

NmLcU/IRU New Mexico State Univ



1011202590

Monographs on
Applied Probability
and Statistics

Identification of Outliers

D.M. Hawkins

QA
276
.H388

Chapman and Hall

Identification of Outliers

D. M. HAWKINS

*Senior Consultant in Operations Research and Statistics,
Council for Scientific and Industrial Research, South Africa*



1980

LONDON NEW YORK
CHAPMAN AND HALL
150TH ANNIVERSARY

QA
276
.H388

Content

Preface

1 Introduction

- 1.1 What is an outlier?
- 1.2 Genesis of outliers
- 1.3 Treatment of outliers
- 1.4 Slippage tests
- 1.5 The meaning of significance level
- 1.6 Brief early history of outlier tests

2 General theoretical principles

- 2.1 Measures of performance
- 2.2 Specific optimal tests
- 2.3 Use of 'departure from model'
- 2.4 Inadequacies of current optimality criteria
- 2.5 Discrete random variables
- 2.6 Relationship with tolerance limits
- 2.7 Null hypothesis distributions
- 2.8 Relations between measures of performance

3 A single outlier in normal samples

- 3.1 Notation
- 3.2 Optimal statistics
- 3.3 Non-optimal statistics
- 3.4 Performance of the outlier tests
- 3.5 The use of normal approximation

4 The gamma distribution

- 4.1 Introduction
- 4.2 The problem of unequal variances
- 4.3 Performance of the tests
- 4.4 Comparison with the Bartlett test
- 4.5 The maximum F -ratio

First published 1980 by
Chapman and Hall Ltd
11 New Fetter Lane, London EC4P 4EE

Published in the USA by
Chapman and Hall
in association with Methuen, Inc.
733 Third Avenue, New York, NY 10017

© 1980 D. M. Hawkins

Printed in Great Britain
at the University Printing House, Cambridge

ISBN 0 412 21900 X

All rights reserved. No part of
this book may be reprinted, or reproduced
or utilized in any form or by any electronic,
mechanical or other means, now known or hereafter
invented, including photocopying and recording,
or in any information storage or retrieval
system, without permission in writing
from the publisher

British Library Cataloguing in Publication Data

Hawkins, D M

Identification of outliers. – (Monographs on
applied probability and statistics).

1. Outliers (Statistics)

I. Title II. Series

519.5 QA276

ISBN 0-412-21900-X

832814

NEW MEXICO STATE UNIVERSITY

Preface

The problem of outliers is one of the oldest in statistics, and during the last century and a half interest in it has waxed and waned several times. Currently it is once again an active research area after some years of relative neglect, and recent work has solved a number of old problems in outlier theory, and identified new ones. The major results are, however, scattered amongst many journal articles, and for some time there has been a clear need to bring them together in one place. That was the original intention of this monograph: but during execution it became clear that the existing theory of outliers was deficient in several areas, and so the monograph also contains a number of new results and conjectures.

In view of the enormous volume of literature on the outlier problem and its cousins, no attempt has been made to make the coverage exhaustive. The material is concerned almost entirely with the use of outlier tests that are known (or may reasonably be expected) to be optimal in some way. Such topics as robust estimation are largely ignored, being covered more adequately in other sources. The numerous *ad hoc* statistics proposed in the early work on the grounds of intuitive appeal or computational simplicity also are not discussed in any detail.

The main emphasis of the monograph is on the salient theoretical aspects of the problem, but the appendices of fractiles of many of the tests discussed will also be of use to practitioners interested in applying outlier tests to data.

I wish to thank the Universities of the Witwatersrand and North Carolina – the former for granting a year's sabbatical leave in which to write this monograph, and the latter for providing facilities for preparing it and carrying out some of the original research it embodies. The final preparation was carried out under the aegis of the Council for Scientific and Industrial Research, and all three bodies have provided me with invaluable facilities and considerable amounts of computer time for the computation of tables.

PREFACE

publishers of the following journals for
light material from their pages:
for Appendix 1;
Computation and Simulation for Appendix 3;
for Appendix 5;
al Statistics for Appendix 6.
leagues who helped with the preparation of
buted ideas, in particular Drs R. Carroll and
ussions of early drafts, Mrs J. Galpin for the
es in Appendices 8 and 10, and Miss S. van
cial symbols into the manuscript.
e gratefully the help of my wife, Greer, who
manuscript for computer formatting, helped
rams and cheerfully tolerated the disruption
monograph.

D. M. Hawkins

CHAPTER 1

Introduction

1.1 What is an outlier?

Any applied statistician who has analysed a number of sets of real data is likely to have come across 'outliers'. The intuitive definition of an outlier would be 'an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism'. An inspection of a sample containing outliers would show up such characteristics as large gaps between 'outlying' and 'inlying' observations and the deviation between the outliers and the group of inliers, as measured on some suitably standardized scale.

1.2 Genesis of outliers

There are two basic mechanisms which give rise to samples which appear to have outliers. It is a matter of some importance which of the mechanisms generated any particular set of observations since this consideration certainly affects, or should affect, one's subsequent analysis of the data.

Mechanism (i): The data come from some heavy tailed distribution such as Student's t . There is no question that any observation is in any way erroneous.

Formalizing this model, Green (1976) has introduced a classification of families of statistical distributions into those that are 'outlier-prone' and those that are 'outlier-resistant'. The 'outlier-prone' families have tails which go to zero slowly: a distribution is said to be absolutely outlier-prone if (letting $X_{n,i}$ be the i th order statistic based on a sample of size n) there exists $\varepsilon, \delta > 0$ and an integer n_0 such that

$$\Pr[X_{n,n} - X_{n,n-1} > \varepsilon] \geq \delta \quad \text{for all } n > n_0 \quad (1.1)$$

and is relatively outlier-prone if there exist $c > 1, \delta > 0$ and n_0 such

that

$$\Pr[X_{n,n}/X_{n,n-1} > c] \geq \delta \quad \text{for all } n > n_0 \quad (1.2)$$

Clearly if either of these situations holds, then there will be a tendency for the largest order statistic to be suspiciously large relative to its predecessor, and so samples generated by outlier-prone distributions will tend to contain visual outliers.

Absolutely and relatively outlier-resistant distributions are those which are not absolutely and relatively outlier-prone respectively. The former class includes the normal family of distributions and the latter the gamma family.

Mechanism (ii): The data arise from two distributions. One of these, the 'basic distribution', generates 'good' observations, while another, the 'contaminating distribution', generates 'contaminants'. If the contaminating distribution has tails which are heavier than those of the basic distribution, then there will be a tendency for the contaminants to be outliers – that is, to separate visibly from the good observations, which will then constitute the inliers.

When this mechanism is appropriate, it may often be invoked in either of two ways. The first, *mechanism (iia)*, specifies that in the sample of size n , exactly $n - k$ observations come from the basic distribution, and k from the contaminating distribution. For this model to be of use presupposes a knowledge of k – the number of contaminants in the sample.

More commonly, k is not known. A suitable model here is the Dixon (1950) *mechanism (iib)*: with probability p , any given observation comes from the contaminating distribution, while with probability $1 - p$ it comes from the basic distribution. The value of k is then a random variable following a binomial distribution. In most cases of interest, p is assumed to be near zero, even if its actual value is not known.

We may note that if the basic distribution is $f_0(\cdot)$ and the contaminating distribution is $f_1(\cdot)$ then the distribution of an arbitrary observation is

$$f(\cdot) = (1 - p)f_0(\cdot) + pf_1(\cdot) \quad (1.3)$$

Viewed in this way, the mechanism iib thus reduces to a particular case of mechanism (i), the heavy-tailed distribution model. It is, however, a case of especial interest, since problems of estimation will generally involve the individual parameters of the distributions $f_0(\cdot)$

and $f_1(\cdot)$, rather than those of $f(\cdot)$. The distinction between the two models.

1.3 Treatment of outliers

1.3.1 Estimation

The effect of outliers on the analysis of data on the mechanism by which the outliers arise. If mechanism (i) is assumed, then the outlier-prone observations from the distribution are not valid observations from the distribution. The major objective of the analysis will be to estimate the location – of this distribution. If mechanism (i) is known (for example, if it is a Cauchy distribution), estimation poses no especial difficulties. The estimator is best for that particular distribution.

If the distributional form is not known, it is to use a robust estimator. To estimate the location might use a trimmed mean – the mean of the sample. While the trimmed mean is a good estimator of the location of a certain, highly specific distribution, it is obviously good under conditions in which a heavy tail gives rise to some outliers. Statistics behave like a sample from the distribution, likely to be at least approximately true for distributions with heavy tails.

We may note that the use of a trimmed mean for $n - k$ values consists of discarding the k values on each side, regardless of their magnitude. Contrast this procedure with the 'outliers' procedure – 'discard all those observations with a deviation from the remainder of the sample greater than a certain location estimator the arithmetic mean'.

This comparison shows that the 'rejection' procedure may be regarded as a method of obtaining the advantage over the trimmed mean. If the data are normal, then the fully efficient arithmetic mean can be used as the estimator. Thus it is a suitable selection of a procedure for dealing with outliers. The extreme values are sufficiently large that a procedure can be made to give estimates for distributions with heavier tails than wa

$$[x_{n-k} - 1 > c] \geq \delta \quad \text{for all } n > n_0 \quad (1.2)$$

these situations holds, then there will be a order statistic to be suspiciously large relative and so samples generated by outlier-prone to contain visual outliers.

vely outlier-resistant distributions are those ly and relatively outlier-prone respectively. des the normal family of distributions and family.

arise from two distributions. One of these, generates 'good' observations, while another, 'distribution', generates 'contaminants'. If the on has tails which are heavier than those of then there will be a tendency for the cons - that is, to separate visibly from the good l then constitute the inliers.

n is appropriate, it may often be invoked in e first, *mechanism (iia)*, specifies that in the ly $n - k$ observations come from the basic n the contaminating distribution. For this supposes a knowledge of k - the number of mple.

s not known. A suitable model here is the (iib): with probability p , any given obser- e contaminating distribution, while with s from the basic distribution. The value of k is following a binomial distribution. In most med to be near zero, even if its actual value is

if the basic distribution is $f_0(\cdot)$ and the ion is $f_1(\cdot)$ then the distribution of an

$$f = (1 - p)f_0(\cdot) + pf_1(\cdot) \quad (1.3)$$

e mechanism iib thus reduces to a particular the heavy-tailed distribution model. It is, al interest, since problems of estimation will individual parameters of the distributions $f_0(\cdot)$

and $f_1(\cdot)$, rather than those of $f(\cdot)$. Thus it is well to maintain the distinction between the two models.

1.3 Treatment of outliers

1.3.1 Estimation

The effect of outliers on the analysis of a set of data depends strongly on the mechanism by which the outliers are believed to be generated. If mechanism (i) is assumed, then the outliers, despite appearances, are valid observations from the distribution under study. Usually, the major objective of the analysis will be to estimate a parameter - for example location - of this distribution. If the form of the distribution is known (for example, if it is a Cauchy distribution), then this estimation poses no especial difficulties - one uses whatever estimator is best for that particular distributional form.

If the distributional form is not known, then one would be tempted to use a robust estimator. To estimate location, for example, one might use a trimmed mean - the mean of the central order statistics of the sample. While the trimmed mean is an optimal estimator of the location of a certain, highly specific distributional family, it is an estimator that is obviously good under very broad conditions - those in which a heavy tail gives rise to some outliers, while the central order statistics behave like a sample from the normal distribution. This is likely to be at least approximately true for most symmetric unimodal distributions with heavy tails.

We may note that the use of a trimmed mean using, say, the central $n - k$ values consists of discarding the $k/2$ most extreme observations on each side, regardless of their magnitude, and finding the mean of those left. Contrast this procedure with the following 'rejection of outliers' procedure - 'discard all those extreme order statistics whose deviation from the remainder of the sample is sufficiently large. Use as location estimator the arithmetic mean of the data retained.'

This comparison shows that the 'rejection of outliers' approach may be regarded as a method of obtaining a robust estimator. It has the advantage over the trimmed mean that if the sample is truly normal, then the fully efficient arithmetic mean of all observations will often be used as the estimator. Thus it is very plausible that, with suitable selection of a procedure for deciding whether the deviations of the extreme values are sufficiently large, the 'rejection of outliers' procedure can be made to give estimators which are both robust for distributions with heavier tails than was supposed; and efficient, in

that, if the distribution assumed is correct, use is frequently made of the fully efficient estimator for that distribution.

The idea of using rejection as a method of obtaining robust estimators was set out formally by Anscombe (1960). Using a mixture model, mechanism (iib), he set up a rejection rule in terms of an 'insurance policy': As a result of being willing to occasionally reject a good observation, one incurs a cost in that the mean squared error of the resulting estimate applied to uncontaminated samples will be larger than would otherwise be the case. If we regard this increase as being like an insurance premium, then the protection that it buys is a decrease in mean squared error when the sample is actually contaminated. Some simulation studies have shown that a large amount of protection may be bought for a small premium.

While these comments apply in quite general terms, there is an especially compelling interest in the rejection approach if mechanism (ii) (in either of its variants) is believed to hold, for then, one can regard rejection as a method of partitioning the data from $f_0(\cdot)$ and those from $f_1(\cdot)$ prior to using the partitioned data to estimate the parameters of $f_0(\cdot)$ and $f_1(\cdot)$.

In the multivariate context, of course, mechanism (iib), mixtures of distributions, is familiar as the mathematical model underlying many problems in cluster analysis (see for example Hartigan, 1975). The standard methods of cluster analysis include two broad classes: one seeks to estimate p and the parameters of $f_0(\cdot)$ and $f_1(\cdot)$ from the entire sample, after which a discriminant analysis may be used to partition the data; the other starts by finding an optimal partitioning of the data – from this the parameters may be estimated. The similarity between the latter approach and rejection of outliers is unmistakable.

There is another area in which the use of rejection of outliers represents a more viable approach to robust estimation, and that is the situation in which the data are not identically distributed. A good example of this is multiple regression. While robust multiple regression procedures exist, they are still rather awkward computationally. The approach of eliminating several outliers and estimating on the basis of the remainder of the sample involves quite simple matrix operations, and can lead to considerable insight into the structure of the underlying process (see Chapter 7).

1.3.2 Smoothly varying weights

A valid objection to the outlier rejection approach is the influence on

the final estimator of an individual observation. The location estimator consisting of the observations retained. The derivative of the particular observation is a constant until a cutoff point at which it is rejected; the derivative is zero. The undesirability of this is clear, and Glaisher (1872) proposed an alternative location of the underlying distribution as the mean of the data. The weights are computed from a Bayesian model in which the standard deviations, but are normally mean.

This proposal seems to be the ancestor suggested more recently (see Chapter 9) for the basic distribution in the face of outliers.

Another approach to estimation is the idea underlying this is that all observations excessively outlying are moved to a prefix of observations. In this way, the outlying observations are entirely in estimating the parameters of the basic distribution.

When deciding whether to use Winsorization of outlying observations, one should be aware of the model of the data. If the observations are from a heavy-tailed distribution, and one wants to estimate the parameters of this distribution, then the outlying observations. Thus one should be reluctant to use Winsorization, which is a partial use of the outliers.

If, on the other hand, one believes that the data are from a basic distribution, and one is interested in estimating the parameters of this distribution, then Winsorization should be avoided. Outliers may be presumed to come from a heavy-tailed distribution, and hence to contain no information about the basic distribution. Thus to the extent that one is interested in the basic distribution, one should ignore the outliers in the classical way by deleting them as soon as they are identified. A Bayesian way by assigning them smaller weights than those from the basic distribution.

Of course this distinction, quite clear in theory, is somewhat blurred in practice – for example,

IDENTIFICATION OF OUTLIERS

assumed is correct, use is frequently made of the estimator for that distribution.

Rejection as a method of obtaining robust estimates was formally proposed by Anscombe (1960). Using a mixture model, he set up a rejection rule in terms of an increase in the result of being willing to occasionally reject a hypothesis. It incurs a cost in that the mean squared error of the estimator applied to uncontaminated samples will be increased. Otherwise, it will be the case. If we regard this increase as a premium, then the protection that it buys is a reduction in the error when the sample is actually contaminated. Empirical studies have shown that a large amount of protection is bought for a small premium.

When we apply in quite general terms, there is an interest in the rejection approach if mechanism (i) is believed to hold, for then, one can regard the data as partitioning the data from $f_0(\cdot)$ and those from $f_1(\cdot)$. Using the partitioned data to estimate the parameters of $f_0(\cdot)$ and $f_1(\cdot)$.

In this context, of course, mechanism (iib), mixtures of distributions, is the mathematical model underlying many statistical analyses (see for example Hartigan, 1975). The latter analysis include two broad classes: one in which the parameters of $f_0(\cdot)$ and $f_1(\cdot)$ are estimated from the data, and a discriminant analysis may be used to decide whether the data starts by finding an optimal partitioning of the data. The parameters may be estimated. The latter approach and rejection of outliers is

in which the use of rejection of outliers is an alternative approach to robust estimation, and that is when the data are not identically distributed. A good example is multiple regression. While robust multiple regression, they are still rather awkward compared with the approach of eliminating several outliers and then estimating the remainder of the sample involves quite a bit of work, and can lead to considerable insight into the underlying process (see Chapter 7).

The influence of the outlier rejection approach is the influence on

INTRODUCTION

5

the final estimator of an individual observation. Take, for example, the location estimator consisting of the arithmetic mean of the observations retained. The derivative of this quantity with respect to a particular observation is a constant until that observation reaches the cutoff point at which it is rejected; the derivative then drops at once to zero. The undesirability of this is clear, and more than a century ago, Glaisher (1872) proposed an alternative. His suggestion is that the location of the underlying distribution be estimated by a weighted mean of the data. The weights are computed iteratively, and follow from a Bayesian model in which the observations have different standard deviations, but are normally distributed with common mean.

This proposal seems to be the ancestor of Bayesian approaches suggested more recently (see Chapter 9) for estimating the parameters of the basic distribution in the face of outlier contamination.

Another approach to estimation is the use of Winsorization. The idea underlying this is that all observations which are regarded as excessively outlying are moved to a prefixed position near the central observations. In this way, the outlying observations are not discarded entirely in estimating the parameters of interest.

When deciding whether to use Winsorization or outright rejection of outlying observations, one should be guided by the underlying model of the data. If the observations are generated by mechanism (i), the heavy-tailed distribution, and one wishes to estimate the parameters of this distribution, then the outliers represent valid observations. Thus one should be reluctant to discard them entirely, and hence prefer to use Winsorization, which is robust, but does make partial use of the outliers.

If, on the other hand, one believes that mechanism (ii) is operative, and one is interested in estimating the parameters of the basic distribution, then Winsorization should not be used. In this case, the outliers may be presumed to come from the contaminating distribution, and hence to contain no information about the basic distribution. Thus to the extent that one is sure they are not from the basic distribution, one should ignore them. This may be done in a classical way by deleting them as soon as one rejects the null hypothesis that they come from the basic distribution, or in a Bayesian way by assigning them smaller weights as they deviate more from the basic distribution.

Of course this distinction, quite clear in theory, may become somewhat blurred in practice – for example, when the basic and the

contaminating distributions are believed to share a common parameter. If this is the case, then it is no longer optimal to discard contaminants, but the manner in which they are incorporated into the estimation will depend on the specific distributions in the model. In the light of this comment it is of interest to note that many studies of the performance of estimators for contaminated samples have assumed a parameter in common between the two distributions.

1.3.3 'Sounding a warning'

Another possible use of a procedure for identifying outliers is to provide a check on model assumptions. For example, if one wishes to carry out some parametric analysis assuming normality of data, then various checks will be applied to the data to ensure that the model fits well enough. One such test could well be an outlier test, and such a test will be sensitive to the tail behaviour of the true distribution. This may be an advantage if the parametric analysis is sensitive to departures from model in the tails of the distribution, but not elsewhere. This is the case for many techniques (for example analysis of variance) where skewness and high kurtosis are the most damaging departures from model.

As the phrase 'sounding a warning' implies, this use of an outlier identification procedure is made in much the same way as a preliminary test for homogeneity of variance in an analysis of variance. The test is regarded as a screen – if the data pass, then the standard parametric procedure will be applied; otherwise some other, unspecified, action will be taken. The range of actions includes, but is by no means restricted to, removing any outliers from the sample and carrying out the analysis proposed originally on the remaining 'clean' observations.

The fact that any outlier test may be used to test for departures from model has a corollary: since the presence of outlier is indicative of some departure from the model assumed for the data, a goodness-of-fit test statistic may, in principle, be used to test for the presence of outliers. Two goodness-of-fit tests which are very effective for detecting the presence of outliers in a normal sample are the moment coefficients of skewness and kurtosis, and both in fact are optimal in a particular sense.

A more general class of goodness-of-fit tests is that based on the sample distribution function. Provided all parameters are known (or can be estimated from a sufficiently large sample) the probability integral transformation may be used to bring this to a standard form.

INTRODUCTION

The Anderson–Darling test, which is es-
tures from model in the tails of the distr
appeal as an outlier test statistic. Its use
proposed in the literature, but it is to be l
cover its performance thoroughly.

1.3.4 Classification of observations

Finally, as noted earlier, when the data a
(ii), one may be interested in first testing w
contaminants, and if so, how many and w
This basic framework supports much of t
tests.

The supposition underlying the desire to
that the contaminants are not an unmi
effects on the sample are to be eliminate
interest to identify the observations fro
taminating distributions, and hence ma
generating mechanism.

It is common knowledge that a number
astronomy have been made as a result o
residuals from fitted orbits. While the a
important effect that has shown up in th
opposed to a systematic lack of fit) it is nev
that this could occur. In such a case, the e
(a) testing for the presence of outliers in
locating and identifying the outliers, an
explanation for them. This procedure is
procedure set out above.

1.4 Slippage tests

A topic related quite closely to outlier the
usual model for this is that one is given o
 $i = 1$ to n . All are independently distrib
distribution $f_i(\cdot)$. The slippage hypothesis
the $f_i(\cdot)$ are identical to some common di
number of them have large probabilities in
are said to have 'slipped'.

When one inspects a sample generated b
tends to find clusters of outliers, these outl
the slipped population or populations. O

IDENTIFICATION OF OUTLIERS

Observations are believed to share a common parameter case, then it is no longer optimal to discard them in the manner in which they are incorporated into the model based on the specific distributions in the model. In fact it is of interest to note that many studies of estimators for contaminated samples have shown that there is common between the two distributions.

One of a procedure for identifying outliers is to test the model assumptions. For example, if one wishes to use a parametric analysis assuming normality of data, then one can apply the data to ensure that the model fits the test could well be an outlier test, and such a test can detect the tail behaviour of the true distribution. This may be a parametric analysis is sensitive to departures from normality of the distribution, but not elsewhere. This is why techniques (for example analysis of variance) where departures from normality are the most damaging departures from

giving a warning' implies, this use of an outlier test is made in much the same way as a test of homogeneity of variance in an analysis of variance regarded as a screen - if the data pass, then the outlier procedure will be applied; otherwise some action will be taken. The range of actions includes, but is not restricted to, removing any outliers from the sample and the analysis proposed originally on the basis of the observations.

An outlier test may be used to test for departures from normality: since the presence of outlier is indicative of a departure from the model assumed for the data, a goodness-of-fit test, in principle, be used to test for the presence of outliers. Goodness-of-fit tests which are very effective for testing for outliers in a normal sample are the moment tests and kurtosis, and both in fact are optimal in a

One of goodness-of-fit tests is that based on the chi-square test. Provided all parameters are known (or estimated from a sufficiently large sample) the probability of a test may be used to bring this to a standard form.

INTRODUCTION

7

The Anderson-Darling test, which is especially sensitive to departures from model in the tails of the distribution, has great intuitive appeal as an outlier test statistic. Its use has not, apparently, been proposed in the literature, but it is to be hoped that future work will cover its performance thoroughly.

1.3.4 Classification of observations

Finally, as noted earlier, when the data are generated by mechanism (ii), one may be interested in first testing whether any observations are contaminants, and if so, how many and which observations they are. This basic framework supports much of the theory of optimal outlier tests.

The supposition underlying the desire to classify the observations is that the contaminants are not an unmitigated nuisance whose ill effects on the sample are to be eliminated somehow; rather it is of interest to identify the observations from the basic and the contaminating distributions, and hence make deductions about the generating mechanism.

It is common knowledge that a number of important discoveries in astronomy have been made as a result of a careful study of sets of residuals from fitted orbits. While the author is not aware of any important effect that has shown up in the form of a few outliers (as opposed to a systematic lack of fit) it is nevertheless quite conceivable that this could occur. In such a case, the effect would be detected by: (a) testing for the presence of outliers in the sample, and then (b) locating and identifying the outliers, and searching for a possible explanation for them. This procedure is precisely the classification procedure set out above.

1.4 Slippage tests

A topic related quite closely to outlier theory is slippage testing. The usual model for this is that one is given observations $X_{ij}, j = 1 \text{ to } m_i, i = 1 \text{ to } n$. All are independently distributed, X_{ij} coming from a distribution $f_i(\cdot)$. The slippage hypothesis is that while the majority of the $f_i(\cdot)$ are identical to some common distribution $f(\cdot)$, some small number of them have large probabilities in the tail regions of $f(\cdot)$, and are said to have 'slipped'.

When one inspects a sample generated by a slippage hypothesis, one tends to find clusters of outliers, these outliers being the samples from the slipped population or populations. Of course if all m_i equal 1, then

IDENTIFICATION OF OUTLIERS

It becomes formally identical to an outlier slippage problem may be regarded quite usefully as the outlier problem, and it is no surprise to learn that the theory is common to both. It is often a good one to use in an analysis of a large number of possible treatments are being tested, where some small number are markedly more extreme. In such a case, the use of a slippage test will give more power than the usual F -test of analysis of variance. The consequence of the use of a relatively restricted test, and may be quite substantial (see Chapter 5). The similarity between the use of a slippage test here and the outlier problem. The latter problem, starting from the case where the $f_i(\cdot)$ are identical, seeks to isolate the 'best' observation, the stochastically largest $f_i(\cdot)$.

Significance levels

It is, data analysts undoubtedly had much more observations they were processing and probably the displays of data and grouping of data for the statistics. In the process of doing this, they were often of whether the sample seemed to contain many. It is also probably true that no sample, which appeared to be good, would ever be subjected to

Box (1976) point this fact out, and suggest as a notion of the significance level of a formal test, since the test is unlikely ever to be applied to an outlier obviously present.

In addition, they present the results of a study in which subjects were presented with samples containing outliers of varying magnitude, and were asked to decide whether or not they contained outliers. The study showed clearly that this decision was influenced by such irrelevant factors as the scale of the outliers and the value of ancillary statistics.

It was made to this study that the subjects were influenced in statistical analysis. It is this author's contention that statisticians tend to detect outliers that are not significant, and the non-significance of outlier test statistics as a

INTRODUCTION

9

reflection on the poor power of the test rather than an indication that the suspicious-looking observation is statistically quite plausible.

A degree of support for the view that statisticians are quite good at locating outliers visually is provided by Relles and Rogers (1977). They provided experienced statisticians with random samples from several heavy-tailed t -distributions and asked them to provide estimates of location. The results showed that rejecting the observations considered by the statisticians to be outliers, and finding the mean of the remainder, produced quite good estimators of location. Somewhat counter to this is the observation that rejecting a further observation on each side yields a slightly better estimator.

A more pressing problem with regard to outliers, however, is that now being felt with computerized data bases. While the statistician plotting data values on graph paper may be quite good at finding outliers, the numbers in large data bases are seldom displayed, and so outliers run a real risk of remaining undetected because unsuspected. For this reason, fully automated outlier detection procedures should be an integral part of any large data base maintenance system, and all computer-based files of data should be checked for outliers before use. In this way, one can hope to trap anomalous cases, as well as recording blunders, and so ensure the integrity of the data base.

If this procedure is followed then it is meaningful to discuss the significance levels of an outlier test, since all samples are tested for outliers, and not just those which almost certainly contain outliers.

Further, if we believe that the average data analyst who would use an outlier test at all, will apply one whenever a sample appears at all suspicious, then once again, the concept of the significance level is meaningful as a statement applying to all sets of data seen by that analyst. This is because the samples which are not screened at all, would not yield a significance if they were tested for outliers.

1.6 Brief early history of outlier rejection

Surveys of the 18th- and 19th-century literature of statistics have shown that awareness of the outlier problem arose very early. The genesis of research into outlier rejection is bound up with that of the method of least squares, and the very extensive history of least squares given by Harter (1974, 1975, 1976) contains a full discussion of early writings on outlier rejection. Another account of the literature up to 1931 is to be found in Rider (1933), and so it is appropriate to outline only a few major discoveries and concepts, up to the major paper by Pearson and Chandra Sekar (1936).

IDENTIFICATION OF OUTLIERS

ice of rejecting observations which appeared to
er data has so much intuitive appeal that it was
rmally, but merely came into use, possibly
s of osmosis. In 1778, we find Daniel Bernoulli
rably on the practice (already widespread
s of his time) of discarding any values that
nd behaving as if the remaining observations
sample.

f outlier rejection, no formal statistical criterion
on of outliers was made on the basis of the
whether the sample values were mutually
t results of Relles and Rogers suggest that they
er job in this than they are generally credited

out whether or not one should reject outliers
e, and in fact remains an issue two centuries
jective nature of outlier rejection was ame-
rice's proposal of an objective criterion for the
is criterion is based on the model that a mixture
unknown p generated the data. By making a
on p and supposing that the underlying basic
with σ known, he sets up what is essentially a
ny residuals exceeding $c\sigma$ in absolute value are
and as follows:

he cumulative distribution function of the
tribution. Let s and s' denote the standard
ample, and the sample stripped of N outliers,
ervations retained. Then c solves the equation

$$\frac{1}{2}(c^2 - 1)\{2\Phi(-c)\}^n = [N^N n^{n'} / n^n]^{n/N} \quad (1.4)$$

ised theoretical and practical objections to this
et (1866) proposed a different one. Noting that
of observations in a sample of size n from
in absolute value is $n\Phi(-c)$, he proposed that
ng $c\sigma$ in absolute value be rejected, where c

$$n\Phi(-c) = 0.5 \quad (1.5)$$

tion is so set as to reject, on average, half an
ata per sample, regardless of n .
evenet's test is especially significant, since it

INTRODUCTION

11

focuses on the experimentwise rejection rate, and not on the proportion of data values rejected. As n increases, so does c , and so the proportion of observations rejected decreases. It is easy by hindsight to see in this idea the kernel of the concept of simultaneous statistical inference, and indeed the test is not far from one with a preset experimentwise significance level.

By contrast, Stone (1868) proposed that any observer has a probability of $1/m$ of making a blunder on any particular observation, and proposed that c be chosen by

$$m\Phi(-c) = 0.5 \quad (1.6)$$

The apparent similarity between this and Chauvenet's rule is deceptive. Stone's rule rejects a fixed proportion of good observations, and so the number of outliers increases in proportion to the sample size.

A variant approach due to Irwin (1925) provided the basis for a single test for several outliers in the sample. Considering again a set of data distributed as $N(\xi^2, \sigma^2)$, σ known, Irwin considered the order statistics $X_{(1)} < X_{(2)} < \dots < X_{(n)}$, and derived the distribution of the k th gap $(X_{(n-k+1)} - X_{(n-k)})/\sigma$ as a rather complicated infinite series. He also derived computationally tractable expressions for the moments of this k th gap, and showed that for $k=1$ and 2, the distribution of the gap is well approximated by normal distributions whose means and variances he tabulated for several values of n . From this basic idea has sprung the modern work on gap tests.

All these criteria referred to the situation of observations distributed as $N(\xi, \sigma^2)$ with both ξ and σ known. In practice, of course, $\bar{X}$ was substituted for ξ , and s for σ , but the effect of this studentization depended on theory not yet extant.

Goodwin (1913) proposed that instead of the criterion $(X_i - \bar{X})/s$ used by earlier authors, one should substitute for $\bar{X}$ and s , the mean and standard deviation obtained omitting the outlier from the sample. Some years were to pass before it was shown that this test statistic is a monotonic function of $(X_i - \bar{X})/s$. The latter discovery by Thompson (1935) led to his derivation of the null distribution of an arbitrary studentized residual $(X_i - \bar{X})/s$. From this, he was able to deduce an outlier screening procedure that would reject a fixed proportion of all good data values and to produce a table of suitable critical values.

The final paper discussed in this short history is the 1936 paper of Pearson and Chandra Sekar. Using Thompson's distribution and some mathematical lemmas on extrema of the set of studentized

residuals $(X_i - \bar{X})/s$, they showed that for sufficiently large c , the event $|X_i - \bar{X}|/s > c$ implies that for all $j \neq i$, $|X_j - \bar{X}|/s < c$. Thus if c is chosen to be the $\alpha/2n$ fractile of the distribution of $(X_i - \bar{X})/s$, then an outlier test rejecting any X_i satisfying $|X_i - \bar{X}|/s > c$ has an experimentwise significance level of α . Thus, in contrast to Thompson's work, which was in the spirit of Stone, theirs was in the spirit of Chauvenet.

This extremely significant paper also pointed out the existence of what has come to be termed the 'masking effect'. This effect arises when a sample contains more than one outlier. These outliers so increase the spread of the sample that the removal of a single outlier makes little improvement in the appearance of the sample, and in particular, all values of $(X_i - \bar{X})/s$ are near zero because of the very large value of s . The practical consequences of the masking effect is that any attempt to remove these outliers one at a time will prove fruitless, and so there is a need for more sophisticated methods that will detect all outliers.

Despite the shortness of this list of papers, they provide the basic ideas from which the subsequent work is derived. The details of this subsequent development will be sketched further in the chapters on various specific aspects of, and problems in outlier theory.

General theoretical

2.1 Measures of performance

In Chapter 1, a distinction was made between the primary concern is to estimate one of the parameters of an underlying distribution, and that in which the data are divided into 'good' data and outliers (if any).

Concentrating on the second of these questions, the question of how effective an outlier test is, Paulson (1965) list five measures of performance for an outlier. Two of these are not applicable in the present context and will not be considered here. The remaining three are:

- (i) β_1 , the probability of a correct decision that the test concludes correctly that the outlier is identified correctly.
- (ii) β_2 , the power function. This is the probability that the test concludes that there is an outlier when there is one, whether or not it identifies the right outlier. The difference between β_1 and β_2 is that β_1 is the probability that one of the central values in the sample is an outlier observation may be good. In this case, a statistic would warn us correctly that there is an outlier; however, we acted on it and rejected the null hypothesis, then we would make an error (a type I error in terminology).

These two measures are the prime measures of performance. The third measure is:

- (iii) β_3 , the conditional power – that is, the probability that the test is significant given that the contaminant is present.

Clearly,

$$\beta_3 = \beta_1 / \Pr(\text{contaminant present})$$

The appeal of this measure lies in reasoning that if the test is significant, then the contaminant is present.