.00154	0.3278	0.4964	0.4444	0.4241	0.4145	0.40941	0.40645	0.40461	0.40259	0.40129	0.40530	
	3	0.02500	0.00837	0.00418	0.00251	0.00167	0.00119	0.00895	0.00696	0.00557	0.00380	0.00238	0.003132
4	0.06083	0.02489	0.01401	0.00916	0.00653	0.00493	0.00387	0.00314	0.00261	0.00189	0.00128	0.00781	
	5	0.09430	0.04262	0.02546	0.01736	0.01280	0.00992	0.00799	0.00661	0.00558	0.00418	0.00294	0.00188
6	0.12275	0.05892	0.03647	0.02550	0.01917	0.01512	0.01234	0.01033	0.00882	0.00673	0.00484	0.00318	
	7	0.14663	0.07331	0.04647	0.03306	0.02518	0.02008	0.01654	0.01395	0.01200	0.00926	0.00676	0.00453

*n* = number of samples; *m* = sample size.

Table XXV Approximate critical values and sizes for tests for upward slippage in one of  $n$  Poisson distributions  $\begin{pmatrix} .05 \\ .01 \end{pmatrix}$  (reproduced by permission of Mathematische Centrum, Amsterdam)

$t \backslash n$	2	3	4	5	6	7	8	9	10
3				.040	.028	.020	.016	.012	.010
4		.037	.016	.008	.005	.003	.002	.045	.037
5		.012	.004	.034	.020	.013	.009	.006	.005
6	.031	.004	.019	.008	.004	.035	.024	.017	.013
7	.016	.021	.005	.023	.012	.007	.004	.037	.027
8	.008	.008	.017	.006	.028	.016	.010	.006	.004
9	.008	.008	.002	.006	.003	.001	.010	.006	.004
10	.021	.010	.014	.032	.015	.008	.036	.024	.016
11	.012	.027	.030	.010	.028	.015	.008	.040	.028
12	.039	.012	.011	.020	.048	.026	.015	.009	.043
13	.022	.027	.023	.035	.015	.042	.024	.015	.009
14	.013	.012	.041	.012	.025	.012	.038	.023	.015
15	.035	.026	.017	.021	.040	.019	.010	.035	.022
16	.021	.048	.030	.035	.013	.030	.016	.009	.033
17	.049	.024	.050	.013	.021	.045	.024	.013	.047
18	.031	.044	.022	.021	.032	.014	.035	.020	.012
19	.019	.022	.036	.033	.048	.021	.050	.028	.017
20	.041	.039	.016	.050	.017	.031	.015	.040	.024
21	.027	.021	.026	.020	.026	.044	.022	.011	.033
22	.017	.035	.040	.031	.037	.015	.031	.016	.044
23	.035	.019	.019	.045	.014	.022	.042	.022	.012
24	.023	.031	.029	.019	.020	.030	.014	.030	.017
25	.043	.049	.043	.028	.029	.041	.019	.040	.023
	.004	.005	.004	.008	.008	.002	.004	.009	.005

$t$  = sum of all observations.

Table XXVI Critical values for 5% and 1% tests of discordancy of a single outlier in a multivariate normal sample where  $V$  is known and the test statistic is  $\max_{j=1, 2, \dots, n} (\mathbf{x}_j - \bar{\mathbf{x}})' V^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$  (reproduced by permission of The Institute of Statistical Mathematics)

n	$p = 2$		$p = 3$		$p = 4$	
	5%	1%	5%	1%	5%	1%
3	5.32	7.53	6.69	9.07	7.92	10.45
4	6.48	8.95	8.05	10.70	9.47	12.28
5	7.29	9.92	9.00	11.81	10.54	13.51
6	7.91	10.64	9.72	12.63	11.34	14.41
7	8.41	11.21	10.28	13.28	11.97	15.12
8	8.82	11.68	10.74	13.80	12.49	15.70
9	9.18	12.08	11.15	14.24	12.93	16.19
10	9.48	12.42	11.49	14.62	13.31	16.61
12	9.99	12.98	12.05	15.26	13.94	17.29
14	10.40	13.44	12.53	15.76	14.45	17.83
16	10.77	13.88	12.93	16.18	14.87	18.28
18	11.06	14.13	13.26	16.53	15.23	18.66
20	11.32	14.42	13.55	16.84	15.55	18.99
25	11.88	15.02	14.15	17.47	16.19	19.67
30	12.31	15.49	14.63	17.96	16.70	20.21

n = number of observations; p = dimension.

Table XXVII Critical values for 5% and 1% tests of discordancy of a single outlier in a multivariate normal sample where  $\mu$  and  $V$  are known, and the test statistic is

$$R_{(n)}(\mu, V) = \max_{j=1, 2, \dots, n} (\mathbf{x}_j - \mu)' V^{-1} (\mathbf{x}_j - \mu)$$

n	$p = 2$		$p = 4$		$p = 6$		$p = 8$		$p = 10$	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
3	8.15	11.40	12.05	15.77	15.46	19.54	18.63	23.02	21.66	26.31
4	8.73	11.98	12.72	16.42	16.20	20.24	19.43	23.76	22.50	27.10
5	9.17	12.42	13.23	16.91	16.76	20.78	20.03	24.34	23.15	27.71
6	9.53	12.79	13.65	17.32	17.22	21.22	20.53	24.81	23.67	28.21
7	9.84	13.09	14.00	17.66	17.60	21.59	20.94	25.20	24.12	28.62
8	10.11	13.36	14.30	17.96	17.94	21.91	21.30	25.54	24.49	28.98
9	10.34	13.61	14.57	18.24	18.23	22.21	21.61	25.86	24.83	29.31
10	10.55	13.81	14.81	18.46	18.49	22.45	21.89	26.11	25.12	29.58
25	12.38	15.64	16.87	20.48	20.73	24.62	24.29	28.41	27.65	31.99
50	13.77	17.02	18.41	21.99	22.40	26.24	26.06	30.12	29.52	33.78
100	15.15	18.41	19.94	23.50	24.04	27.84	27.80	31.82	31.35	35.55
200	16.54	19.80	21.46	25.00	25.67	29.44	29.52	33.49	33.16	37.30
500	18.37	21.63	23.46	26.98	27.80	31.95	31.77	35.68	35.50	39.58
1000	19.76	23.03	24.96	28.47	29.39	33.11	33.44	39.33	37.25	41.30

n = number of observations; p = dimension.

Table XXVIII Critical values for 5% and 1% tests of discordancy of a single outlier in a multivariate normal sample where  $\mu$  and  $V$  are unknown and the test statistic is

$$R_{(n)}(\tilde{\mathbf{x}}, S) = \max_{j=1, 2, \dots, n} (\mathbf{x}_j - \tilde{\mathbf{x}})' S^{-1} (\mathbf{x}_j - \tilde{\mathbf{x}})$$

n	$p = 2$		$p = 3$		$p = 4$		$p = 5$	
	5%	1%	5%	1%	5%	1%	5%	1%
5	3.17	3.19						
6	4.00	4.11	4.14	4.16				
7	4.71	4.95	5.01	5.10	5.12	5.14		
8	5.32	5.70	5.77	5.97	6.01	6.09	6.11	6.12
9	5.85	6.37	6.43	6.76	6.80	6.97	7.01	7.08
10	6.32	6.97	7.01	7.47	7.50	7.79	7.82	7.98
12	7.10	8.00	7.99	8.70	8.67	9.20	9.19	9.57
14	7.74	8.84	8.78	9.71	9.61	10.37	10.29	10.90
16	8.27	9.54	9.44	10.56	10.39	11.36	11.20	12.02
18	8.73	10.15	10.00	11.28	11.06	12.20	11.96	12.98
20	9.13	10.67	10.49	11.91	11.63	12.93	12.62	13.81
25	9.94	11.73	11.48	13.18	12.78	14.40	13.94	15.47
30	10.58	12.54	12.24	14.14	13.67	15.51	14.95	16.73
35	11.10	13.20	12.85	14.92	14.37	16.40	15.75	17.73
40	11.53	13.74	13.36	15.56	14.96	17.13	16.41	18.55
45	11.90	14.20	13.80	16.10	15.46	17.74	16.97	19.24
50	12.23	14.60	14.18	16.56	15.89	18.27	17.45	19.83
100	14.22	16.95	16.45	19.26	18.43	21.30	20.26	23.17
200	15.99	18.94	18.42	21.47	20.59	23.72	22.59	25.82
500	18.12	21.22	20.75	23.95	23.06	26.37	25.21	28.62

n = number of observations; p = dimension.

Table XXIX Critical values of  $\sqrt{r_2}$  for 5% and 1% tests of discordancy for a pair of outliers in a multivariate normal sample where  $\mu$  and  $V$  are unknown, using the test statistic

$$r_2 = \min_{i_1, i_2} R_{j_1, j_2}$$

n	$p = 2$		$p = 3$		$p = 4$		$p = 5$	
	5%	1%	5%	1%	5%	1%	5%	1%
5	0.0025	0.0005	0.0000					
6	0.0337	0.0150	0.0011	0.0002				
7	0.0860	0.0498	0.0202	0.0090	0.0006	0.0001		
8	0.1417	0.0937	0.0580	0.0335	0.0136	0.0060	0.0004	0.0001
9	0.1942	0.1393	0.1024	0.0674	0.0425	0.0245	0.0098	0.0043
10	0.2419	0.1831	0.1470	0.1049	0.0788	0.0518	0.0327	0.0189
12	0.3229	0.2616	0.2288	0.1791	0.1549	0.1163	0.0966	0.0686
14	0.3879	0.3276	0.2982	0.2460	0.2246	0.1804	0.1631	0.1270
16	0.4410	0.3828	0.3563	0.3040	0.2853	0.2389	0.2242	0.1838
18	0.4850	0.4295	0.4054	0.3542	0.3376	0.2908	0.2782	0.2360
20	0.5221	0.4694	0.4472	0.3976	0.3828	0.3366	0.3257	0.2830
25	0.5935	0.5472	0.5288	0.4839	0.4722	0.4290	0.4211	0.3798
30	0.6451	0.6041	0.5882	0.5478	0.5380	0.4984	0.4923	0.4537
35	0.6842	0.6475	0.6335	0.5969	0.5885	0.5523	0.5473	0.5116
40	0.7150	0.6818	0.6693	0.6360	0.6285	0.5953	0.5911	0.5580
45	0.7399	0.7097	0.6982	0.6677	0.6610	0.6304	0.6267	0.5961
50	0.7605	0.7328	0.7222	0.6941	0.6880	0.6596	0.6564	0.6279
100	0.8629	0.8477	0.8417	0.8260	0.8225	0.8065	0.8047	0.7883
200	0.9232	0.9152	0.9118	0.9035	0.9015	0.8929	0.8918	0.8830
500	0.9650	0.9618	0.9602	0.9568	0.9558	0.9523	0.9517	0.9480

n = number of observations; p = dimension.

**Table XXX** Critical values for 5% and 1% tests of discordancy of a single outlier in a bivariate normal sample where  $\mu$  and  $V$  are unknown, and the test statistic is

$$R_{(n)}(\bar{\mathbf{x}}, S_v) = \max_{j=1, 2, \dots, n} (\mathbf{x}_j - \bar{\mathbf{x}})' S_v^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$$

where  $S_v$  is an 'external' estimate of  $V$  (reproduced by permission of The Institute of Statistical Mathematics)

5% test

$n$	3	4	5	6	7	8	9	10	11	12	14
$\nu$											
20	6.88	8.53	9.72	10.64	11.44	12.10	12.67	13.18	13.56	14.04	14.76
22	6.72	8.30	9.45	10.36	11.10	11.73	12.28	12.76	13.19	13.58	14.26
24	6.58	8.13	9.24	10.12	10.83	11.44	11.97	12.43	12.84	13.21	13.86
26	6.47	7.98	9.06	9.92	10.61	11.20	11.71	12.16	12.55	12.91	13.54
28	6.37	7.86	8.92	9.75	10.42	11.00	11.49	11.93	12.31	12.66	13.28
30	6.29	7.75	8.79	9.61	10.27	10.83	11.31	11.74	12.11	12.45	13.05
32	6.23	7.66	8.69	9.49	10.14	10.69	11.16	11.57	11.94	12.27	12.86
34	6.17	7.58	8.60	9.38	10.02	10.56	11.02	11.43	11.79	12.12	12.69
36	6.12	7.51	8.51	9.29	9.92	10.45	10.91	11.31	11.67	11.98	12.55
38	6.07	7.45	8.44	9.21	9.83	10.35	10.81	11.20	11.55	11.87	12.42
40	6.03	7.40	8.38	9.14	9.75	10.27	10.71	11.10	11.45	11.76	12.30
45	5.94	7.29	8.25	8.99	9.59	10.09	10.53	10.91	11.24	11.54	12.07
50	5.88	7.20	8.14	8.87	9.46	9.95	10.38	10.75	11.08	11.37	11.89
55	5.82	7.13	8.06	8.78	9.36	9.84	10.26	10.63	10.95	11.24	11.74
60	5.78	7.07	7.99	8.70	9.27	9.75	10.16	10.52	10.84	11.13	11.62
100	5.59	6.82	7.70	8.37	8.91	9.36	9.75	10.09	10.39	10.65	11.12
150	5.50	6.71	7.56	8.21	8.74	9.18	9.55	9.88	10.17	10.43	10.88
200	5.45	6.65	7.49	8.13	8.64	9.09	9.46	9.78	10.06	10.32	10.76

1% test

Table XXX (Continued)

$\nu \backslash n$	3	4	5	6	7	8	9	10	11	12	14
20	10.72	12.99	14.61	15.85	16.86	17.71	18.44	19.08	19.66	20.17	21.07
22	10.36	12.53	14.07	15.24	16.20	17.00	17.69	18.29	18.83	19.31	20.16
24	10.07	12.16	13.63	14.76	15.68	16.44	17.10	17.67	18.18	18.64	19.44
26	9.84	11.86	13.28	14.37	15.25	15.98	16.62	17.16	17.66	18.09	18.86
28	9.64	11.63	12.99	14.05	14.90	15.62	16.22	16.75	17.22	17.64	18.37
30	9.47	11.40	12.74	13.77	14.60	15.30	15.88	16.40	16.85	17.26	17.97
32	9.33	11.22	12.54	13.54	14.35	15.02	15.60	16.10	16.54	16.94	17.63
34	9.21	11.06	12.36	13.34	14.13	14.79	15.35	15.84	16.28	16.66	17.34
36	9.10	10.93	12.20	13.17	13.95	14.59	15.14	15.62	16.04	16.42	17.08
38	9.01	10.81	12.06	13.01	13.78	14.41	14.95	15.42	15.84	16.21	16.86
40	8.93	10.70	11.94	12.88	13.63	14.26	14.79	15.25	15.66	16.03	16.66
45	8.75	10.48	11.69	12.60	13.33	13.93	14.45	14.89	15.29	15.64	16.25
50	8.62	10.31	11.49	12.38	13.09	13.68	14.18	14.62	15.00	15.34	15.93
55	8.51	10.18	11.33	12.21	12.90	13.48	13.97	14.39	14.77	15.10	15.68
60	8.42	10.07	11.20	12.06	12.75	13.32	13.80	14.21	14.58	14.91	15.48
100	8.05	9.59	10.66	11.46	12.10	12.63	13.07	13.45	13.79	14.09	14.61
150	7.87	9.37	10.40	11.18	11.79	12.30	12.73	13.10	13.42	13.71	14.21
200	7.79	9.26	10.28	11.04	11.62	12.14	12.57	12.92	13.24	13.52	14.01

 $n$  = number of observations;  $\nu$  = number of degrees of freedom.

Table XXXI 5% and 1% critical values for the maximum normed residual, for testing the discordancy of a single outlier in a  $r \times c$  factorial experiment (reproduced by permission of the American Statistical Association and the American Society for Quality Control)

*5% critical values*

$c \backslash r$	3	4	5	6	7	8	9
3	0.648	0.645	0.624	0.600	0.577	0.555	0.535
4		0.621	0.590	0.561	0.535	0.513	0.493
5			0.555	0.525	0.499	0.477	0.457
6				0.495	0.469	0.447	0.428
7					0.444	0.423	0.405
8						0.402	0.385
9							0.368

*1% critical values*

$c \backslash r$	3	4	5	6	7	8	9
3	0.660	0.675	0.664	0.646	0.626	0.606	0.587
4		0.665	0.640	0.613	0.588	0.565	0.544
5			0.608	0.578	0.551	0.527	0.506
6				0.546	0.519	0.495	0.475
7					0.492	0.469	0.449
8						0.446	0.426
9							0.407

Table XXXII Critical values for 5% and 1% tests of discordancy for a single outlier in a general linear model with normal error structure, using the studentized residual as test statistic (reproduced by permission of the American Statistical Association and the American Society for Quality Control)

5% critical values

$n \backslash q$	1	2	3	4	5	6	8	10	15	25
5	1.92									
6	2.07	1.93								
7	2.19	2.08	1.94							
8	2.28	2.20	2.10	1.94						
9	2.35	2.29	2.21	2.10	1.95					
10	2.42	2.37	2.31	2.22	2.11	1.95				
12	2.52	2.49	2.45	2.39	2.33	2.24	1.96			
14	2.61	2.58	2.55	2.51	2.47	2.41	2.25	1.96		
16	2.68	2.66	2.63	2.60	2.57	2.53	2.43	2.26		
18	2.73	2.72	2.70	2.68	2.65	2.62	2.55	2.44		
20	2.78	2.77	2.76	2.74	2.72	2.70	2.64	2.57	2.15	
25	2.89	2.88	2.87	2.86	2.84	2.83	2.80	2.76	2.60	
30	2.96	2.96	2.95	2.94	2.93	2.93	2.90	2.88	2.79	2.17
35	3.03	3.02	3.02	3.01	3.00	3.00	2.93	2.97	2.91	2.64
40	3.08	3.08	3.07	3.07	3.06	3.06	3.05	3.03	3.00	2.84
45	3.13	3.12	3.12	3.12	3.11	3.11	3.10	3.09	3.06	2.96
50	3.17	3.16	3.16	3.16	3.15	3.15	3.14	3.14	3.11	3.04
60	3.23	3.23	3.23	3.23	3.22	3.22	3.22	3.21	3.20	3.15
70	3.29	3.29	3.28	3.28	3.28	3.28	3.27	3.27	3.26	3.23
80	3.33	3.33	3.33	3.33	3.33	3.33	3.32	3.32	3.31	3.29
90	3.37	3.37	3.37	3.37	3.37	3.37	3.36	3.36	3.36	3.34
100	3.41	3.41	3.40	3.40	3.40	3.40	3.40	3.40	3.39	3.38

Table XXXII (Continued)

1% critical values

$n \backslash q$	1	2	3	4	5	6	8	10	15	25
5	1.98									
6	2.17	1.98								
7	2.32	2.17	1.98							
8	2.44	2.32	2.18	1.98						
9	2.54	2.44	2.33	2.18	1.99					
10	2.62	2.55	2.45	2.33	2.18	1.99				
12	2.76	2.70	2.64	2.56	2.46	2.34	1.99			
14	2.86	2.82	2.78	2.72	2.65	2.57	2.35	1.99		
16	2.95	2.92	2.88	2.84	2.79	2.73	2.58	2.35		
18	3.02	3.00	2.97	2.94	2.90	2.85	2.75	2.59		
20	3.08	3.06	3.04	3.01	2.98	2.95	2.87	2.76	2.20	
25	3.21	3.19	3.18	3.16	3.14	3.12	3.07	3.01	2.75	
30	3.30	3.29	3.28	3.26	3.25	3.24	3.21	3.17	3.04	2.21
35	3.37	3.36	3.35	3.34	3.34	3.33	3.30	3.25	3.19	2.81
40	3.43	3.42	3.42	3.41	3.40	3.40	3.38	3.36	3.30	3.05
45	3.48	3.47	3.47	3.46	3.46	3.45	3.44	3.43	3.38	3.23
50	3.52	3.52	3.51	3.51	3.51	3.50	3.49	3.48	3.45	3.34
60	3.60	3.59	3.59	3.59	3.58	3.58	3.57	3.56	3.54	3.48
70	3.65	3.65	3.65	3.65	3.64	3.64	3.64	3.63	3.61	3.57
80	3.70	3.70	3.70	3.70	3.69	3.69	3.69	3.68	3.67	3.64
90	3.74	3.74	3.74	3.74	3.74	3.74	3.73	3.73	3.72	3.70
100	3.78	3.78	3.78	3.77	3.77	3.77	3.77	3.77	3.76	3.74

 $n$  = number of observations. $q$  = number of independent variables (including count for intercept if fitted).

## *References and Bibliography*

Most of the works listed here have been referred to in the text; the pages on which principal or substantial mention is made of any work are shown in parentheses at the end of the reference. An example is:

- Behnken, D. W., and Draper, N. R. (1972). 'Residuals and their variance patterns'. *Technometrics*, **14**, 101–111. (253, 255, 261)

Additional works which have not been specifically mentioned in the text, but which are likely to assist with further study of the history or development of outlier problems are also listed. Where appropriate we have indicated the area of relevance by means of a chapter reference, accompanied by the symbol *H* if the work is of particular historical interest, for example:

- Glaisher, J. W. L. (1874). Note on a paper by Mr. Stone 'On the rejection of discordant observations'. *Monthly Notices Roy. Astr. Soc.*, **34**, 251. (Ch. 2, *H*)

- Acton, F. S. (1959). *Analysis of Straight-Line Data*, Wiley, New York. (Ch. 7)  
Adichie, J. N. (1967a). 'Asymptotic efficiency of a class of non-parametric tests for regression parameters'. *Ann. Math. Statist.*, **38**, 884–893. (256)  
Adichie, J. N. (1967b). 'Estimates of regression parameters based on rank tests'. *Ann. Math. Statist.*, **38**, 894–903. (256)  
Airy, G. B. (1856). Letter from Professor Airy, Astronomer Royal, to the Editor. *Astr. J.*, **4**, 137–138. (Ch. 2, *H*)  
Airy, G. B. (1861). *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations*. MacMillan, London. (Ch. 2, *H*)  
Allen, G. C. (1961). See Bernoulli, D. (1777).  
Andrews, D. F. (1971). 'Significance tests based on residuals'. *Biometrika*, **58**, 139–148. (260)  
Andrews, D. F. (1972). 'Plots of high-dimensional data'. *Biometrics*, **28**, 125–136. (227)  
Andrews, D. F. (1973). 'Robust estimation for multiple linear regression models'. *Bull. Int. Statist. Inst.*, **45**, 105–111. (Ch. 7)  
Andrews, D. F. (1974). 'A robust method for multiple linear regression'. *Technometrics*, **16**, 523–531. (Ch. 7)

- Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, N.J. (48, 135, 150–155, 157, 158, 163–165)
- Anonymous (1821). ‘Dissertation sur la recherche du milieu le plus probable, entre les résultats de plusieurs observations ou expériences’. *Annales de Mathématiques Pure et Appliquées*, **12**, 181–204. (Ch. 2, *H*)
- Anscombe, F. J. (1960a). ‘Rejection of outliers’. *Technometrics*, **2**, 123–147. (26, 27, 34, 50, 127, 131, 231, 240, 246, 247, 260)
- Anscombe, F. J. (1960b). ‘Discussion by Kruskal, W., Ferguson, T. S., Tukey, J. W., Gumbel, E. J. and Anscombe, F. J.’. *Technometrics*, **2**, 157–166. (Ch. 2)
- Anscombe, F. J. (1961). ‘Examination of residuals’. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 1–36. (247, 259)
- Anscombe, F. J. (1968). ‘Statistical analysis, special problems of outliers’. In *International Encyclopaedia Social Sciences*. MacMillan, New York, Vol. 15, pp. 178–182. (Ch. 2)
- Anscombe, F. J., and Barron, B. A. (1966). ‘Treatment of outliers in samples of size three’. *J. Res. Nat. Bur. Standards, B*, **70**, 141–147. (51, 126, 132, 135)
- Anscombe, F. J., and Tukey, J. W. (1963). ‘The examination and analysis of residuals’. *Technometrics*, **5**, 141–160. (247, 248)
- Ansell, M. (1973). ‘Robustness of location estimators to asymmetry’. *Applied Statistics*, **22**, 249–254. (Ch. 4)
- Antille, A. (1974). ‘A linearized version of the Hodges–Lehmann estimator’. *Ann. Statist.*, **2**, 1308–1313 (Ch. 4)
- Arley, N. (1940). ‘On the distribution of relative errors from a normal population of errors. A discussion of some problems in the theory of errors’. *Mathematisk-Fysiske Meddelelser udgivet af det Kgl. Danske Videnskabernes Selskab*, **18**. (Ch. 2)
- Arley, N., and Buch, K. (1950). *Introduction to the Theory of Probability and Statistics*. Chapman & Hall, London; Wiley, New York. (Ch. 2)
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York. (187)
- Barnett, V. D. (1966). ‘Order statistics estimators of the location of the Cauchy distribution’. *J. Amer. Statist. Ass.*, **61**, 1205–1218. (48)
- Barnett, V. D. (1976). ‘The ordering of multivariate data’ (with Discussion). *J. Roy. Statist. Soc. A*, **139**, (208)
- Barnett, V. D., and Lewis, T. (1967). ‘A study of low-temperature probabilities in the context of an industrial problem’ (with Discussion). *J. Roy. Statist. Soc. A*, **130**, 177–206. (6)
- Basu, A. P. (1965). ‘On some tests of hypotheses relating to the exponential distribution when some outliers are present’. *J. Amer. Statist. Ass.*, **60**, 548–559. *Corr. J. Amer. Statist. Ass.*, **60**, 1249. (77)
- Bechhofer, R. E., Kiefer, J., and Sobel, M. (1968). *Sequential Identification and Ranking Procedures*. The University of Chicago Press, Chicago. (187)
- Beckman, R. J., and Trussell, H. J. (1974). ‘The distribution of an arbitrary studentized residual and the effects of updating in multiple regression’. *J. Amer. Statist. Ass.*, **69**, 199–201. (Ch. 7)
- Begg, T. B., Preston, S. R., and Healy, M. J. R. (1966). ‘The dietary habits of patients with occlusive arterial disease’. *Atti V Conv. internat. Asp. diet. Inf. Senesc.*, pp. 66–79. (212)
- Behnken, D. W., and Draper, N. R. (1972). ‘Residuals and their variance patterns’. *Technometrics*, **14**, 101–111. (253, 255, 261)
- Beran, R. (1974). ‘Asymptotically efficient adaptive rank estimates in location models’. *Ann. Statist.*, **2**, 63–74. (Ch. 4)

- Berman, S. (1962). 'Limiting distribution of the Studentized largest observation'. *Skand. Akt.*, **45**, 154–161. (Ch. 3)
- Bernoulli, D. (1777). 'Djudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda'. *Acta Academiae Scientiarum Petropolitanae*, **1**, 3–33. English translation by C. G. Allen (1961), *Biometrika*, **48**, 3–13. (18)
- Bernoulli, J. III (1785). 'Milieu'. *Encyclopédie Méthodique*, II, 404–409. (Ch. 2, H)
- Bertrand, J. (1888a). 'Sur la loi de probabilité des erreurs d'observation'. *C. R. Acad. Sci. Paris*, **106**, 153–156. (Ch. 2, H)
- Bertrand, J. (1888b). 'Sur la combinaison des mesures d'une même grandeur'. *C. R. Acad. Sci. Paris*, **106**, 701–704. (Ch. 2, H)
- Bertrand, J. (1889). *Calcul des Probabilités*. Gauthier-Villars, Paris. (Ch. 2, H)
- Bessel, F. W., and Baeuer, J. J. (1838). *Gradmessung in Ostpreussen und ihre Verbindung mit Preussischen und Russischen Dreiecksketten*. Berlin. Reprinted in *Abhandlungen von F. W. Bessel* (ed. R. Engleman), Leipzig, 1876.
- Bickel, P. J. (1965). 'On some robust estimates of location'. *Ann. Math. Statist.*, **36**, 847–858. (46, 135, 152, 156, 157)
- Bickel, P. J. (1967). 'Some contributions to the theory of order statistics'. *Proc. Fifth Berkely Symp. Math. Statist. Prob.* Vol. 1, 575–591. (167)
- Bickel, P. J., and Hodges, J. L. Jr. (1967). 'The asymptotic theory of Galton's test and a related simple estimate of location'. *Ann. Math. Statist.*, **38**, 73–89. (49, 165)
- Birnbaum, A. (1959). 'On the analysis of factorial experiments without replication'. *Technometrics*, **1**, 343–357. (248)
- Birnbaum, A., and Laska, E. M. (1967). 'Optimal robustness: A general method with applications to linear estimators of location'. *J. Amer. Statist. Ass.*, **62**, 1230–1240. (135)
- Birnbaum, A., and Miké, V. (1970). 'Asymptotically robust estimators of location'. *J. Amer. Statist. Ass.*, **65**, 1265–1282. (153)
- Birnbaum, A., Laska, E. M., and Meisner, M. (1971). 'Optimally robust linear estimators of location'. *J. Amer. Statist. Ass.*, **66**, 302–310. (135)
- Bliss, C. I., Cochran, W. G., and Tukey, J. W. (1956). 'A rejection criterion based upon the range'. *Biometrika*, **43**, 418–422. (197)
- Bofinger, V. J. (1965). 'The  $k$ -sample slippage problem'. *Austral. J. Statist.*, **7**, 20–31. (178, 179, 182, 184)
- Boscovich, R. J. (1757). 'De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa. Bononiensi Scientiarum et Artum Instituto Atque Academia Commentarii', **4**, 353–396. (Ch. 2, H)
- Bowley, A. L. (1928). *F. Y. Edgeworth's Contributions to Mathematical Statistics*. Royal Statistical Society, London. (Ch. 2, H)
- Box, G. E. P., and Tiao, G. C. (1962). 'A further look at robustness via Bayes' theorem'. *Biometrika*, **49**, 419–432. (279)
- Box, G. E. P., and Tiao, G. C. (1968). 'A Bayesian approach to some outlier problems'. *Biometrika*, **55**, 119–129. (32, 46, 155, 252, 275, 276, 277)
- Bross, I. D. J. (1961). 'Outliers in patterned experiments: a strategic re-appraisal'. *Technometrics*, **3**, 91–102. (235, 238, 251)
- Brown, B. M. (1975). 'A short-cut test for outliers using residuals'. *Biometrika*, **62**, 623–629. (252)
- Brunt, D. (1917). *The Combination of Observations*. University Press, Cambridge. (2nd edn. 1931). (Ch. 2, H)
- Cacoullos, T. (1968). 'A sequential scheme for detecting outliers'. *Bulletin de la Société Mathématique de Grèce*, **9**, 113–123. (Ch. 3)
- Calvin, M., Heidelberger, C., Reid, J. C., Tolbert, B. M., and Yankwich, P. F.

- (1949). *Isotopic Carbon: Techniques in its measurement and chemical manipulation*. Wiley, New York. (2)
- Chambers, C. (1967). 'Extension of tables of percentage points of the largest variance ratio  $S_{\max}^2/S_0^2$ '. *Biometrika*, **54**, 225–228. (191)
- Chandra Sekar, C., and Francis, M. G. (1941). 'A method to get the significance limit of a type of test criteria'. *Sankhya*, **5**, 165–168. (191)
- Chase, G. R., and Bulgren, W. G. (1971). 'A Monte Carlo investigation of the robustness of  $T^2$ '. *J. Amer. Statist. Ass.*, **66**, 499–503. (Ch. 4)
- Chatfield, C. (1974). Personal correspondence. (7)
- Chatfield, C. (1975). *The Analysis of Time Series: Theory and Practice*. Chapman and Hall, London (p. 102). (11)
- Chatfield, C., Ehrenberg, A. S. C., and Goodhardt, G. J. (1966). 'Progress on a simplified model of stationary purchasing behaviour' (with Discussion). *J. Roy. Statist. Soc. A*, **129**, 317–367. (9)
- Chauvenet, W. (1863). 'Method of least squares'. Appendix to *Manual of Spherical and Practical Astronomy*, Vol. 2, Lippincott, Philadelphia, pp. 469–566; tables 593–599. Reprinted (1960) 5th edn. Dover, New York. (2, 19, 24)
- Chedzoy, O. B. (1973). Paper read at annual conference of Royal Statistical Society, Newcastle upon Tyne. (13)
- Chen, E. H. (1971). 'The power of the Shapiro–Wilks  $W$ -test for normality in samples from contaminated distributions'. *J. Amer. Statist. Ass.*, **66**, 760–762. (249)
- Chen, E. H., and Dixon, W. J. (1972). 'Estimates of parameters of a censored regression sample'. *J. Amer. Statist. Ass.*, **67**, 664–671. (Ch. 7)
- Chernoff, H., Gastwirth, J. L., and Johns, M. V. Jr. (1967). 'Asymptotic distribution of linear combinations of order statistics'. *Ann. Math. Statist.*, **31**, 52–72. (153)
- Chew, V. (1964). 'Tests for the rejection of outlying observations'. *RCA Systems Analysis Technical Memorandum No. 64-7*, Missile Test Project, Patrick Air Force Base, Florida. (Ch. 3)
- Claridge, P. N., and Potter, I. C. (1974). 'Heart ratios at different stages in the life cycle of lampreys'. *Acta Zoologica*, **55**, 61–69. (175)
- Cochran, W. G. (1941). 'The distribution of the largest of a set of estimated variances as a fraction of their total'. *Ann. Eugen.*, **11**, 47–52. (79, 108, 191)
- Cochran, W. G. (1968). 'Errors of measurement in statistics'. *Technometrics*, **10**, 637–666. (Ch. 3)
- Collett, D., and Lewis, T. (1976). 'The subjective nature of outlier rejection procedures'. *Applied Statistics*, **25**, 228–237. (64)
- Conover, W. J. (1965). 'Several  $k$ -sample Kolmogorov–Smirnov tests'. *Ann. Math. Statist.*, **36**, 1019–1026. (Ch. 5)
- Conover, W. J. (1968). 'Two  $k$ -sample slippage tests'. *J. Amer. Statist. Ass.*, **63**, 614–626. (183, 184)
- Coolidge, J. L. (1925). *An Introduction to Mathematical Probability*. Oxford University Press, London. Reprint (1962) Dover, New York. (Ch. 2, H)
- Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London. (48)
- Cox, D. R., and Snell, E. J. (1968). 'A general definition of residuals' (with Discussion). *J. Roy. Statist. Soc. B*, **30**, 248–275. (261)
- Cox, D. R., and Snell, E. J. (1971). 'On test statistics calculated from residuals'. *Biometrika*, **58**, 589–594. (261)
- Crow, E. L., and Siddiqui, M. M. (1967). 'Robust estimation of location'. *J. Amer. Statist. Ass.*, **62**, 353–389. (135, 156, 157)
- Cucconi, O. (1962). 'Un criterio per il rigetto delle osservazioni spurie'. *Scuola in Azione*, **21**, 92–106. (Ch. 3)
- Czuber, E. (1891). *Theorie der Beobachtungsfehler*. Teubner, Leipzig. (Ch. 2, H)

- Daniel, C. (1959). 'Use of half-normal plots in interpreting factorial two-level experiments'. *Technometrics*, **1**, 311–341. (25, 248, 274)
- Daniel, C. (1960). 'Locating outliers in factorial experiments'. *Technometrics*, **2**, 149–156. (235, 238, 240)
- Daniell, P. J. (1920). 'Observations weighted according to order'. *Amer. J. Math.*, **42**, 222–236. (Ch. 2, H)
- Darling, D. A. (1952). 'On a test for homogeneity and extreme values'. *Ann. Math. Statist.*, **23**, 450–456. (Ch. 3)
- David, H. A. (1952). 'Upper 5 and 1% points of the maximum *F*-ratio'. *Biometrika*, **39**, 422–424. (84)
- David, H. A. (1956a). 'On the application to statistics of an elementary theorem in probability'. *Biometrika*, **43**, 85–91. (105)
- David, H. A. (1956b). 'Revised upper percentage points of the extreme studentized deviate from the sample mean'. *Biometrika*, **43**, 449–451. (105, 110, 111)
- David, H. A. (1962). 'Order statistics in short-cut tests'. In Sarhan and Greenberg (1962). (107)
- David, H. A. (1970). *Order Statistics*. Wiley, New York. (43, 44, 45, 65–67, 73, 152, 159, 192)
- David, H. A., and Paulson, A. S. (1965). 'The performance of several tests for outliers'. *Biometrika*, **52**, 429–436. (45, 94, 104, 105, 111)
- David, H. A., Hartley, H. O., and Pearson, E. S. (1954). 'The distribution of the ratio, in a single normal sample, of range to standard deviation'. *Biometrika*, **41**, 482–493. (39, 97)
- David, F. N., Barton, D. E., Ganeshalingam, S., Harter, H. L., Kim, P. J., Merrington, M., and Walley, D. (1968). *Normal Centroids, Medians and Scores for Ordinal Data*, Tracts for computers No. XXIX, Cambridge University Press, London. (139)
- Dawid, A. P., Stone, M., and Zidek, J. V. (1973). 'Marginalisation paradoxes in Bayesian and structural inference' (with Discussion). *J. Roy. Statist. Soc. B*, **35**, 189–233. (273)
- DeAlba Guerra, E., and Van Ryzin, J. (1974). 'An empirical Bayes approach to the outlier problem' (Abstract). *Inst. Math. Statist. Bull.*, **3**, 125. (Ch. 8)
- De Finetti, B. (1961). 'The Bayesian approach to the rejection of outliers'. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 199–210. (46, 270)
- Dempster, A. P., and Rosner, B. (1971). 'Detection of outliers'. In Gupta and Yackel (1971). (46, 272)
- Desu, M. M., Gehan, E. A., and Severo, N. C. (1974). 'A two-stage estimation procedure when there may be spurious observations'. *Biometrika*, **61**, 593–599. (51, 132)
- Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1975). 'Robust estimation and outlier detection with correlation coefficients'. *Biometrika*, **62**, 531–545. (228, 233)
- Dixon, W. J. (1950). 'Analysis of extreme values'. *Ann. Math. Statist.*, **21**, 488–506. (34, 54, 66, 69, 70, 87, 94, 96, 98, 99, 100, 107, 110, 111, 114)
- Dixon, W. J. (1951). 'Ratios involving extreme values'. *Ann. Math. Statist.*, **22**, 68–78. (38, 54, 87, 98, 99, 100)
- Dixon, W. J. (1953). 'Processing data for outliers'. *Biometrics*, **9**, 74–89. (33, 41, 73)
- Dixon, W. J. (1960). 'Simplified estimation from censored normal samples'. *Ann. Math. Statist.*, **31**, 385–391. (135, 160, 169)
- Dixon, W. J. (1962). 'Rejection of observations'. In Sarhan and Greenberg (1962). (94, 111, 114)
- Dixon, W. J. (1964). 'Query 4: Rejection of outlying values'. *Technometrics*, **6**, 238. (Ch. 3)

- Dixon, W. J., and Tukey, J. W. (1968). 'Approximate behaviour of the distribution of Winsorized  $t$  (Trimming/Winsorization 2)'. *Technometrics*, **10**, 83–98. (142, 159, 168, 169)
- Doolittle, H. M. (1884). 'The rejection of doubtful observations' (Abstract). *Bulletin of the Philosophical Society of Washington, (Math. Soc.)*, **6**, 153–156. (Ch. 2, *H*)
- Doornbos, R. (1956). 'Significance of the smallest of a set of estimated normal variances'. *Statistica Neerlandica*, **10**, 117–126. (191, 197)
- Doornbos, R. (1966). *Slippage Tests*. Mathematical Centre Tracts, No. 15, Mathematisch Centrum, Amsterdam. (120, 187, 191, 193, 195, 197, 200, 201–204)
- Doornbos, R., and Prins, H. J. (1956). 'Slippage tests for a set of gamma-variates'. *Indag. Math.*, **18**, 329–337. (202)
- Doornbos, R., and Prins, H. J. (1958). 'On slippage tests'. *Indag. Math.*, **20**, 38–55, 438–447. (181, 195, 197)
- Downton, F. D. (1976). 'Nonparametric tests for block experiments'. *Biometrika*, **63**, 137–141. (187)
- Draper, N. R., and Smith, H. (1966). *Applied Regression Analysis*. Wiley, New York. (261)
- Edgeworth, F. Y. (1883). 'The method of least squares'. *Philosophical Magazine*, **16**, 360–375. (20)
- Edgeworth, F. Y. (1887). 'On discordant observations'. *Philosophical Magazine*, **23**, 364–375. (20)
- van Eeden, C. (1970). 'Efficiency-robust estimation of location'. *Ann. Math. Statist.*, **41**, 172–181. (Ch. 4)
- Eisenhart, C., and Solomon, H. (1947). 'Significance of the largest of a set of sample estimates of variance'. In Eisenhart, Hastay and Wallis (1947). (191)
- Eisenhart, C., Hastay, M. W., and Wallis, W. A. (Eds.) (1947). *Selected Techniques of Statistical Analysis*, McGraw-Hill, New York. (79, 108, 191, 196, 197)
- Elashoff, J. D. (1972). 'A model for quadratic outliers in linear regression'. *J. Amer. Statist. Ass.*, **67**, 478–485. (255)
- Ellenberg, J. H. (1973). 'The joint distribution of the standardized least squares residuals from a general linear regression'. *J. Amer. Statist. Ass.*, **68**, 941–943. (255, 262, 264, 265)
- Ellenberg, J. H. (1976). 'Testing for a single outlier from a general linear regression'. *Biometrics*, **32**, 637–645. (264, 265)
- Epstein, B. (1960a). 'Tests for the validity of the assumption that the underlying distribution of life is exponential: Part I'. *Technometrics*, **2**, 83–101. (40, 77)
- Epstein, B. (1960b). 'Tests for the validity of the assumption that the underlying distribution of life is exponential: Part II'. *Technometrics*, **2**, 167–183. (40, 77, 85)
- Faye, H. E. (1888). 'Sur certain points de la théorie des erreurs accidentielles'. *C. R. Acad. Sci. Paris*, **106**, 783–786. (Ch. 3, *H*)
- Fellegi, I. P. (1975). 'Automatic editing and imputation of quantitative data' (Summary). *Bull. Int. Statist. Inst.*, **XLVI**, 249–253. (221, 224)
- Fellegi, I. P., and Holt, D. (1976). 'A systematic approach to automatic edit and imputation'. *J. Amer. Statist. Ass.*, **71**, 17–35. (Ch. 6)
- Feller, W. (1968). *An Introduction to Probability Theory and its Applications*. Vol. I, 3rd edn. Wiley, New York. (192)
- Fenton, R. (1975). Personal correspondence. (13)
- Ferguson, T. S. (1961a). 'On the rejection of outliers'. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 253–287. (1, 34, 39, 41, 42, 58, 60, 94, 95, 98, 101, 109, 210, 218, 240, 276)
- Ferguson, T. S. (1961b). 'Rules for rejection of outliers'. *Rev. Inst. Int. de Statist.*, **29**, 29–43. (73, 94–98, 100, 111)

- Fieller, N. R. J. (1976). *Some Problems related to the Rejection of Outlying Observations*. Ph.D. Thesis, University of Sheffield. (71, 83, 85, 95–97, 106–108, 110–112, 251, 264)
- Finney, D. J. (1974). 'Problems, data and inference: The Address of the President' (with Proceedings). *J. Roy. Statist. Soc. A*, **137**, 1–23, (6)
- Fisher, R. A. (1929). 'Tests of significance in harmonic analysis'. *Proc. Roy. Soc. A*, **125**, 54–59. (64, 79)
- Fisher, R. A. (1936). 'The use of multiple measurements in taxonomic problems'. *Ann. Eugen.*, **7**, 179–188. (227)
- Fisher, R. A. (1960). *The Design of Experiments*. 7th edn. Oliver & Boyd, Edinburgh. (279)
- Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). 'The relation between the number of species and the number of individuals in a random sample of an animal population'. *J. Animal Ecol.*, **12**, 42–57. (10)
- Forsythe, A. B. (1972). 'Robust estimation of straight line regression coefficients by minimizing  $p$ th power deviations'. *Technometrics*, **14**, 159–166. (Ch. 7)
- Fox, A. J. (1972). 'Outliers in time series'. *J. Roy. Statist. Soc. B*, **43**, 350–363. (267)
- Friedman, M. (1940). 'A comparison of alternative tests of significance for the problem of  $m$ -rankings'. *Ann. Math. Statist.*, **11**, 86–92. (187)
- Gastwirth, J. L. (1966). 'On robust procedures'. *J. Amer. Statist. Ass.*, **61**, 929–948. (48, 135, 153, 164, 166)
- Gastwirth, J. L., and Cohen, M. L. (1970). 'Small sample behaviour of some robust linear estimators of location'. *J. Amer. Statist. Ass.*, **65**, 946–973. (135, 153, 157, 164)
- Gastwirth, J. L., and Rubin, H. (1969). 'On robust linear estimators'. *Ann. Math. Statist.*, **40**, 24–39. (156)
- Gebhardt, F. (1964). 'On the risk of some strategies for outlying observations'. *Ann. Math. Statist.*, **35**, 1524–1536. (51)
- Gebhardt, F. (1966). 'On the effect of stragglers on the risk of some mean estimators in small samples'. *Ann. Math. Statist.*, **37**, 441–450. (Ch. 4)
- Gentleman, J. F., and Wilk, M. B. (1975a). 'Detecting Outliers: II Supplementing the direct analysis of residuals'. *Biometrics*, **31**, 387–410. (249, 251, 265)
- Gentleman, J. F., and Wilk, M. B. (1975b). 'Detecting outliers in a two-way table. I. Statistical behaviour of residuals'. *Technometrics*, **17**, 1–14. (249)
- Glaisher, J. W. L. (1872). 'On the law of facility of errors of observations and on the method of least squares'. *Mem. Roy. Astr. Soc.*, **39**, 75–124. (47)
- Glaisher, J. W. L. (1872–73). 'On the rejection of discordant observations'. *Monthly Notices Roy. Astr. Soc.*, **33**, 391–402. (20)
- Glaisher, J. W. L. (1874). Note on a paper by Mr. Stone 'On the rejection of discordant observations'. *Monthly Notices Roy. Astr. Soc.*, **34**, 251. (Ch. 2, H)
- Gnanadesikan, R. (1973). 'Graphical methods for informal inference in multivariate data analysis'. *Bull. Int. Statist. Inst.*, **45**, Book 4, 195–206. (222, 227)
- Gnanadesikan, R., and Kettenring, J. R. (1972). 'Robust estimates, residuals and outlier detection with multiresponse data'. *Biometrics*, **28**, 81–124. (208, 220, 222–224, 226, 227, 232, 249)
- Gnanadesikan, R., and Wilk, M. B. (1968). 'Probability plotting methods for the analysis of data'. *Biometrika*, **55**, 1–17. (Ch. 6)
- Gnanadesikan, R., and Wilk, M. B. (1969). 'Data analytic methods in multivariate statistical analysis'. In Krishnaiah (1969). (222)
- Goldsmith, P. L., and Boddy, R. (1973). 'Critical analysis of factorial experiments and orthogonal fractions'. *Applied Statistics*, **22**, 141–160. (241, 244, 264)
- Golub, G. H., Guttman, I., and Dutter, R. (1973). 'Examination of pseudo-residuals of outliers for detecting spuriousity in the general univariate linear model'. In Kabe and Gupta (1973). (231, 260)

- Goodwin, H. M. (1913). *Elements of the Precision of Measurements and Graphical Methods*. McGraw-Hill, New York. (21)
- Gould, B. A. Jr. (1855). 'On Peirce's criterion for the rejection of doubtful observations, with tables for facilitating its application'. *Astr. J.*, **4**, 81–87. (Ch. 2, H)
- Granger, C. W. J., and Neave, H. R. (1968). 'A quick test for slippage'. *Rev. Int. Statist. Inst.*, **36**, 309–312. (180, 183)
- Green, R. F. (1974). 'A note on outlier-prone families of distributions'. *Ann. Statist.*, **2**, 1293–1295. (37)
- Green, R. F. (1976). 'Outlier-prone and outlier-resistant distributions'. *J. Amer. Statist. Ass.*, **71**, 502–505. (Ch. 2)
- Grubbs, F. E. (1950). 'Sample criteria for testing outlying observations'. *Ann. Math. Statist.*, **21**, 27–58. (23, 34, 39, 40, 54, 73, 94, 96, 97, 110, 111, 2, 15, 254)
- Grubbs, F. E. (1969). 'Procedures for detecting outlying observations in samples'. *Technometrics*, **11**, 1–21. (22, 26, 40, 94, 96, 111)
- Grubbs, F. E., and Beck, G. (1972). 'Extension of sample sizes and percentage points for significance tests of outlying observations'. *Technometrics*, **14**, 847–854. (94, 96)
- Gupta, S. S. (1960). 'Order statistics from the gamma distribution'. *Technometrics*, **2**, 243–262. Correction *Technometrics*, **2**, 523. (211)
- Gupta, S. S. (Ed.) (1975). *Applied Statistics*. North Holland, Amsterdam.
- Gupta, S. S., and Yackel, J. (Eds.) (1971). *Statistical Decision Theory and Related Topics*. Academic Press, New York.
- Guttman, I. (1973a). 'Premium and protection of several procedures for dealing with outliers when sample sizes are moderate to large'. *Technometrics*, **15**, 385–404. (127, 132, 135, 166, 168)
- Guttman, I. (1973b). 'Care and handling of univariate or multivariate outliers in detecting spuriousness—a Bayesian approach'. *Technometrics*, **15**, 723–738. (32, 34, 36, 45, 210, 232, 272, 275)
- Guttman, I., and Smith, D. E. (1969). 'Investigation of rules for dealing with outliers in small samples from the normal distribution. I: Estimation of the mean'. *Technometrics*, **11**, 527–550. (50, 51, 127, 132, 148, 166, 167, 168)
- Guttman, I., and Smith, D. E. (1971). 'Investigation of rules for dealing with outliers in small samples from the normal distribution II: Estimation of the variance'. *Technometrics*, **13**, 101–111. (50, 127, 132, 160, 166, 169, 170)
- Guttman, I., and Tiao, G. C. (1978). 'Effect of correlation on the estimation of a mean in the presence of spurious observations'. To appear. *Technometrics* (268)
- Halperin, M., Greenhouse, S. W., Cornfield, J., and Zalokar, J. (1955). 'Tables of percentage points for the studentized maximum absolute deviate in normal samples'. *J. Amer. Statist. Ass.*, **50**, 185–195. (39, 105, 112)
- Hampel, F. R. (1968). *Contributions to the Theory of Robust Estimation*. Ph.D. dissertation, University of California–Berkeley, University Microfilms Inc., Ann Arbor Mich. (136)
- Hampel, F. R. (1971). 'A generalized qualitative definition of robustness'. *Ann. Math. Statist.*, **42**, 1887–1896. (136, 141, 157)
- Hampel, F. R. (1974). 'The influence curve and its role in robust estimation'. *J. Amer. Statist. Ass.*, **69**, 383–393. (46, 136, 140, 147, 151, 157, 158, 165, 169, 228)
- Harris, T. E., and Tukey, J. W. (1949). 'Measures of location and scale which are relatively insensitive to contamination'. *Memorandum Report No. 31*, Statistical Research Group, Princeton University, Princeton, N.J. (Ch. 4)
- Harter, H. L. (1969a). *Order Statistics and their Use in Testing and Estimation, Vol. 1: Tests Based on Range and Studentized Range of Samples from a Normal*

- Population*. U.S. Air Force, Aerospace Research Laboratories, Washington, D.C. (107, 114)
- Harter, H. L. (1969b). *Order Statistics and their Use in Testing and Estimation, Vol. 2: Estimates Based on Order Statistics of Samples from Various Populations*. U.S. Air Force, Aerospace Research Laboratories, Washington, D.C. (115)
- Harter, H. L. (1974–1976). ‘The method of least squares and some alternatives Parts I–VI’. *Rev. Int. Inst. de Statist.*, **42**, 147–174, (Part I); **42**, 235–264, (Part II); **43**, 1–44, (Part III); **43**, 125–190, (Part IV); **43**, 269–278, (Part V); **44**, 113–159, (Part VI). (21, 22)
- Hartigan, J. A. (1968). ‘Note on discordant observations’. *J. Roy. Statist. Soc. B*, **30**, 545–550. (Ch. 3)
- Hartley, H. O. (1950). ‘The maximum *F*-ratio as a short-cut test for heterogeneity of variance’. *Biometrika*, **37**, 308–312. (84)
- Hawkins, D. M. (1969). ‘On the distribution and power of a test for a single outlier’. *South Afr. Statist. J.*, **3**, 9–15. (Ch. 3)
- Hawkins, D. M. (1973). ‘Repeated testing for outliers’. *Statistica Neerlandica*, **27**, 1–10. (71, 73)
- Hawkins, D. M. (1974). ‘The detection of errors in multivariate data using principal components’. *J. Amer. Statist. Ass.*, **69**, 340–344. (224)
- Healy, M. J. R. (1968). ‘Multivariate normal plotting’. *Appl. Statist.*, **17**, 157–161. (212, 226)
- Henry, F. M. (1950). ‘The loss of precision from discarding discrepant data’. *Res. Qty. Amer. Ass. Health*, **21**, 145–152. (Ch. 3)
- Hinich, M. J., and Talwar, P. P. (1975). ‘A simple method for robust estimation’. *J. Amer. Statist. Ass.*, **70**, 113–119. (256)
- Hodges, J. L. Jr. (1967). ‘Efficiency in normal samples and tolerance of extreme values for some estimates of location’. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California, Berkeley, Calif. (141, 154, 161, 165)
- Hodges, J. L. Jr., and Lehmann, E. L. (1963). ‘Estimates of location based on rank tests’. *Ann. Math. Statist.*, **34**, 598–611. (49, 154, 161)
- Hoening, J., and Crotty, I. M. (1958). *International J. Social Psychiatry*, **3**, 260–277. (55)
- Hogg, R. V. (1967). ‘Some observations on robust estimation’. *J. Amer. Statist. Ass.*, **62**, 1179–1186. (135, 154, 155, 157, 166)
- Hogg, R. V. (1974). ‘Adaptive robust procedures: a partial review and some suggestions for future applications and theory (with comments)’. *J. Amer. Statist. Ass.*, **69**, 909–927. (46, 148, 149, 150, 151, 157)
- Hogg, R. V., Uthoff, V. A., Randles, R. J., and Davenport, A. S. (1972). ‘On the selection of the underlying distribution and adaptive estimation’. *J. Amer. Statist. Ass.*, **67**, 597–600. (Ch. 4)
- Huber, P. J. (1964). ‘Robust estimation of a location parameter’. *Ann. Math. Statist.*, **35**, 73–101. (46, 134, 142, 150, 151, 152, 156, 160, 163, 166, 169, 170)
- Huber, P. J. (1967). ‘The behaviour of maximum likelihood estimates under nonstandard conditions’. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.*, Vol. I, pp. 221–233. (149)
- Huber, P. J. (1968). ‘Robust estimation’. *Mathematical Centre Tracts, Selected Statistical Papers*, **27**, 3–25. Mathematisch Centrum Amsterdam. (142, 162)
- Huber, P. J. (1970). ‘Studentizing robust estimates’. In Puri (1970). (142, 160, 161, 162)
- Huber, P. J. (1972). ‘Robust statistics: a review (The 1972 Wald Lecture)’ *Ann. Math. Statist.*, **43**, 1041–1067. (26, 46, 48, 126, 141, 149, 154, 156, 161, 162, 165, 266)

- Huber, P. J. (1973). 'Robust regression: Asymptotics, conjectures and Monte Carlo'. *Ann. Statist.*, **1**, 799–821. (Ch. 7)
- Irwin, J. O. (1925). 'On a criterion for the rejection of outlying observations'. *Biometrika*, **17**, 238–250. (21, 38, 114, 115)
- Jaeckel, L. A. (1969). *Robust Estimates of Location*. Ph.D. dissertation, University of California-Berkeley, University Microfilms Inc., Ann Arbor, Mich. (166)
- Jaeckel, L. A. (1971a). 'Robust estimates of location: Symmetry and asymmetric contamination'. *Ann. Math. Statist.*, **42**, 1020–1034. (46, 49, 135, 151, 153, 154, 156)
- Jaeckel, L. A. (1971b). 'Some flexible estimates of location'. *Ann. Math. Statist.*, **42**, 1540–1552. (135, 148, 157)
- Jeffreys, H. (1932). 'An alternative to the rejection of observations'. *Proc. Roy. Soc. London, A*, **137**, 78–87. (47)
- Jeffreys, H. (1938). 'The law of error and the combination of observations'. *Phil. Trans. Roy. Soc. London, A*, **237**, 231–271. (Ch. 2)
- Jevons, W. S. (1874). *The Principles of Science*. Macmillan, London, (latest edn. 1958). (Ch. 2, H)
- John, J. A. (1978). 'Outliers in factorial experiments'. *Applied Statistics*, **27**, (246, 251)
- John, J. A., and Draper, N. R. (1978). 'On testing for two or one outliers in two-way tables'. *Technometrics*, **20**, (251, 261)
- John, J. A., and Prescott, P. (1975). 'Critical values of a test to detect outliers in factorial experiments'. *Appl. Statistics*, **24**, 56–59. (241, 244)
- Johns, M. V. Jr. (1974). 'Nonparametric estimates of location'. *J. Amer. Statist. Ass.*, **69**, 453–460. (49)
- Joshi, P. C. (1972a). 'Some slippage tests of mean for a single outlier in linear regression'. *Biometrika*, **59**, 109–120. (262)
- Joshi, P. C. (1972b). 'Efficient estimation of a mean of an exponential distribution when an outlier is present'. *Technometrics*, **14**, 137–144. (36, 77, 171)
- Joshi, P. C. (1975). 'Some distribution theory results for a regression model'. *Ann. Inst. Statist. Math., Tokyo*, **27**, 309–317. (Ch. 7)
- Kabe, D. G. (1970). 'Testing outliers from an exponential population'. *Metrika*, **15**, 15–18. (77, 81, 82, 85, 86, 87)
- Kabe, D. G., and Gupta, R. P. (Eds.) (1973). *Multivariate Statistical Inference*. North-Holland, Amsterdam.
- Kale, B. K. (1974a). 'Detection of outliers'. *Technical Report No. 63*, Department of Statistics, University of Winnipeg, Canada. (77)
- Kale, B. K. (1974b). 'Detection of outliers—a semi-Bayesian approach (preliminary report) (abstract)'. *Inst. Math. Statist. Bull.*, **3**, 153. (272)
- Kale, B. K. (1975a). 'A note on outlier-resistant families and mixtures of distributions'. *Technical Report No. 66*, Department of Statistics, University of Manitoba, Winnipeg, Canada. (38)
- Kale, B. K. (1975b). 'On outlier-proneness of some families of distributions'. *Technical Report No. 68*, Department of Statistics, University of Manitoba, Winnipeg, Canada. (38)
- Kale, B. K. (1975c). 'Trimmed means and the method of maximum likelihood when spurious observations are present'. In Gupta (1975). (50, 51, 77, 172)
- Kale, B. K., and Sinha, S. K. (1971). 'Estimation of expected life in the presence of an outlier observation'. *Technometrics*, **13**, 755–759. (35, 50, 77, 155, 171, 173, 272, 281)
- Kapur, M. N. (1957). 'A property of the optimum solution suggested by Paulson for the  $k$ -sample slippage problem for the normal distribution'. *Ind. Soc. Agric. Statist.*, **9**, 179–190. (43, 193)

- Karlin, S., and Truax, D. R. (1960). 'Slippage Problems'. *Ann. Math. Statist.*, **31**, 296–324. (181, 187, 193, 195, 200, 205)
- Kelleher, G. J. (1974). 'Exact two-sample exceedance tests when one observation possibly an outlier'. *Sankhyā, B*, **36**, 187–193. (Ch. 3)
- King, E. P. (1953). 'On some procedures for the rejection of suspected data'. *J. Amer. Statist. Ass.*, **48**, 531–533. (98)
- Kraft, C. H., and van Eeden, C. (1970). 'Efficient linearized estimates based on ranks'. In Puri (1970). (Ch. 4)
- Kraft, C. H., and van Eeden, C. (1972). 'Asymptotic efficiencies of quick methods of computing efficient estimates'. *J. Amer. Statist. Ass.*, **67**, 199–202. (Ch. 4)
- Krishnaiah, P. R. (Ed.) (1969). *Multivariate Analysis*, Vol. II. Academic Press, New York.
- Kruskal, W. H. (1960a). 'Some remarks on wild observations'. *Technometrics*, **2**, 1–3. (Ch. 2)
- Kruskal, W. H. (1960b). 'Discussion of the papers of Messrs. Anscombe and Daniel'. *Technometrics*, **2**, 157–158. (25, 37)
- Kruskal, W. H., and Wallis, W. A. (1952). 'Use of ranks in one-criterion variance analysis'. *J. Amer. Statist. Ass.*, **47**, 583–612. (187)
- Kudo, A. (1956a). 'On the testing of outlying observations'. *Sankhyā*, **17**, 67–76. (43, 94, 95, 106, 111, 112, 194, 195)
- Kudo, A. (1956b). 'On the invariant multiple decision procedures'. *Bull. Math. Statist.*, **6**, 57–68. (43, 193)
- Kudo, A. (1957). 'The extreme value in a multiple normal sample'. *Mem. Fac. Sci. Kyushu Univ.*, A, **11**, 143–156. (218)
- Larson, W. A., and McCleary, S. J. (1972). 'The use of partial residuals in regression analysis'. *Technometrics*, **14**, 781–790. (Ch. 7)
- Laurent, A. G. (1963). 'Conditional distribution of order statistics and distribution of the reduced  $i$ th order statistic of the exponential model'. *Ann. Math. Statist.*, **34**, 652–657. (77)
- Legendre, A. M. (1805). *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. Courcier, Paris (especially 'Appendice sur la méthode des moindres quarrés', pp. 72–80). (19)
- Legendre, A. M. (1814). 'Méthode des moindres quarrés, pour trouver le milieu le plus probable entre les résultats de différentes observations'. *Mémoires de la Classe des Sciences Mathématiques et Physiques de l'Institut de France*, ANNÉE 1810, 149–154. (Ch. 2, H)
- Lehmann, E. L. (1953). 'The power of rank tests'. *Ann. Math. Statist.*, **24**, 23–43. (182)
- Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods based on Ranks*. McGraw-Hill, New York. (162)
- Leone, F. C., and Moussa-Hamouda, E. (1973). 'Relative efficiencies of 'O-BLUE' estimators in simple linear regression'. *J. Amer. Statist. Ass.*, **68**, 953–959. (Ch. 4, Ch. 7)
- Leone, F. C., Jayachandran, T., and Eisenstat, S. (1967). 'A study of robust estimators'. *Technometrics*, **9**, 652–660. (151, 161, 168)
- Lewis, T., and Fieller, N. R. J. (1978). 'A recursive algorithm for null distributions for outliers: I. Gamma samples. To appear, *Technometrics*, **20**, (40, 73, 77, 82, 85, 108, 110)
- Liberman, G. J., and Owen, D. B. (1961). *Tables of the Hypergeometric Probability Distribution*. University Press, Stanford, California. (204)
- Lieblein, J. (1952). 'Properties of certain statistics involving the closest pair in a sample of three observations'. *J. Res. Nat. Bur. Stands.*, **48**, 225–268. (51)
- Lieblein, J. (1962). 'The closest two out of three observations'. In Sarhan and Greenberg (1962). (51)

- Likeš, J. (1966). 'Distribution of Dixon's statistics in the case of an exponential population'. *Metrika*, **11**, 46–54. (40, 54, 73, 77, 80, 81, 82, 83, 86, 87)
- Lingappaiah, G. S. (1976). 'Effect of outliers in the estimation of parameters'. *Metrika*, **23**, 27–30. (281)
- Lund, R. E. (1975). 'Tables for an approximate test for outliers in linear models'. *Technometrics*, **17**, 473–476. (255, 262)
- McCarthy, P. J. (1972). 'The effects of discarding inliers when binomial data are subject to classification errors'. *J. Amer. Statist. Ass.*, **67**, 515–529. (Ch. 3)
- McKay, A. T. (1935). 'The distribution of the difference between the extreme observation and the sample mean in samples of  $n$  from a normal universe'. *Biometrika*, **27**, 466–471. (111)
- McMillan, R. G. (1971). 'Tests for one or two outliers in normal samples with unknown variance'. *Technometrics*, **13**, 87–100. (34, 40, 44, 71, 73, 94, 95, 96, 104, 105, 106)
- McMillan, R. G., and David, H. A. (1971). 'Tests for one of two outliers in normal samples with known variance'. *Technometrics*, **13**, 75–85. (34, 40, 71, 73, 110, 111, 112)
- Maguire, B. A., Pearson, E. S., and Wynn, A. H. A. (1952). 'The time intervals between industrial accidents'. *Biometrika*, **39**, 168–180. (119)
- Maire, C.; Boscovich, R. J. (1755). *De litteraria Expeditione per Pontificiam ditionem ad dimetiendas duas Meridiani gradus, et corrigendam mappam geographicam, jussu, et auspiciis Benedicti XIV Pont. Max. Suscepit*. Ramae. (French translation: *Voyage Astronomique et Géographique dans l'État de l'Église, entrepis par l'Ordre et sous les Auspices du Pope Benoit XIV, pour mesurer deux degrés du méridien, et corriger la Carte de l'État ecclésiastique*. Paris, 1, 770. (18))
- Martin, R. D., Mareliez, C. J., and Goodfellow, D. M. (1973). 'Robust location estimates and confidence intervals via stochastic approximation: small sample behaviour'. *Inst. Math. Statist. Bull.*, **2**, 138. (Ch. 4)
- Mendeleev, D. I. (1895). 'Course of work on the renewal of prototypes or standard measures of lengths and weights' (Russian). *Vremennik Glavnoi Palaty Mer i Vesov*, **2**, 157–185. (Reprinted 1950; *Collected Writings (Socheneniya)*, **22**, 175–213, izdat. Akad. Nauk, SSSR, Leningrad–Moscow.) (21, 47)
- Mercer, W. B., and Hall, A. D. (1912). 'The experimental error of field trials'. *J. Agric. Sci.*, **4**, 107–132. (2)
- Merriman, M. (1877). 'List of writings relating to the method of least squares with historical and critical notes'. *Transactions of the Connecticut Academy of Arts and Sciences*, **4**, 151–232. (Ch. 2, H)
- Merriman, M. (1884). *A Textbook on the Method of Least Squares*. Wiley, New York. (Ch. 2, H)
- Meshalkin, L. D., Smirnov, N. P., and Sosnovskii, N. N. (1969). 'On the stability of estimates of the distribution center' (Review) (Russian). *Zavodskaya Laboratoriya*, **35**, 51–61. English translation: *Industrial Laboratory*, **35**, 712–716. (Ch. 4)
- Mickey, M. R. (1974). 'Detecting outliers with stepwise regression'. *Communications—UCLA Health Sciences Facility*, **1**, 1. (265)
- Mickey, M. R., Dunn, O. J., and Clark, V. (1967). 'Note on use of stepwise regression in detecting outliers'. *Computers & Biomed. Res.*, **1**, 105–111. (265)
- Miké, V. (1971). 'Efficiency-robust systematic linear estimators of location'. *J. Amer. Statist. Ass.*, **66**, 594–601. (Ch. 4)
- Miké, V. (1973). 'Robust Pitman-type estimators of location'. *Ann. Inst. Statist. Math., Tokyo*, **25**, 65–86. (Ch. 4)
- Moore, P. G. (1957). 'The two-sample  $t$ -test based on range'. *Biometrika*, **44**, 482–485. (107)
- Moran, M. A., and McMillan, R. G. (1973). 'Tests for one or two outliers in normal

- samples with unknown variance: a correction'. *Technometrics*, **15**, 637–640. (41, 94, 104, 105)
- Moshman, J. (1952). 'Testing a straggler mean in a 2-way classification using the range'. *Ann. Math. Statist.*, **23**, 126–132. (Ch. 7)
- Mosteller, F. (1948). 'A  $k$ -sample slippage test for an extreme population'. *Ann. Math. Statist.*, **19**, 58–65. (176, 283)
- Mosteller, F., and Tukey, J. W. (1950). 'Significance levels for a  $k$ -sample slippage test'. *Ann. Math. Statist.*, **21**, 120–123. (177)
- Mount, K. S., and Kale, B. K. (1973). 'On selecting a spurious observation'. *Can. Math. Bull.*, **16**, 75–78. (37, 77)
- Moussa-Hamouda, E., and Leone, F. C. (1974). 'The O-BLUE estimators for complete and censored samples in linear regression'. *Technometrics*, **16**, 441–446. (Ch. 4, Ch. 7)
- Mudrov, V. I., Kushko, V. L., Mikhailov, V. I., and Osovitskii, E. M. (1968). 'Some experiments on the use of the least-moduli method in processing orbital data' (Russian). *Kosmicheskie Issledovaniya*, **6**, 502–504. English translation: *Cosmic Research*, **6**, 421–431. (Ch. 2)
- Muncke, G. W. (1825). 'Beobachtung'. Gehler's *Physikalisches Wörterbuch*, 2nd edn. Leipzig, Vol. I, pp. 884–912. (Ch. 2, H)
- Murphy, R. B. (1951). *On Tests for Outlying Observations*. Ph.D. thesis, Princeton University, University Microfilms Inc., Ann Arbor, Mich. (40, 44, 71, 95)
- Naik, U. D. (1972). 'A Bayesian analysis of certain contaminated samples'. *Research Report No. 104*, Department of Probability and Statistics, Sheffield University. (206)
- Nair, K. R. (1948). 'The distribution of the extreme deviate from the sample mean and its studentized from'. *Biometrika*, **35**, 118–144. (104, 110, 111)
- Nair, K. R. (1952). 'Tables of percentage points of the "Studentized" extreme deviate from the sample mean'. *Biometrika*, **39**, 189–191. (104)
- Neave, H. R. (1972). 'Some quick tests for slippage'. *The Statistician*, **21**, 197–208. (178, 180)
- Neave, H. R. (1973). 'A power study of some tests for slippage'. *The Statistician*, **22**, 269–280. (180)
- Neave, H. R. (1975). 'A quick and simple technique for general slippage problems'. *J. Amer. Statist. Ass.*, **70**, 721–726. (183, 185)
- Newcomb, S. (1886). 'A generalized theory of the combination of observations so as to obtain the best result'. *Amer. J. Math.*, **8**, 343–366. (21, 47)
- Newcomb, S. (1912). 'Researches on the motion of the moon, Part II. The mean motion of the moon and other astronomical elements derived from observations of eclipses and occultations extending from the period of the Babylonians until A.D. 1908'. *Astronomical papers*, **9**, 1–249, U.S. Government Printing Office, Washington. (Ch. 2, H)
- Neyman, J., and Scott, E. L. (1971). 'Outlier proneness of phenomena and of related distribution. In Rustagi (1971). (37)
- Noether, G. E. (1967). 'Wilcoxon confidence intervals for location parameters in the discrete case'. *J. Amer. Statist. Ass.*, **62**, 184–188. (162)
- Noether, G. E. (1973). 'Some simple distribution-free confidence intervals for the center of a symmetric distribution'. *J. Amer. Statist. Ass.*, **68**, 716–719. (162)
- Noether, G. E. (1974). 'Distribution-free confidence intervals based on linear rank statistics'. In Williams (1974) (162)
- Odeh, R. E. (1967). 'The distribution of the maximum sum of ranks'. *Technometrics*, **9**, 271–278. (181, 182)
- Ogrodnikoff, K. (1928). 'On the occurrence of discordant observations and a new method of treating them'. *Monthly Notices Roy. Astr. Soc.*, **88**, 523–532. (Ch. 3, H)

- Olkin, I. (Ed.) (1960). *Contributions to Probability and Statistics*. University Press, Stanford, Calif.
- Owen, D. B. (1962). *Handbook of Statistical Tables*. Addison-Wesley, Reading, Mass. (204)
- Paulson, E. (1952a). 'On the comparison of several experimental categories with a control'. *Ann. Math. Statist.*, **23**, 239–246. (Ch. 5)
- Paulson, E. (1925b). 'A optimum solution to the  $k$ -sample slippage problem for the normal distribution'. *Ann. Math. Statist.*, **23**, 610–616. (192, 200, 240)
- Paulson, E. (1961). 'A non-parametric solution for the  $k$ -sample slippage problem'. In Solomon (1961). (207)
- Paulson, E. (1962). 'A sequential procedure for comparing several experimental categories with a standard or control'. *Ann. Math. Statist.*, **33**, 438–443. (207)
- Pearson, E. S. (1926). 'A further note on the distribution of range in samples taken from a normal population'. *Biometrika*, **18**, 173–194. (114)
- Pearson, E. S. (1932). 'The percentage limits for the distribution of range in samples from a normal population ( $n \leq 100$ )'. *Biometrika*, **24**, 404–417. (114)
- Pearson, E. S., and Chandra Sekar, C. (1936). 'The efficiency of statistical tools and a criterion for the rejection of outlying observations'. *Biometrika*, **28**, 308–320. (22, 40, 54, 71, 73, 94, 243)
- Pearson, E. S., and Hartley, H. O. (1942). 'The probability integral of the range in samples of  $n$  observations from a normal population'. *Biometrika*, **32**, 301–310. (114)
- Pearson, E. S., and Hartley, H. O. (Eds.) (1966). *Biometrika Tables for Statisticians*. Vol. 1, 3rd edn., Cambridge University Press, London. (84, 95, 97, 101, 014, 107, 110, 111, 193, 196, 197, 202)
- Pearson, E. S., and Stephens, M. A. (1964). 'The ratio of range to standard deviation in the same normal sample'. *Biometrika*, **51**, 484–487. (39, 97)
- Pearson, K. (Ed.) (1931). *Tables for Statisticians and Biometricalians*. Biometric Lab., University College, London. (25)
- Pearson, K. (Ed.) (1968). *Tables of the Incomplete Beta-Function*. 2nd edn. (with new Introduction by Pearson, E. S., and Johnson, N. L.). Cambridge University Press, London. (202)
- Peirce, B. (1852). 'Criterion for the rejection of doubtful observations'. *Astr. J.*, **2**, 161–163. (19)
- Peirce, B. (1878). 'On Peirce's criterion' (with remarks by Scott, C. A.). *Proceedings of the American Academy of Arts and Sciences*, **13**, 348–351. (Ch. 2, H)
- Peirce, C. S. (1873). 'On the theory of errors of observations'. *Report of the Superintendent of the United States Coast Survey*, (for the year ending 1 November 1870) U.S. Government Printing Office, Washington. (Ch. 2, H)
- Pfanzagl, J. (1959). 'Ein kombiniertes Test und Klassifikations-Problem'. *Metrika*, **2**, 11–45. (195–197)
- Pillai, K. C. S. (1959). 'Upper percentage points of the extreme studentized deviate from the sample mean'. *Biometrika*, **46**, 473–474. (Ch. 3)
- Pillai, K. C. S., and Tienzo, B. P. (1959). 'On the distribution of the extreme studentized deviate from the sample mean'. *Biometrika*, **46**, 467–472. (Ch. 3)
- Prescott, P. (1975a). 'An approximate test for outliers in linear regression'. *Technometrics*, **17**, 127–128. (261)
- Prescott, P. (1975b). 'An approximate test for outliers in linear models'. *Technometrics*, **17**, 129–132. (255, 261)
- Prescott, P. (1976). 'On a test for normality based on sample entropy'. *J. Roy. Statist. Soc. B*, **38**, 254–256. (249)
- Proschan, F. (1953). 'Rejection of outlying observations'. *Amer. J. Phys.*, **21**, 520–525. (Ch. 3)

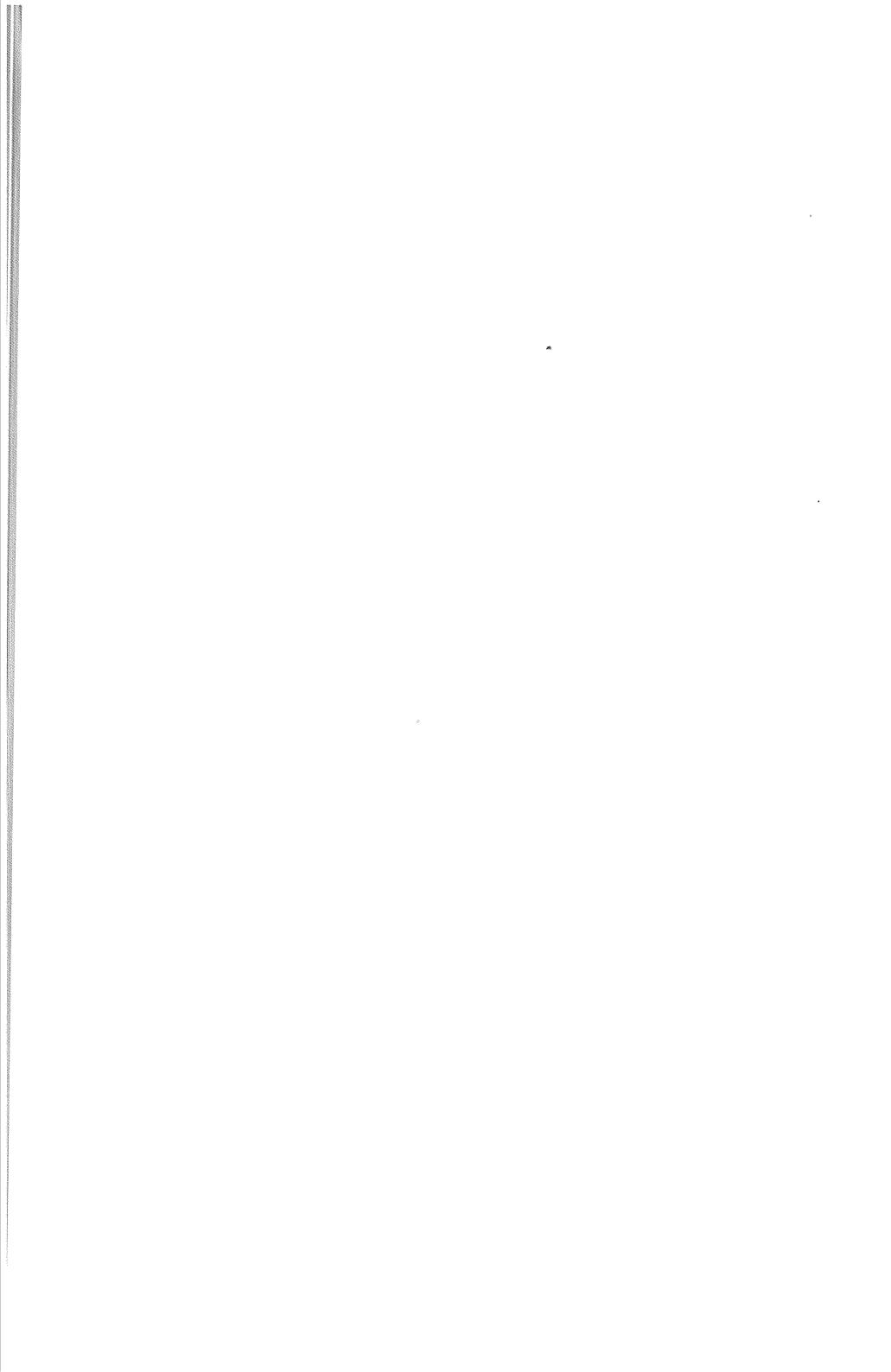
- Proschan, F. (1975a). 'Testing suspected observations'. *Sankhyā A*, **17**, 67–76. (Ch. 3)
- Proschan, F. (1957b). 'Testing suspected observations'. *Ind. Qual. C.XIII*, 14–19. (Ch. 3)
- Puri, M. L. (Ed.) (1970). *Nonparametric Techniques in Statistical Inference*. Cambridge University Press, London.
- Quenouille, M. H. (1953). *The Design and Analysis of Experiment*. Griffin, London. (11, 241)
- Quenouille, M. H. (1956). 'Notes on bias in estimation'. *Biometrika*, **43**, 353–360. (48)
- Quesenberry, C. P., and David, H. A. (1961). 'Some tests for outliers'. *Biometrika*, **48**, 379–387. (94, 95, 104, 105, 106, 193, 195, 200)
- Rahman, N. A. (1972). *Practical Exercises in Probability and Statistics*. Griffin, London. (11)
- Ramachandran, K. V., and Khatri, C. G. (1957). 'On a decision procedure based on the Tukey statistic'. *Ann. Math. Statist.*, **28**, 802–806. (43, 205)
- Randles, R. H., Ramberg, J. S., and Hogg, R. V. (1973). 'An adaptive procedure for selecting the population with the largest location parameter'. *Technometrics*, **15**, 769–778. (Ch. 5)
- Rao, C. R. (1964). 'The use and interpretation of principal component analysis in applied research'. *Sankhyā A*, **26**, 329–358. (226)
- Rao, P. V. (1972). 'Robust estimation for a simple exponential model'. *Australian J. Statist.*, **14**, 54–62. (Ch. 4)
- Rao, P. V., and Thornby, J. I. (1969). 'A robust point estimator in a generalized regression model'. *Ann. Math. Statist.*, **40**, 1784–1790. (Ch. 7)
- Rider, P. R. (1933). 'Criteria for rejection of observations'. *Washington University Studies—New Series, Science and Technology*, **8**, 3–23. (20)
- Rohlf, F. J. (1975). 'Generalisation of the gap test for the detection of multivariate outliers'. *Biometrics*, **31**, 93–101. (221, 229)
- Rosner, B. (1975). 'On the detection of many outliers'. *Technometrics*, **17**, 221–227. (Ch. 3)
- Rustagi, J. (Ed.) (1971). *Optimising Methods in Statistics*. Academic Press, New York.
- Sacks, J., and Ylvisaker, D. (1972). 'A note on Huber's robust estimation of a location parameter'. *Ann. Math. Statist.*, **43**, 1068–1075. (Ch. 4)
- Samuelson, P. A. (1968). 'How deviant can you be?'. *J. Amer. Statist. Ass.*, **63**, 1522–1525. (Ch. 2)
- Sarhan, A. E., and Greenberg, B. G. (Eds.) (1962). *Contributions to Order Statistics*. Wiley, New York. (159)
- Saunders, S. A. (1903). 'Note on the use of Peirce's criterion for the rejection of doubtful observations'. *Monthly Notices Roy. Astr. Soc.*, **63**, 432–436. (19)
- Scholz, F. (1974). 'A comparison of efficient location estimators'. *Ann. Statist.*, **2**, 1323–1326. (Ch. 4)
- Schuster, E. F., and Narvarte, J. A. (1973). 'A new nonparametric estimation of the center of a symmetric distribution'. *Ann. Statist.*, **1**, 1096–1104. (Ch. 4)
- Schweder, T. (1973). 'Window estimation of the asymptotic variance of the Hodges–Lehmann estimator'. *Inst. Math. Statist. Bull.*, **2**, 92. (Ch. 4)
- Schweder, T. (1976). 'Some "optimal" methods to detect structural shift or outliers in regression'. *J. Amer. Statist. Ass.*, **71**, 491–501. (256)
- Searls, D. T. (1966). 'An estimator for a population mean which reduces the effect of large true observations'. *J. Amer. Statist. Ass.*, **61**, 1200–1204. (Ch. 4)
- Sen, P. K. (1968). 'On a further robustness property of the test and estimator based on Wilcoxon's signed rank statistic'. *Ann. Math. Statist.*, **39**, 282–285. (Ch. 4)

- Sen, P. K., and Puri, M. L. (1969). 'On robust nonparametric estimation in some multivariate linear models'. In Krishnaiah (1969). (Ch. 7)
- Seth, G. R. (1950). 'On the distribution of the two closest among a set of three observations'. *Ann. Math. Statist.*, **21**, 298–301. (51)
- Shapiro, S. S., and Wilk, M. B. (1965). 'An analysis of variance test for normality (complete samples)'. *Biometrika*, **52**, 591–611. (31, 40, 88, 103, 249)
- Shapiro, S. S., and Wilk, M. B. (1972). 'An analysis of variance test for the exponential distribution (complete samples)'. *Technometrics*, **14**, 355–370. (31, 40)
- Shapiro, S. S., Wilk, M. B., and Chen, M. J. (1968). 'A comparative study of various tests for normality'. *J. Amer. Statist. Ass.*, **63**, 1343–1372. (31, 40, 97, 101, 103)
- Sheynin, O. B. (1966a). 'Origin of the theory of errors'. *Nature*, **211**, 1003–1004. (Ch. 2, H)
- Sheynin, O. B. (1966b). 'On selection and adjustment of direct observations' (Russian). *Izvestiya Vysshikh Uchebnykh Zavedenii. Geodeziia i Aerofotos'ema*, **1966**. English translation: *Geodesy and Aero-photography*, **1966** (1967), 114–117. (Ch. 2, H)
- Sheynin, O. B. (1971). 'J. H. Lambert's work on probability'. *Archive for History of Exact Sciences*, **7**, 244–256. (Ch. 2, H)
- Shorack, G. R. (1974). 'Random means'. *Ann. Statist.*, **2**, 661–675. (Ch. 4)
- Siddiqui, M. M., and Raghunandanan, K. (1967). 'Asymptotically robust estimators of location'. *J. Amer. Statist. Ass.*, **62**, 950–953. (135, 157, 164)
- Sinha, S. K. (1972). 'Reliability estimation in life testing in the presence of an outlier observation'. *Op. Res.*, **20**, 888–894. (51, 77, 155, 272, 279, 281)
- Sinha, S. K. (1973a). 'Distributions of order statistics and estimation of mean life when an outlier may be present'. *Canad. J. Statist.*, **1**, 119–121. (50, 51, 77, 172)
- Sinha, S. K. (1973b). 'Lifetesting and reliability estimation for non-homogeneous data—a Bayesian approach'. *Comm. Statist.*, **2**, 235–243. (50, 51, 77, 155, 272, 279, 281)
- Sinha, S. K. (1973c). 'Estimation of the parameters of a two-parameter exponential distribution when an outlier may be present'. *Utilitas Mathematica*, **3**, 75–82. Correction (1974), *Utilitas Mathematica*, **4**, 333–334. (51, 77, 172)
- Sinha, S. K. (1973d). 'Some distributions relevant in life testing when an outlier may be present'. *Technical Report No. 42*, Department of Statistics, University of Manitoba, Winnipeg, Canada. (172)
- Siotani, M. (1959). 'The extreme value of the generalised distances of the individual points in the multivariate normal sample'. *Ann. Inst. Statist. Math. Tokyo*, **10**, 183–208. (211, 219)
- Snedecor, G. W., and Cochran, W. G. (1967). *Statistical Methods*. 6th edn. The Iowa State University Press, Ames, Iowa. (265)
- Solomon, H. (Ed.) (1961). *Studies in Item Analysis and Prediction*. University Press, Stanford Calif.
- Srikantan, K. S. (1961). 'Testing for the single outlier in a regression model'. *Sankhyā, A*, **23**, 251–260. (254, 257, 259, 266)
- Srivastava, M. S. (1973). 'The performance of a sequential procedure for a slippage problem'. *J. Roy. Statist. Soc. B*, **35**, 97–103. (206)
- Stampfer, S. (1839). 'Ueber das Verhältniss der Wiener Klafter zum Meter'. *Jahrbücher des K. K. Polytechnisches Institutes (Vienna)*, **20**, 145–176. (Ch. 2, H)
- Stefansky, W. (1971). 'Rejecting outliers by maximum normal residual'. *Ann. Math. Statist.*, **42**, 35–45. (94, 242, 254, 255)
- Stefansky, W. (1972). 'Rejecting outliers in factorial designs'. *Technometrics*, **14**, 469–479. (242, 254, 255)
- Stewart, R. M. (1920a). 'Pierce's criterion'. *Popular Astronomy*, **28**, 2–3. (Ch. 2, H)
- Stewart, R. M. (1920b). 'The treatment of discordant observations'. *Popular Astronomy*, **28**, 4–6. (Ch. 2, H)

- Stigler, S. M. (1973a). 'The asymptotic distribution of the trimmed mean'. *Ann. Statist.*, **1**, 472–477. (Ch. 4)
- Stigler, S. M. (1973b). 'Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920'. *J. Amer. Statist. Ass.*, **68**, 872–879. (21, 158)
- Stigler, S. M. (1974). 'Linear functions of order statistics with smooth weight functions'. *Ann. Statist.*, **2**, 676–693. (Ch. 4)
- Stone, E. J. (1868). 'On the rejection of discordant observations'. *Monthly Notices Roy. Astr. Soc.*, **28**, 165–168. (19)
- Stone, E. J. (1873). 'On the rejection of discordant observations'. *Monthly Notices Roy. Astr. Soc.*, **34**, 9–15. (20)
- Stone, E. J. (1874). 'Note on a discussion relating to the rejection of discordant observations'. *Monthly Notices Roy. Astr. Soc.*, **35**, 107–108. (Ch. 2, H)
- Student (1927). 'Errors of routine analysis'. *Biometrika*, **19**, 151–164. (47)
- Sukhov, A. N. (1971). 'Comparison of the median and the arithmetic mean in the case of a small sample' (Russian). *Izvestiya Vysshikh Uchebnykh Zavedenii. Geodeziia i Aerofotos'emka*, **1971**, 59–65. English translation: *Geodesy and Aerophotography*, **1971**, (1973), 326–329. (Ch. 4)
- Swaroop, R., and Winter, W. R. (1971). 'A statistical technique for computer identification of outliers in multivariate data'. *NASA TN D-6472*. National Aeronautics and Space Administration, Washington, D.C. (Ch. 7)
- Swaroop, R., West, K. A., and Lewis, C. E. Jr. (1969). 'A simple technique for automatic computer editing of biodata'. *NASA TN D-5275*. National Aeronautics and Space Administration, Washington, D.C. (Ch. 3)
- Takeuchi, K. (1971). 'A uniformly asymptotically efficient estimator of a location parameter'. *J. Amer. Statist. Ass.*, **66**, 292–301. (49, 142, 153, 166)
- Thomas, J. (1969). 'Monte Carlo investigation of the robustness of Dixon's criteria for testing outlying observations'. *Proceedings of the Fourteenth Conference on the Design of Experiments in Army Research, Development and Testing*, pp. 437–483. (Ch. 3)
- Thompson, G. W. (1955). 'Bounds for the ratio of range to standard deviation'. *Biometrika*, **42**, 268–269. (107)
- Thompson, W. A. Jr., and Willke, T. A. (1963). 'On an extreme rank sum test for outliers'. *Biometrika*, **50**, 375–383. (252)
- Thompson, W. R. (1935). 'On a criterion for the rejection of observations and the distribution of the ratio of the deviation to the sample standard deviation'. *Ann. Math. Statist.*, **6**, 214–219. (22, 254)
- Thompson, W. R. (1942). 'On a criterion of the difference between the extreme observation and the sample mean in samples of  $n$  from a normal universe'. *Biometrika*, **32**, 301–310. (Ch. 3)
- Tiao, G. C., and Guttman, I. (1967). 'Analysis of outliers with adjusted residuals'. *Technometrics*, **9**, 541–559. (127, 132, 168, 260)
- Tietjen, G. L., and Moore, R. H. (1972). 'Some Grubbs-type statistics for the detection of several outliers'. *Technometrics*, **14**, 583–597. (38, 40, 73, 95, 96, 102)
- Tietjen, G. L., Moore, R. H., and Beckman, R. J. (1973). 'Testing for a single outlier in simple linear regression'. *Technometrics*, **15**, 717–721. (254, 261)
- Tippett, L. H. C. (1925). 'On the extreme individuals and the range of samples taken from a normal population'. *Biometrika*, **17**, 364–387. (114)
- Torgerson, E. N. (1971). 'A counterexample on translation invariant estimators'. *Ann. Math. Statist.*, **42**, 1450–1451. (Ch. 4)
- Truax, D. R. (1953). 'An optimum slippage test for the variances of  $k$  normal distributions'. *Ann. Math. Statist.*, **24**, 669–674. (43, 196, 200)
- Tukey, J. W. (1949). 'The truncated mean in moderately large samples'. *Memorandum*

- dum Report No. 32.* Statistical Research Group, Princeton University, Princeton, N.J. (Reports Nos. 25, 33, and 34 relate.) (Ch. 4)
- Tukey, J. W. (1960). 'A survey of sampling from contaminated distributions'. In Olkin (1960). (33, 46, 127, 130, 158)
- Tukey, J. W. (1962). 'The future of data analysis'. *Ann. Math. Statist.*, **3**, 1–67. (249)
- Tukey, J. W. (1977). *Exploratory Data Analysis*, Vol. 1. Addison-Wesley, Reading, Mass. (153)
- Tukey, J. W., and McLaughlin, D. M. (1963). 'Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization'. *Sankhyā*, A, **25**, 331–352. (48, 142, 161)
- Veale, J. R., and Huntsberger, D. V. (1969). 'Estimation of a mean when one observation may be spurious'. *Technometrics*, **11**, 331–339. (51, 127)
- Veale, J. R., and Kale, B. K. (1972). 'Tests of hypotheses for expected life in the presence of a spurious observation'. *Utilitas Mathematica*, **2**, 9–23. (35, 50, 51, 77, 144, 172)
- Walsh, J. E. (1950). 'Some nonparametric tests of whether the largest observations of a set are too large or too small'. *Ann. Math. Statist.*, **21**, 583–592. Correction (1953), *Ann. Math. Statist.*, **24**, 134–135. (44, 283)
- Walsh, J. E. (1959). 'Large sample non-parametric rejection of outlying observations'. *Ann. Inst. Statist. Math. Tokyo*, **10**, 223–232. (44, 283, 285)
- Walsh, J. E. (1965). *Handbook of Non-parametric Statistics, II*. Van Nostrand, Princeton N.J. (44, 283)
- Walsh, J. E., and Kelleher, G. J. (1973). 'Nonparametric estimation of mean and variance when a few "sample" values possibly outliers'. *Ann. Inst. Statist. Math. Tokyo*, **25**, 87–90. (285)
- Wani, J. K., and Kabe, D. G. (1971). 'Distributions of Dixon's statistics for the truncated exponential, rectangular, and random intervals population'. *Metron*, **29**, 151–160. (124, 125)
- West, S. A. (1975). 'Bias in the estimator of Kendall's rank correlation when extreme pairs are removed from the sample'. *J. Amer. Statist. Ass.*, **70**, 439–442. (Ch. 6)
- Wilk, M. B., and Gnanadesikan, R. (1964). 'Graphical methods for internal comparisons in multiresponse experiments'. *Ann. Math. Statist.*, **35**, 613–631. (222, 226, 230)
- Wilk, M. B., Gnanadesikan, R., and Huyett, M. J. (1962a). 'Probability plots for the gamma distribution'. *Technometrics*, **4**, 1–20. (226, 230)
- Wilk, M. B., Gnanadesikan, R., and Huyett, M. J. (1962b). 'Estimation of parameters of the gamma distribution using order statistics'. *Biometrika*, **49**, 525–545. (226)
- Wilks, S. S. (1962). *Mathematical Statistics*. Wiley, New York. (215)
- Wilks, S. S. (1963). 'Multivariate statistical outliers' *Sankhyā*, A, **25**, 407–426. (215, 226)
- Williams, E. J. (Ed.) (1974). *Studies in Probability and Statistics*. Jerusalem Academic Press, Jerusalem.
- Willke, T. A. (1966). 'A note on contaminated samples of size three'. *Journal of Research of the National Bureau of Standards*, B, **70**, 149–151. (51, 127, 135)
- Winlock, J. (1856). 'On Professor Airy's objections to Peirce's criterion'. *Astr. J.*, **4**, 145–147. (Ch. 2, H)
- Wooding, W. M. (1969). 'The computation and use of residuals in the analysis of experiment data'. *J. Quality Technology*, **1**, 175–188. Correction, **1**, 294 (226)
- Wright, T. W. (1884). *A Treatise on the Adjustment of Observations by the Method of Least Squares*. Van Nostrand, New York. (2, 21)
- Wright, T. W., and Hayford, J. F. (1906). *Adjustment of Observations*. Van Nostrand, New York. (21)

- Yanagawa, T. (1969). 'A small sample robust competitor of Hodges-Lehmann estimate'. *Bull. Math. Statist.* (Fukuoka), **13**, (3–4), 1–14. (Ch. 4)
- Yhap, E. F. (1967). *An Asymptotic Optimally Robust Linear Estimation of Location for Symmetric Shapes*. Doctoral dissertation, New York University, University Microfilms Inc., Ann Arbor, Mich. (Ch. 4)
- Yohai, V. J. (1974). 'Robust estimation in the linear model'. *Ann. Statist.*, **2**, 562–567. (Ch. 7)
- Youden, W. J. (1949). 'The fallacy of the best two out of three'. *National Bureau of Standards Technical Bulletin*, **33**, 77–78. (Ch. 2)
- Youden, W. J. (1953). 'Sets of three measurements'. *The Scientific Monthly*, **77**, 143–147. (Ch. 2)
- von Zach, F. X. (1805). 'Versuch einer auf Erfahrung gegründeten Bestimmung terrestrischer Refractionen'. *Monatliche Correspondenz zur Beförderung der Erd- und Himmels-Kunde*, **11**, 389–415. (Ch. 2, H)
- Zelenen'kiy, V. P. (1969). 'Application of statistical decision theory to the exclusion of anomalous measurements'. *Izvestiya Akademii Nauk SSR, Tekhnicheskaya Kibernetika*, **1969**, (2), 139–142 (Russian). Translated in *Engineering Cybernetics* (**1969**) (2), 122–126. (Ch. 3, Ch. 7)
- Zinger, A. (1961). 'Detection of best and outlying normal distributions with known variances'. *Biometrika*, **48**, 457. (Ch. 3)



# *Index*

- 'Aberrant' observation, 36, 51, 53, 208, 237 (*but see* Outlier)  
'Abnormal' observation, 19 (*but see* Outlier)  
Accommodation of outliers (robust inference), 4, 20, 26, 46–51  
adaptive estimators, 49  
blanket procedures, 47–49  
correlation coefficient, 233  
exponential samples, 171–173  
Hodges–Lehmann estimator, 49, 154  
in confidence intervals, 141–142, 160–162  
in designed experiments, 246–247  
in estimating dispersion, 129, 133–134, 150, 158–160, 169–171  
in estimating location, 47–49, 128–129, 130, 144–158, 163–169  
in general linear model, 246–247, 260  
in regression, 256  
in significance tests, 142–144, 160–162  
in time-series, 268  
influence functions and influence curves, 136–141, 228–229  
linear order statistics estimators ( $L$ -estimators), 48, 152–153, 171–172  
maximin robust estimator, 135, 156, 157  
maximum likelihood type estimators ( $M$ -estimators), 48–49, 149–152, 163–166, 169–170  
median, 46  
minimax robust estimator, 135, 156, 157  
multivariate data, 231–233  
multivariate mean, 231–232  
multivariate normal data, 231–232  
performance criteria (univariate inference), 130–144  
premium-protection rules, 50, 131  
rank test estimators ( $R$ -estimators), 49, 153–154  
robustness of performance, 126  
robustness of validity, 126  
specific procedures, 47, 49–51  
studentized location estimators, 160–162, 168–169  
trimming, 21, 26, 48  
univariate, 126–173  
univariate normal data, 163–171  
using Bayesian methods, 274–275, 277–282  
using non-parametric methods, 285  
variance–covariance matrix, 232–233  
Winsorization, 26, 48, 50  
(*For further details see under specific headings:* Trimming, etc.)  
Adaptive inference procedures, 49, 148–149, 157  
Adjusted residuals, 168, 232, 260  
Andrews' Fourier-type plot, 227, 228  
'Anomalous' observation, 8, 18, 19, 230 (*but see* Outlier)  
'Ballooning', 253  
Basic model, 26, 29, 285  
Bayesian treatment of outliers, 15, 45, 46, 269–282  
accommodation, 45, 274–275, 277–282  
'detection of spuriousity', 275–277  
exchangeable model, 36, 272, 279

- in general linear model, 252
- in multivariate data, 277
- multiple decision approach, 279
- multiple outliers, 272–275
- philosophical considerations, 269–271, 283
- ‘semi-Bayesian’ methods, 272, 280, 281
- slippage model, 33, 36
- slippage tests, 192, 193, 200–201, 206
- ‘tests of discordancy’, 269–277
- Bickel-Hodges estimator (*see* Folded median)
- Binomial distribution**
  - details of discordancy tests for practical use, 75–76, 115–116, 123–124; guide to use of tables, 75–76; worksheets, 123–124; tables, 320–322
  - slippage tests, 203–204; tables, 320–322
- Block procedure for slippage, 184
- Block tests of discordancy** (*see* Multiple outliers)
- Breakdown point, 140–141
- Carelessness (modulus of), 19
- Cauchy distribution**
  - compared with normal distribution, 9
  - L*-estimators, 48
  - outlier proneness, 37
- Chauvenet’s rejection test, 2, 19
- ‘Cloaking’ of outliers by other non-null manifestations in structured data, 238, 247, 249, 265, 287
- Confidence intervals, robust against outliers, 141–142, 160–162
  - error frequency, 142
  - error probability, 142
- Consecutive procedures for slippage, 184
- Consecutive tests of discordancy (*see* Multiple outliers)
- Contaminant, 65, 127
- Contaminated distribution, 127, 255
- Contamination, 47, 127–130
  - asymmetric, 128, 132–133, 135–136
  - effect on estimating normal mean and variance, 128–129
  - symmetric, 128
  - undetectability in normal mixture, 129–130
- Designed experiments (outliers in), 14, 234–252
- accommodation, 246–247
- Bayesian method, example, 274–275
- detected by two-stage maximum likelihood procedure, 60
- effect confounded with non-normality, non-additivity, etc., 238, 247, 249
- graphical methods, 247–249
- half-normal plots, 248
- indicated by pattern disruption, 235, 238, 245
- masking, 251
- multiple decision approach, 240
- multiple outliers, 244, 247, 248, 249–251
- non-parametric methods, 251–252
- non-residual-based methods, 249–251
- sensitivity contours, 249
- swamping, 251
- table, 334
- tests of discordancy, 238–246
- use of residuals, 60, 236–237, 238–247 (*see also* Residuals)
- Detection of outliers**, 52
  - in multivariate data, 208–209, 210–211, 214–215
  - in time-series, 266–267
  - pattern disruption in designed experiments, 235, 238, 245
  - two-stage maximum likelihood procedure, 59, 60–61, 210–211, 214–215
- Deterministic model for explaining outliers**, 6–7, 14, 18, 23, 30–31
- Discordancy**, 23
  - discordancy test, 24 (*see also* Tests of discordancy)
- Discordant observations**, 23
  - multiple, 37
- Distance measures** (multivariate data), 208, 209
  - generalized distances, 224–227
  - graphical plots, 212–213
  - use in outlier detection, 208, 209, 215
- Distribution** (*see under specific heading*:
  - Binomial distribution, etc.)
- Dixon statistics** for tests of discordancy, 54, 55–56, 60
  - for exponential and gamma data, 78–79, 81, 82, 83, 85, 86–88
  - for univariate normal data, 91, 97–100

- 'Doubtful' observation, 18, 19 (*but see Outlier*)
- 'Error', 32
- Evil, 21
- Exchangeable model for outliers, 35–37, 50
- explosive distribution, 36
  - in Bayesian context, 272, 279
- Execution error, 27–28, 46
- Exponential distribution
- accommodation of outliers, 171–173
  - Bayesian accommodation, 279–281
  - Bayesian procedure for slippage, 206
  - details of discordancy tests for practical use, 75–88, 124–125; guide to use of tests, 75–76; contents list, 77–79; worksheets, 79–88, 124–125; tables, 290–297
  - estimating the mean, 35–36, 50, 51
  - exchangeable model for outliers, 35–36, 50
  - $L$ -estimator, 51
  - labelled slippage model, 72–73
  - locally optimal test for outlier, 59
  - maximum likelihood ratio test, 56–58
  - multiple decision procedure, 58–59
  - multiple outliers, 36–37
  - recursive derivation of discordancy test statistic, 62–64
  - robust tests, 35–36, 50, 51
  - shifted origin, 77–78
  - trimmed mean, 51
  - truncated, 124–125
  - two-stage maximum likelihood ratio test, 59
  - Winsorized mean, 51
- Extreme-value (Gumbel, Fréchet, Weibull) distributions
- discordancy tests for practical use, 115–116, 117–118
  - in Bayesian treatment of (Weibull) outliers; 281, 282
- Ferguson's slippage models *A* and *B* for normal outliers, 34
- model *A*, 34, 42, 167–168, 276
  - model *B*, 34, 42, 167–168, 276, 277
  - multivariate model *A*, 210–218, 219
  - multivariate model *B*, 59, 210, 218–219
- Folded median (Bickel–Hodges estimator), 154
- in normal sample, 165–166
- Fréchet distribution (*see* Extreme-value distributions)
- Gamma distribution
- Bayesian treatment of outliers, 281, 282
  - details of discordancy tests for practical use, 75–88; guide to use of tests, 75–76; contents list, 77–79; worksheets, 79–88; tables, 290–291, 294–295
  - distribution of  $x_{(n)}$ , 211
  - extreme/location statistic, 40
  - in relation to test for slippage of normal variance, 201
  - outlier proneness, 37
  - shifted origin, 77–78
  - slippage tests, 197, 201–202
- Gamma-type probability plots, 226, 230
- Glaisher's accommodation procedure, 20
- Goodwin's rejection test, 21
- Gross-error sensitivity, 140
- Gumbel distribution (*see* Extreme-value distributions)
- Hinge, 153
- Historical background, 18–22, 47
- Hodges–Lehmann estimator, 49, 154
- in normal data, 164, 169
- Huber's proposal 2, 151–152, 158
- Bickel one-step modification, 158
  - in normal data, 163–164, 169
- Inclusive and exclusive measures of discordancy, 61–64, 72
- Individual outliers in slipped samples, 183
- Influence curve, 136–141
- for correlation coefficient, 228–229
  - for trimmed mean, 145–147
  - for Winsorized mean, 146–147
  - mean-squared value as asymptotic variance, 139
- Influence function (*see* Influence curve)
- Inherent model for outliers, 9–10, 23, 31
- Jackknifing, 48, 161
- Kurtosis, 39, 42, 102, 109, 266

- L*-estimators (*see* Linear order statistics estimators)
- Labelled slippage model (*see* Slippage model for outliers)
- Linear model (outlier in), 253
  - accommodation, 260
  - Bayesian methods, 277
  - equivalence of use of residuals and of residual sums of squares, 262–265
  - graphical methods, 261
  - multiple decision approach, 240
  - multiple outliers, 249–251, 261, 265
  - non-residual-based methods, 264–265
  - residual-based methods, 257–264
  - slippage model, 262–264
  - table, 335–336
  - tests of discordancy, 257–265
  - two-stage maximum likelihood ratio procedure, 262–264
- (*see also* Designed experiments, Regression)
- Linear order statistics estimators (*L*-estimators), 48, 152–153
  - asymptotic normality, 162
  - Gastwirth's estimator, 153, 164
  - in exponential samples, 171–172
  - modified trimmed mean, 152
  - modified Winsorized mean, 152
  - Trimean, 153
- Local-shift sensitivity, 140
- Log-normal distribution
  - discordancy tests for practical use, 115–116, 118
  - outlier proneness, 37
- M*-estimators (*see* Maximum likelihood type estimators)
- Masking, 38, 40, 43, 71, 74, 251
  - generalized, 287 (*see also* 'Cloaking')
- Maximin robust estimator, 135, 156, 157
- Maximum likelihood ratio principle
  - in linear model, 262–264
  - in tests for multivariate outliers, 210–211, 214–215, 219
  - in tests of discordancy, 41, 56–58
  - in time-series, 267–268
  - two-stage, 59, 210–211, 214–215, 262–264
- Maximum likelihood type estimators (*M*-estimators), 48–49, 149–152
  - asymptotic normality, 162
  - for dispersion in normal data, 169–170
  - for location in normal data, 163–166
- Huber's proposal 2, 151–152
- one-step Huber estimators, 151
- three-part descending *M*-estimators, 151
  - trimmed mean, 150
  - Winsorized mean, 150
- Mean deviation, 159
- Measurement error, 27–28, 37, 46
- Median, 46
  - influence of contaminants on median and mean, 136–137, 138–139, 140
- Median deviation, 150, 158
- Mendeleev's accommodation procedure, 21
- Mid-mean, 145
- Minimax robust estimator, 134, 156, 157, 169
- Mixture model for outliers, 23, 31–33
  - as model for contamination: accommodation aspect, 127–130
  - normal, 33
- 'Modified' residuals, 249
- Multinomial distribution (slippage test), 203
- Multiple decision procedure
  - for outlier in general linear model, 240
  - in Bayesian treatment, 279
  - in multivariate normal tests of discordancy, 218–219
  - in slippage tests, 192, 193–194, 196–197, 200
  - in univariate tests of discordancy, 43, 44, 58–59
- Multiple outliers, 44, 68–74
  - block tests of discordancy, 40, 69–73
  - consecutive tests of discordancy, 40–41, 69–71, 73–74
  - in Bayesian context, 272–275
  - in designed experiments, 244, 247, 248, 249–251
  - in exponential samples, 36–37
  - in linear models, 249–251, 265
  - in multivariate normal samples, 215–218
  - masking effect (*see* Masking)
  - number of outliers, 68–69
  - relative performance of block and consecutive tests, 70–71
  - 'sequential' tests of discordancy, 40, 41, 69, 73
- swamping effect (*see* Swamping)
- Multivariate normal distribution

- tables for tests of discordancy, 329–333  
 tests of discordancy for outliers, 209–219, 230–231  
**Multivariate outliers**, 14, 59, 208–233  
 accommodation, 231–233  
 Andrews' Fourier-type plot, 217, 218  
 Bayesian method, 277  
 detection, 208–209  
 detection by two-stage maximum likelihood procedure, 60–61, 210–211, 214–215  
 distance measures, 208, 209, 224–227  
 estimation of correlation coefficient, 233  
 estimation of mean, 231–232  
 estimation of variance–covariance matrix, 232–233  
**Ferguson models A and B**, 210–219  
 gamma-type plots, 226, 230  
 ‘gap test’, 229–231  
 generalized distances, 224–227  
 graphical methods, 221–223, 224–225, 226, 230  
 influence function for correlation coefficient, 228–229  
 informal methods, 219–233  
 internal scatter, 214, 215  
 linear constraints, 221  
 marginal outliers, 220  
 maximum likelihood ratio procedure, 210–211, 214–215, 219  
 minimum spanning tree, 229–230  
 multiple decision approach, 218–219  
 multiple outliers, 215–218  
 normal samples, 209–219  
 outlier scatter ratio, 215, 216  
 plots of distance measures, 212–213  
 premium-protection procedures (normal samples), 231–232  
 tables for tests of discordancy, 329–333  
 use of correlation coefficients, 227–229  
 use of principal components, 223–226, 227  
  
**Negative binomial distribution (slippage test)**, 204  
**Newcomb's accommodation procedure**, 21  
**Non-parametric slippage tests of location**, 176–187  
 block procedure, 184  
 consecutive procedures, 184–185  
 equal sized samples, 176–177, 178–183  
 models, 182–183  
 multiple sample slippage, 183–186  
 performance (power), 179  
 rank methods, 180–183  
 single sample slippage, 176–183  
 tables, 324–326  
 unequal sized samples, 177–178, 179  
**Non-parametric treatment of outliers**, 15, 44, 282–285  
 accommodation, 285  
 in designed experiments, 251–252  
**non-parametric slippage tests**, 176–187 (*see Slippage tests*)  
 philosophical considerations, 282–283, 285  
 tests of discordancy, 283–285  
**Normal distribution**  
 accommodation of outliers in multivariate samples, 209, 231–233  
 accommodation of outliers in univariate samples, 163–171  
 Bayesian accommodation of outliers, 33, 277–282  
 Bayesian ‘detection of spuriousity’, 275–277  
 details of univariate discordancy tests for practical use, 75–76, 89–115; guide to use of tests, 75–76; content list, 91–93; worksheets, 93–115; tables, 298–315  
 effect of contaminant on estimation of mean and variance, 128–129, 132–133  
**Ferguson models A and B** (*see Ferguson models*)  
 maximum likelihood ratio test, 59–60  
 multiple decision procedure, 59–60  
 multivariate discordancy tests (*see Multivariate normal distribution*)  
 multivariate slippage model, 59  
 outlier proneness, 37  
 slippage models, 33, 34–35, 42–44, 45–46, 59–60  
 slippage tests (mean and variance), 188–197; tables, 290–291, 302, 303, 327  
 testing of outliers in multivariate samples, 209–219  
 undetectability of contamination, 129–130

- Old French custom, 26
- One-step Huber estimators, 151
- Ordering in multivariate data, 208, 209
  - distance measures, 208, 209
  - marginal ordering, 209
  - reduced ordering, 209
  - sub-ordering, 208–209
- (*see also* Multivariate outliers)
- Outlier**
  - accommodation, 4, 20, 24–26, 236, 271 (*see also* Accommodation of outliers)
  - as distinct from discordant observation, 23
  - as distinct from extreme observation, 23
  - basic model (working hypothesis) against which outliers are examined, 26, 29, 285
  - Bayesian methods, 200–201, 269–282 (*see also* Bayesian treatment of outliers)
  - causes, 26–28
  - classification of outlier problems, 5, 16
  - definition, 4, 22–23, 286–287
  - detection (non-subjective), 52 (*see* Detection of outliers)
  - different aims in handling outliers (schematic diagram), 28
  - identification, 8, 24–26, 236, 271
  - in designed experiment (*see* Designed experiments)
  - in regression (*see* Regression)
  - in time-series (*see* Time-series)
  - masking effect (*see* Masking)
  - models for explaining outliers (*see* Outlier-generating models)
  - multiple, 40, 44, 68–74 (*see also* Multiple outliers)
  - multivariate (*see* Multivariate outliers)
  - non-parametric methods, 176–187, 282–285 (*see also* Non-parametric treatment of outliers)
  - outlying sub-sample (*see* Slippage tests)
  - rejection, 1, 18–19, 24–26, 236, 271
  - relative nature (to model), 5, 7–10, 16
  - significance level (subjective aspect), 64
  - subjective nature, 4, 15, 22–23, 45, 64, 286–287
  - ‘surprise’, 1, 10, 14, 32, 37, 234, 270, 273, 282, 285, 286–288
  - Outlier-generating models, 23, 28–37
  - contamination models, 47
  - deterministic, 6–7, 14, 18, 23, 27–28, 30–31
  - exchangeable, 35–37, 50, 272
  - in time-series, 266–267
  - inherent, 9–10, 23, 31
  - intangibility in non-parametric context, 285
  - irrelevant in Bayesian approach?, 15
  - mixture, 23, 31–33, 278
  - non-specific with respect to the outlier, 33, 45
  - slippage, 34–35, 42, 43
  - (*see also* under specific headings)
  - Outlier-robust methods (*see* Accommodation of outliers)
  - Outlier proneness, 37
  - outlier-prone, 38
  - outlier-prone completely, 38
  - outlier-resistant, 38
  - Outlying sub-sample, 236, 282 (*see* Slippage tests)
  - Parametric slippage tests, 187–205
  - Bayes solution, 192, 193
  - binomial samples, 203–204
  - gamma samples, 197, 201–202, 206
  - group tests (multiple slipped samples), 204–205
  - multinomial samples, 203
  - multiple decision approach, 192, 193–194, 196–197, 200
  - negative binomial samples, 204
  - normal samples, 188–197
  - normal samples, mean, 188–190, 192–196
  - normal samples, variance, 190–191, 196–197
  - optimality of Paulson procedure, 192
  - Poisson samples, 202–203
  - relationship with single outlier tests, 188–189
  - slippage in unspecified direction, 194
  - tables, 290–291, 302, 303, 327, 328
  - unequal sample sizes, 194–195, 197
  - use of ranges, 197
  - Pareto distribution, discordancy tests for practical use, 115–116
  - Peirce’s rejection test, 19

- Performance criteria for discordancy tests, 30, 41, 43–44, 45, 64–68
- Single outliers:*  
David's measures  $P_1$  (power),  $P_2$ ,  $P_3$ ,  $P_4$ ,  $P_5$   
for slippage model, 65–67  
measures for inherent model, 68  
measures for mixture model, 67–68
- Multiple outliers:*  
comparison of consecutive and block tests, 70–71  
measures for block tests, 73  
measures for consecutive tests, 73–74
- Performance criteria in outlier-robust inference  
asymptotic measures, 134–136, 137  
breakdown point, 140–141  
error frequency, 142  
error probability, 142  
exponential samples, 171–173  
finite-sample measures, 131–133, 135, 137  
for confidence intervals, 141–142  
for location estimators, 155–158  
for significance tests, 142–144  
gross-error sensitivity, 140  
influence curve, 136–141  
local-shift sensitivity, 140  
rejection point, 140  
univariate normal samples, 163–171  
univariate samples (general), 130–144
- Poisson distribution  
details of discordancy tests for practical use, 75–76, 115–116, 120–122; guide to use of tests, 75–76; worksheets, 121–122; tables, 316–319, 328  
modified, 8  
slippage tests for Poisson samples, 202–203; table, 328
- Power of discordancy test, 65–66, 70  
(*see also* Performance criteria for discordancy tests)
- Premium and protection  
for multivariate normal samples, 231–232  
in designed experiments, 246–247  
in general linear models, 246–247, 260  
in relation to hypothesis testing, 51, 144, 172  
in time-series, 268
- location estimates (normal samples), 167–168  
premium, 50, 131–132  
premium-protection rules, 50, 131, 231–232  
protection, 50, 131–132  
scale estimates (exponential samples), 173
- Principal component analysis (in study of multivariate outliers), 223–226, 227
- Protection (*see* Premium and protection)
- R-estimators (*see* Rank test estimators)
- Rank test estimators (R-estimators), 49, 153–154  
asymptotic normality, 162  
folded-median estimator (Bickel–Hodges estimator), 154  
Hodges–Lehmann estimator, 154
- Recursive algorithm for null distribution of discordancy test statistic, 62–64, 72, 75
- Regression (outliers in), 11, 252–257  
accommodation, 256  
detection by two-stage maximum likelihood procedure, 60–61  
example, 11–12  
linear regression, 252–255  
multiple regression, 256–257  
tests of discordancy, 252–257  
use of residuals, 11, 60–61, 252–257
- Rejection point, 140
- Residuals, 11, 60, 236–237, 238–247, 257–264  
adjusted residuals, 168, 232, 260  
'ballooning', 253  
estimated residuals, 236, 239, 250, 253  
graphical display, 247–249  
in designed experiments, 236–237, 238–247  
in linear model, 253, 257, 264  
largest absolute residual, 240, 242, 246  
maximum absolute studentized residual, 240  
maximum normal residual, 242–243  
'modified' residuals, 249  
normalized residuals, 260  
studentized residuals, 240, 258–264  
(*see also* Designed experiments, Linear model, Regression)

- Robust inference procedures (robust against outliers) (*see* Accommodation of outliers)
- Robustness of efficiency, 142
- Robustness of performance, 126, 142, 161
- Robustness of validity, 126, 142, 161
- 'Rogue' observation, 1 (*but see* Outlier)
- 'Scatter' in multivariate data  
  internal scatter, 214, 215  
  outlier scatter-ratios, 215, 216
- Self-camouflaging effect of outliers (*see* 'Cloaking')
- Semi-interquartile range, 158
- Sensitivity contours, 249
- Sensitivity curve, 139
- Sequential methods for slippage, 206–207
- 'Sequential' tests of discordancy (*see* Multiple outliers)
- 'Shorth', 153
- Significance tests robust against outliers, 142–144, 160–162  
  performance criteria, 142–144  
  'premium' and 'protection', 144
- Skewness, 39, 42, 100–101, 266
- Skipping, 153  
  iterative, 153  
  multiple, 153
- Slippage model for outliers, 34–35  
  gamma, 44  
  in linear models, 262–264  
  in time-series, 266–267  
  labelled and unlabelled slippage models, 57–58, 59–60, 72–73, 259  
  normal, 33, 34–35  
  slippage in dispersion, 34–35, 43  
  slippage in location, 34–35, 43  
  (*see also* Ferguson's slippage models A and B)
- Slippage tests, 42–43, 174–207  
  Bayesian approach, 192, 193, 200–201, 206  
  general method for constructing slippage tests, 197–200  
  individual outliers in slipped samples, 183  
  masking, 186  
  non-parametric tests of location, 176–187 (*see* Non-parametric slippage tests of location)
- parametric tests, 187–205 (*see* Parametric slippage tests)
- sequential approach, 206–207
- slippage of mean, 174
- slippage of variance, 174
- the slippage model, 186–187
- 'Spurious' observation, 1, 8, 27, 45–46, 275–277 (*but see* Outlier)
- Stability aspect of robustness, 141
- Stone's rejection test, 19–20
- 'Suspect' observation, 23 (*but see* Outlier)
- 'Suspicious' observation, 4, 10, 15, 29, 220, 222 (*but see* Outlier)
- Swamping, 71, 251
- Tests of discordancy for outliers, 24, 29, 47, 52–56
- block tests (*see* Multiple outliers)
- consecutive tests (*see* Multiple outliers)
- details of univariate tests for practical use, 75–125; tables, 290–322, 328
- Dixon statistics (*see* Dixon statistics)  
in context of Bayesian approach, 269–277
- in designed experiments, 238–246
- in linear models, 257–265
- in regression, 252–257
- in relation to slippage test, 201
- in time-series, 267–268
- intuitively constructed tests, 52–56, 72
- invariance considerations, 52–56
- maximum likelihood ratio procedure, 41
- measures of performance (e.g. power), 30, 41, 43–44, 45, 64–68 (*see also* Performance criteria for discordancy tests)
- multiple decision procedure, 43, 44
- multivariate normal samples, 209, 219, 230–231
- non-parametric, 283–285
- one-sided versus two-sided tests, 44–45
- optimality criteria, 41–42
- recursive algorithm for null distribution of test statistic, 62–64
- 'sequential' (*see* Multiple outliers)
- statistical principles, 41–46, 56–61
- types of test statistic, 38–41;

- deviation/spread, 39–61;  
 excess/spread, 38–39, 61;  
 extreme/location, 40–41; high-order, 39–40; inclusive and exclusive measures, 61–64; omnibus, 40; range/spread, 39; sums of squares, 39
- univariate samples, 52–125
- Time-series (outliers in), 11, 266–268  
 accommodation, 268  
 ‘bumps and quakes’, 266–267  
 detection, 266–267  
 maximum likelihood ratio tests, 267–268  
 tests of discordancy, 267–268  
 types of outlier, 266–267
- Transformation of variables for discordancy testing, 77, 88, 89–90, 115, 116, 117, 118, 120–121, 123
- Trimean, 153
- Trimmed mean (*see* Trimming)
- Trimming, 21, 26, 144  
 as *L*-estimator, 48  
 $\alpha$ -trimmed mean, 26, 48  
 in exponential samples, 173  
 in normal samples, 163–166, 167–169, 170–171, 232, 233  
 influence curve for trimmed mean, 145–147  
 mid-mean, 145  
 modified trimming, 147–148, 166–168  
 $r$ -fold symmetrically trimmed mean, 145  
 $(r, s)$ -fold trimmed mean, 144  
 scale estimators, 162, 170–171  
 trimmed mean, 26, 51
- Unifrom distribution, details of discordancy tests for practical use, 115–116, 118–120
- ‘Unreasonable’ observation, 23 (*but see Outlier*)
- ‘Unrepresentative’ observation, 1, 10 (*but see Outlier*)
- W**-statistics of Shapiro and Wilk  
 for exponential sample, 88  
 for normal sample, 40, 102
- Weibull distribution (*see* Extreme-value distributions)
- Weighting of observations, 20
- Winsorization, 26, 50, 144  
 and *L*-estimators, 48  
 $\alpha$ -Winsorized mean, 145  
 in exponential samples, 171–172  
 in normal samples, 163–166, 167–169, 170–171, 232, 233  
 influence function for Winsorized mean, 146–147  
 modified Winsorization, 50, 147–148, 166–168, 170–171  
 $r$ -fold symmetrically Winsorized mean, 145  
 $(r, s)$ -fold Winsorized mean, 145  
 scale estimators, 159  
 semi-Winsorization, 148, 166–169, 170–171, 232  
 Winsorized mean, 48, 51
- Winsorized mean (*see* Winsorization)
- Working hypothesis (*see* Basic model)
- Wright’s rejection test, 2, 21