THE UNIVERSITY OF HULL

Some Problems Related to the Rejection of Outlying Observations

being a Thesis submitted for the Degree of

Doctor of Philosophy

in the University of Hull

by

Nicholas Rodney James Fieller, M.A.(Cambridge), M.Sc.(Birmingham)

July 1976.

TEXT CUT OFF IN THE ORIGINAL

Some Problems Related to the Rejection of

Outlying Observations

FOREWORD

The general problems in terminology and methodology encountered in testing for outliers are discussed in Chapter 1.

Chapter 2 is concerned with outliers in gamma samples. The null distribution of the statistic $T_{(n)}$, used for testing for a single upper outlier, is a known result; the result, given in section 2.2, that it is the likelihood-ratio statistic for an appropriate alternative hypothesis and the method of derivation of the distribution given in section 2.2.1 are new. All the results given in sections 2.3 to 2.7 are new. These concern the derivations and distributions of likelihood-based criteria, for testing for single and multiple lower outliers, for multiple upper outliers, and also for testing simultaneously the largest and smallest observations in the sample. The null distributions of the 'Dixon' criteria considered in section 2.8 are known results in the particular case of exponential parent populations; the derivations of the distributions and the extensions to a more general gamma parent population are new results. In section 2.9 some of the results are applied to practical examples.

Chapter 3 is concerned with single outliers in univariate normal samples. The various cases of known and unknown mean and variance are considered in sections 3.2 to 3.5. The results that the various criteria are likelihood-based for certain appropriate alternative hypotheses are new. The null distributions of these criteria and of their extensions to criteria incorporating external estimates of the variance are known, but the methods of derivation presented using recursive procedures, are new. Tests for outliers in normal samples with known mean and unknown variance (section 3.4)

have not previously been considered. In section 3.6 'two-sided' test criteria are considered. The distinction drawn between one-sided and two-sided criteria in terms of the alternative hypothesis is new, the method of obtaining upper and lower bounds for the percentage points is a known result. An error in a paper by Tietjen and Moore (1972) is corrected.

Chapter 4 is concerned with multiple outliers in univariate normal samples. A general discussion of the problems of testing for multiple outliers is given and the phenomenon of "swamping" is identified. An error in a paper by Tietjen and Moore is identified. Section 4.1 considers multiple upper outliers. The derivation of the test criteria, in the various cases of known and unknown mean and variance, as likelihood-ratio statistics for certain appropriate alternative hypotheses is new. The null distributions of these criteria for the cases of two upper outliers when the population variance is known and unknown (and the mean is unknown) are known; the derivations of these distributions presented is new, and the extensions to criteria incorporating independent estimates of the variance and to criteria for a general number of upper outliers is also new. A minor error of sign in a paper by Quesenberry and David (1961) is corrected. Section 4.2 considers tests for the largest and smallest observations simultaneously as outlying. All the results in this section relating to the criteria shewn to be likelihood-based for appropriate alternative hypotheses are new. In section 4.2.1 known results relating to the 'internally' studentized range are extended to cases when an external estimate of the population variance may be incorporated in the statistic. In section 4.3 tests for outliers at unspecified ends of the sample are considered. In section 4.3.1 it is shewn that in the case when the population mean is unknown, the procedures and criteria proposed and investigated empirically by Tietjen and Moore (1972) have undesirable

properties, and that these criteria are essentially different from the criteria derived as likelihood-ratio statistics for appropriate alternative hypotheses. The results of section 4.3.2 relating to tests for outliers in samples from a population with known mean are new.

Chapter 5 considers outliers in linear models. The derivation of the statistic $T_{(n)}$ as a likelihood-ratio criterion given in section 5.1 is new. The main results of section 5.2 relating to upper bounds on $T_{(n-1)}$ are new; it is shewn that these may be used to derive the rather weaker results of Srikantan (1961) and Stefansky (1971 and 1972). The calculation of the upper bounds for the percentage points of $T_{(n)}$, given in Table 5.1, was performed independently of Lund (1975) who gives tables of an equivalent quantity for a smaller range of significance levels but for a larger range of sample sizes. The other results of section 5.3 are new. Section 5.4 is concerned with outliers in polynomial regression. The results relating to linear regression are extensions to a larger range of significance levels of equivalent results of Srikantan (1961); the results relating to quadratic and cubic regression are new. Numerous important errors in a paper by Tietjen, Moore and Beckman (1973) are identified and corrected. All the results of sections 5.6 and 5.7 concerning multiple outliers and tests for outliers incorporating independent estimates of the variance are new. In section 5.7 applications of the results to practical examples are given.

Chapter 6 considers outliers in multivariate normal samples. Some of the general problems of detecting outliers in multivariate data are discussed in section 6.1, and some deficiencies in a paper by Rohlf (1975) are discussed.

In sections 6.2 and 6.3 the derivation of the test statistics for single and multiple outliers in all the various cases of known and unknown mean and variance as likelihood-ratio criteria is new. The results

concerning single and multiple outliers in samples with unknown mean and variance are known, but in the case of a single outlier (section 6.7.4) they are derived by a simpler method. All the results concerning single and multiple outliers in samples where either the mean or the variance or both are known, given in sections 6.2.1 to 6.2.3 and the early paragraphs of 6.3, are new. All the results of section 6.4 concerning criteria incorporating independent estimates of the variance are new. The results of section 6.5, which provide a new interpretation of the 'one-outlier scatter ratio', of Wilks (1963), are new.

Summary of Thesis for Ph.D. degree

by N.R.J. Fieller

on

Some Problems Related to the Rejection of Outlying Observations

The thesis consists of six chapters. The introductory first chapter considers some of the more general problems involved in the detection and rejection of outlying observations, and describes the general form of the tests discussed in detail in the later chapters.

In Chapter 2 likelihood-based criteria are derived for testing for single and multiple outliers at both the upper and the lower ends of samples from gamma distributions. The null distributions of these criteria are obtained by use of a recursive algorithm and the methods are extended to criteria appropriate for testing for multiple outliers occurring at both ends of the sample and to various 'Dixon' criteria. The results are applied to some practical examples.

In Chapter 3 likelihood-based tests and criteria for single outliers in univariate normal samples are considered. The null distributions of the criteria are obtained by recursive algorithms. The cases of known and unknown mean and variance are considered separately and the methods are extended to cases where independent estimates of the variance are available. These methods and results are extended in Chapter 4 to tests and criteria for multiple outliers in univariate normal samples. The extensions of the results of both of these chapters to single and multiple outliers in multivariate normal samples are considered in Chapter 6.

In Chapter 5 problems of single and multiple outliers in data following a linear model are discussed. A likelihood-based criterion is derived and the extreme tail of the null distribution of this criterion is obtained. Some practical examples on data from a series of chemical experiments are given.

Preface

I wish to thank Professor W.C.E. Higginson and Dr M.H. Hutchinson of the Department of Chemistry in the University of Hull for the data and the many helpful and illuminating discussions relating to the examples discussed in Chapter 5.

I particularly wish to express my sincere thanks to Professor T. Lewis of the Department of Mathematical Statistics in the University of Hull, for his inspiring help and guidance throughout the course of this work.

N.R.J.F.

Department of Probability and Statistics
The University of Sheffield
July 1976.

Contents

Chapter 1

The Problems of Outlying Observations

1.1 Introduction

It has long been recognised that the occurrence of spurious observations in sets of recorded and collected data is an inevitable hazard in most, if not all, statistical investigations.

One of the earliest statements which contains in essence the two basic aspects of the problems encountered in such situations is that by the astronomer Benjamin Peirce. In 1852 he wrote, "In almost every true series of observations, some are found, which differ so much from the others as to indicate some abnormal source of error not contemplated in the theoretical discussions, and the introduction of which into the investigations can only serve, in the present state of science, to perplex and mislead the inquirer".

The first aspect which may be identified here is the detection of spurious observations. Sets of experimental data may, and frequently do, contain some observations which are not of the same statistical population as the overall majority of the sets. Whether the population of primary interest in the investigation is that of the overall majority or that of the aberrant observations, it is important to detect and identify the aberrant observations in the set, either with a view to correcting them if possible (and appropriate), or to excluding them entirely from the analysis of the bulk of the data, or to focussing attention upon abnormal values. Indeed the latter may be the sole purpose of the experiment.

The second aspect is the accommodation of spurious observations. If the population of primary interest is that of the overall majority of the data set then the inclusion of any spurious observations in an analysis may lead to erroneous inferences being made, unless, of course, a method of analysis is employed which affords protection against the presence of aberrant observations; since 1852 "the state of science" has developed and advanced and now includes a great many techniques and procedures which do offer such protection.

The particular aspect discussed in detail in the later chapters is the first of these; various criteria and methods will be derived and developed for the detection of outliers in a wide variety of statistical situations. The following sections of this chapter consider some of the more general problems associated with the occurrence of spurious observations in sets of data, problems whose consideration is of importance not only when the primary aim is their detection but also when the aim is their accommodation .

1.2 Preliminary Considerations

Many different terms have been used by different authors in the literature to refer to spurious observations; as well as the adjectives 'doubtful', 'suspicious', 'surprising', 'discordant', 'unrepresentative', 'aberrant', 'wild', 'rogue' and 'spurious' and the nouns 'stragglers' and 'mavericks', there are the more commonly used terms 'outliers' and 'outlying observations', terms which have been in use certainly since the beginning of this century. See for example Pearson (1902). There is a lack of consistency not only in terminology but also in interpretation; while some authors may use a given term to refer to an observation which in actuality arises from a distribution different from that of the remainder of the sample, others may use the very same term to refer to an observation which merely appears, subjectively to the analyst, to deviate from the majority parent population. This inconsistency is not a mere matter of semantics; it reveals a deeper distinction in the lines of approach to the problems of spurious observations (here the adjective 'spurious' is taken to have its common usage meaning of "not proceeding from the pretended source"). Some authors, such as Ferguson, specifically stipulate that the decision of whether or not to apply statistical tests and procedures for the presence of spurious observations should not be made in the light of the data in question, and that any tests should be regarded as "part of the data screening process for every set of data which the experimenter may encounter". (Ferguson (1961a)). Other authors are less clear on the subject and some even imply that the various techniques will only

be applied when the sample contains observations which appear to the experimenter to be spurious just on subjective grounds. For example Guttman and Smith (1969) say, "The problem of how to deal with data which contain 'outliers', i.e. observations which look suspicious in some way, has long been a source of concern to experimenters and data analysts". Indeed not even each individual author is entirely consistent on the subject. For example Ferguson (1961) referring to the problem of the rejection of outlying observations says, "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". This would imply a certain degree of subjectivity on the part of the analyst as to whether or not to scrutinise the 'surprising' observations by employing a statistical test or procedure, in apparent contradiction to his (1961a) statement quoted above.

The position taken throughout the following chapters is that the various tests and procedures discussed should be regarded implicitly as part of the routine analysis to be performed upon all data sets. Of course it has to be recognised that many analysts will not in practice apply these tests in complete detail in cases where it is 'clear' or 'obvious' that the sets of data concerned contain no spurious observations. This is no novelty in statistical methodology; many statistical analysts would not actually pursue the detailed calculations necessary to perform a 't-test' for the equality of normal population means in cases where it is 'clear' or 'obvious' that no statistical difference exists, 'clear' or 'obvious' that is on the basis of the extensive statistical experience of the analyst. If on the other hand the various techniques and methods advocated in the following chapters were to be regarded as applicable only to those data

sets which appeared, subjectively to the experimenters, to contain spurious observations then the interpretation of the results of any statistical tests performed would have to be treated with extreme caution. Collett and Lewis (1976) shew that in such cases the conventional frequency interpretation of both the type I and type II errors are invalid and that the correct interpretations depend upon the subjective judgement of the analysts concerned. They shew further, from an experiment involving undergraduate and postgraduate students, that these subjective judgements differ not only from analyst to analyst but also from occasion to occasion for the same analyst, and that the decision of whether or not to apply a test depends upon both the particular scale and configuration and the method of presentation of the data.

Throughout the following chapters, in conformity with the position taken above, the terms 'outlier' and 'outlying observation' will be reserved to denote those observations, if any, in a sample which arise from a population other than that of the rest of the sample. The phrases 'suspected outlier', 'possible outlier' and the like will refer to those observations which are to be tested as outlying (or equivalently tested for 'discordancy'), by a statistical test; that is they are the 'extreme' observations in the sample as defined by some objective criterion. The definitions of the 'extreme' observations in the various statistical situations considered in the later chapters are given as the occasion arises. In many situations these definitions conform to the intuitive notions of 'extremeness'. Certainly this is so for single suspected outliers in univariate samples from unimodel parent populations where the 'extreme' observation is either the maximum or the minimum in the sample. In cases of

suspected multiple outliers or in cases of data following a linear model or from a multivariate population the definitions of the 'extreme' observations may not necessarily conform with intuitive considerations. For example in Chapter 5 it is shewn that the observation which must be regarded as most 'extreme' in data following a simple linear regression on an independent variable is not necessarily that observation lying furthest from the 'fitted' line. It may be noted in this context that considerable complications would arise if observations in such sets of data were to be tested as outlying or discordant only if they were thought to be 'surprising' by the particular experimenter concerned. Intuitively the 'most surprising' observation would usually be that lying furthest from the 'fitted' line; calculation of the true type I and type II errors of a test of this observation as outlying (using say, the test criterion discussed in Chapter 5) would be extremely difficult as allowance would have to be made for the possibility that this observation was not the most extreme as judged by an objective criterion. Similar considerations apply to most complex situations; in multivariate samples, for example, it is not even at all clear which is the 'most surprising' observation as judged intuitively. The only consistent procedure for avoiding these difficulties is to regard tests for the presence of outliers in data as part of the routine analysis to be applied to all data sets.

If, on the other hand, the aim of the statistical analysis of the sample of data is not primarily the detection of outliers but instead is say the estimation of some population parameter, so that it is desired to 'accommodate' any outliers in the sample by using a 'robust' method of estimation which affords protection against their presence,

then similar considerations to those outlined above would suggest that the analyst should decide in advance of the data (and not because the data does or does not contain 'surprising' or 'suspicious' observations) whether or not to apply such 'robust' procedures.

1.3 The Form of Tests for Outliers

All the various tests for outliers considered in detail in the later chapters have the form of a hypothesis test. The null hypothesis is always that all the observations in the data set form a random sample of some specified distribution (such as gamma, normal or multivariate normal). This is tested against an alternative hypothesis of the form that all but a 'small' number of the sample arise from that same specified distribution, the remaining 'small' number of observations arising from a different distribution (or distributions), typically this second distribution is of the same family as the first but with a change in scale or location. There are several issues in this formulation which need emphasis and clarification.

It must be noted firstly that it is crucial to specify the null distribution of the sample. An observation which is judged to be an outlier in relation to the rest of the sample on the null assumption of normality might not be so judged if the null assumption were that the sample was from a distribution with 'fatter tails', such as a Cauchy distribution. Or again if it is assumed that the observations arise from a normal distribution with known mean and variance then outliers might be discovered which would not be detected without the assumption of those known values of the mean and variance. Indeed it is known that certain families of distributions, such as the gamma and log-normal families, are what has been termed "outlier-prone" (Neyman and Scott (1971) and Green (1974)), that is samples from distributions in these families are 'more likely' (in a strict

probabilistic sense) to contain observations which are well separated from the bulk of the observations than samples from distributions in families which are "outlier-resistant" (such as the normal and, surprisingly, the Cauchy families).

It may be noted that there have been some attempts to devise tests for outliers which are 'distribution-free' or 'non-parametric', that is tests which require the minimum of assumptions on the particular form of the parent distribution of the majority of the sample. Typically the only assumption made is that the parent distribution is symmetric. Such tests have been considered by Walsh (1950, 1958) in the case of univariate samples. However these tests have great disadvantages since, as he indicates, they are rather insensitive to the presence of outliers unless the sample sizes are large and the number of suspected outliers is greater than four or five, all of which are assumed to occur at the same end of the sample. For the detection of outliers in designed experiments, tests have been derived by Bross (1961) and Brown (1975) based upon the 'pattern' of signs of the residuals, tests which again require a minimal assumption of a symmetric parent distribution. Presumably these latter tests are restricted to designs which have the property that the residuals have a common variance, though this is not made clear; the particular examples considered by both authors were two-way classification designs. A rather different form of non-parametric test has been discussed by Thompson and Wilke (1963), again for a two-way classification design, which is based entirely upon ranking the data in each of the 'columns' and summing these ranks across the 'rows', the purpose being to detect an 'outlying row' rather than an outlying

individual observation. Although these various distribution-free tests for outliers have the advantage of requiring only minimal assumptions about the form of the parent distribution, the inevitable loss of power entailed severely restricts their application. Further, it would seem difficult to define the term 'outlier' (at least in a form similar to that given in the preceding section) unless the form of the distribution of remainder of the sample is known with some precision.

A second crucial point in the formulation of tests for outliers given above is that the number of observations tested as outlying, k say, should be 'small' in relation to the total sample size, n say. If k is not 'small' in relation to n then the test becomes one of the adequacy of the original model rather than one for outliers, a point which will be returned to below. Exactly how 'small' is difficult to say; it would clearly be unreasonable to declare more than half the sample as outlying, so k should certainly be less than $\frac{1}{2}n$, and probably less than a rather smaller fraction of n, (maybe k should be less than a fractional power of n?). Further the number of observations accepted as genuine (i.e. not tested as outlying) clearly should be sufficient to estimate all the unknown parameters in the model, a possible restriction in samples from multivariate normal distributions or in data following a linear model. Naturally the omission from the sample of those observations tested as outlying should not render any parameters inestimable.

Most of the tests discussed in detail in the later chapters require the number of observations in the sample tested as outlying to be stipulated in advance. It is well known (see for example

Pearson and Chandra Sekar (1936)) that a test designed for a single outlier may fail to detect any outliers at all if in fact the samples contain two or more outliers. That is, the presence of a second outlier may mask the presence of the first. Complementally a single outlier may deviate so greatly from the rest of the sample as to 'swamp' other observations, that is a test on a pair of observations as outlying might result in the erroneous declaration of both observations as outlying when in fact only one is an outlier. The phenomena of masking and 'swamping' are returned to in Chapter 4. It can easily be seen, in view of these two phenomena, that the decision of how many outliers to test can be critical. In some situations external considerations might determine the number of observations to be tested as outlying. For example there might be some knowledge of the probability of occurrence of outliers for the particular experiment concerned, or there may be knowledge that some specific number of observations are in doubt because of some known temporary aberration in the experimental procedure at an unknown stage in the course of the experiment. In many cases however the number of outliers to be expected will not be known, and some preliminary analysis on the data has to be performed. This will affect the calculation of the probabilities of the type I and, type II errors. Some work has been done on this problem by Daniel (1959), Dempster and Rosner (1972) and Rosner (1975), though note that the work of Rosner (1975) suffers from the same defect as that of Tietjen and Moore (1972), discussed in detail in Section 4.3.1, in not necessarily identifying the 'most extreme' set of observations in the sample as defined by a likelihood-based criterion. These and related points are discussed more fully in Chapter 4.

The third essential point of note in the formulation of outlier tests, given at the beginning of this section, is the distinction between outlier tests and tests on the overall validity of the model. When a test for outliers in normal samples of size n, for example, is performed the hypothesis that all observations arise from a common normal distribution $N(\mu,\sigma^2)$ is tested against the alternative hypothesis, that a 'small' number, k say, of the sample arise from a different distribution, but the remaining n-k observations (the majority of the sample) arise from the null distribution $N(\mu,\sigma^2)$. The essential point is that the distribution of the overall majority of the sample is not in doubt; the only question is whether or not a 'small' proportion of the sample deviates from this distribution. On the other hand a test of the overall validity of the model would test the same null hypothesis against the alternative that the entire sample arises from a different distribution. There is a possibility of confusion between these two forms of test, particularly in the case of normal samples, because certain criteria which have been proposed for testing for outliers are also commonly used as overall tests of normality. For example the coefficients of skewness and kurtosis, which have been shewn by Ferguson (1961) to have certain locally optimal properties when used to test for outliers, are familiar as criteria used for testing for normality. Again the studentized range (whether 'externally' or 'internally' studentized) considered in Chapter 4 as a test statistic for the detection of a pair of outliers (one each of the pair occurring at each end of the sample) was originally studied as a 'short-cut' test for normality, (see for example, Pillai (1952), David, Hartley & Pearson (1954),

Pearson and Stephens (1964) and the introduction to Biometrika Tables for Statisticians Volume 1, p. 59). These, and various other criteria which may be applied to the problem of the detection of outliers, are discussed in the context of testing for normality by Shapiro and Wilk (1965) and Shapiro, Wilk and Chen (1965). Clearly there is some danger, when testing for outliers, of committing an error of the third kind (i.e. correctly rejecting the null hypothesis, but for the wrong reason, Kendall and Buckland (1957)). It can only be assumed, when testing for outliers, that an extreme value of a test statistic indeed indicates the presence of an outlying observation, rather than a departure of the entire sample from the null distribution, when there are a priori reasons for believing in the validity of the null distribution, at least for the great majority of the sample.

Considerations similar to those commented upon above apply when the particular aspect of the problem of the occurrence of outliers is their accommodation rather than their detection. Just as there are tests for outliers and tests for normality (or more generally the overall validity of the model) which employ the same test criteria, so there are robust procedures of estimation which offer protection not only against the presence of outliers but also against departures from normality (or whatever the null distribution of the sample may be). Typically these departures are assumed to be in the direction of the distribution being fatter tailed. An example of such a procedure when estimating the mean of a univariate population would be 'data trimming', i.e. invariably discarding both maximum and minimum observations in the sample before calculating the estimate, the

principle being that the discarding of genuine observations will affect the estimate of the mean very much less than the inclusion of spurious ones. Further, just as there are test criteria (such as the studentized extreme deviates from the sample discussed in the later chapters) which are designed specifically for the detection of outliers (though they could presumably be used as tests on the overall validity of the model) so there are methods of estimation which are designed specifically to be robust against the presence of outliers. Many of these procedures involve first the detection of any outliers, using one or other of the many criteria designed for that purpose, and then discarding any outliers detected and estimating the population parameters on the reduced sample. Other procedures, robust against the presence of outliers, which do not involve their rejection as such, include 'Winsorization' (replacing the extreme observations by values equal to the second-most extremes) and various Bayesian procedures which essentially give little weight in the estimation procedure to the extreme observations. Various methods of estimating of the mean and variance in normal samples, robust against the presence of outliers, are discussed fully by Anscombe (1960 and 1961), Guttman and Smith (1969 and 1971) and Gebhardt (1964 and 1966). In the case of the estimation of parameters in an exponential population the literature is equally extensive and includes that of Veale and Hutsberger (1969), and Mount and Kale (1973) as well as the numerous papers discussed in the following chapter. Approaching the problems of estimation in the presence of outliers from an essentially Bayesian viewpoint is the work of de Finetti (1961), Tiao and Guttman (1967), Box and Tiao (1968), Guttman (1973) ,and

Guttman and Khatri (1975).

The final issue of import arising from the formulation of the test of outliers given above is that the particular observation (or observations) singled out as 'extreme' and tested as outlying depends upon the form of the alternative hypothesis. For example, in the case of normal samples, if the alternative hypothesis is that any outlier arises from a population with a larger mean than that of the rest of the sample then the 'extreme' observation is clearly the maximum of the sample. Correspondingly the 'extreme' observation is the minimum of the sample if the alternative hypothesis is that any outlying observation arises from a population with smaller mean than that of the remainder of the sample. If, however, the alternative is 'two-sided' in the sense that it is specified only that the aberrant observation arises from a population with a different mean then the 'extreme' observation is the maximum or minimum of the sample, whichever is more distant from the mean. This important point, first emphasised by King (1953), will be discussed further in the following chapters.

1.4 Probabilistic Models for Outliers

The formulation of outlier tests given in the previous section presupposes a specific probabilistic model for the occurrence of outliers, a model which is contained in the statement of the alternative hypothesis. It is assumed that a sample of size n containing k outliers (where k is assumed to be known) may be described probabilistically by saying that the sample is composed of n-k observations arising from one distribution, F, say, and k observation arising from some different distribution G, say, (or distributions G_i). This may be described as a contamination model; that is the sample of observations from the distribution F is contaminated by the inclusion of observations from the distribution G.

An intrinsically different form of model for 'outliers' which may be termed a 'mixture' model, has been proposed by some authors, for example by Elashoff (1972) in the case of outliers in linear regression. This model does not require the number of outliers in the sample to be known. In its most general form it describes a sample containing outliers as one where there is a constant probability of (1-p) that any particular observation in the sample arises from a distribution F say, and a probability of p_i that the observation arises from a distribution G_i, where $p=\sum p_i$ and where usually it is assumed that p is small. This is taken as the alternative to the null hypothesis that all observations in the sample arise from the distribution F.

This second model has some considerable intuitive appeal since it reflects the 'generation' of any outliers in the sample; it assumes that there is a small probability p that any particular observation in

the sample is spurious, i.e. that it arises from one of the distributions G_i rather than the null distribution F.

It is easily seen, however, that the assumption of a 'mixture' model for outliers is equivalent to testing the null hypothesis that the set of observations is a random sample from the distribution F against the alternative that it is a random sample from the 'mixed' distribution $(1-p)F+\sum p_i G_i$: that is the assumption of a 'mixture' model for outliers implies that the test is not so much one for outliers, in the sense described in the previous section, as one on the overall validity of the model. The essential distinction between the 'contamination' and 'mixture' models is that the former merely supposes that the sample contains a specified number of spurious observations while the latter involves the construction of a probabilistic mechanism to describe the actual generation of the spurious observations.

The model assumed throughout the following chapters is the 'contamination' model. In Chapter 2 the distribution F is taken to be a gamma distribution $\Gamma(\lambda,r)$ where the degrees of freedom parameter r is assumed known. In Chapters 3 and 4 it is taken to be a normal distribution $N(\mu,\sigma^2)$, the four distinct cases involving μ and σ^2 known and unknown are considered separately. In Chapter 5 the distribution F is taken to be $N(X\beta,\sigma^2)$ where X is a known matrix and β is an unknown vector, and in Chapter 6 it is taken to be the multivariate normal distribution $N(\mu,\Lambda)$, where again the four distinct cases of μ and Λ known and unknown are considered separately. In the derivations of the various likelihood ratio test criteria the contaminating distribution F is taken to be of the same family as F but with a change in scale or location, or both. In cases where the distribution G is of the same

family as F but with a change in location some of the tests for outliers become particular cases of the so-called 'slippage tests' discussed in the cases of univariate normal samples by Paulson (1952) and more generally in the cases of multivariate normal samples by Karlin and Truax (1960).

It may appear somewhat unrealistic to assume that any outliers in a sample arise from the same form of distribution as the rest of the sample but with a change in location or scale. If the outliers reflect gross errors in the actual recording or copying of data, such as the interchanging of a pair of adjacent digits, or even some quite arbitrary misrecording, then indeed it may not be possible to regard the outliers as arising from the same distribution, except for a shift in location or a change in scale, as the rest of the sample. If however the recording error is say the omission or addition of a terminal zero, or if the outlier reflects the measurement of a unit from a set similar to, though distinct from, the rest of the experimental units, then the outliers may well be regarded as arising from a distribution shifted in location or scale from that of the rest of the sample. Only some external knowledge of the likely causes of any outliers in the data can settle this question. Some examples are discussed in Chapter 5 where a likely cause of outliers was identified which indicated that outliers in the data could be regarded as arising from a normal distribution with a shift in mean, of unknown magnitude, but with the same variance, as the rest of the sample. Another interesting example of this kind has been discussed by Finney (1974), and relates to a study performed by his former student B.K. Thompson, now of the University of Toronto. The data were the weights of a very large

number of chickens (measured in kilograms), data which could be assumed to be normally distributed. The chickens had been weighed on balance scales at regular intervals in their growth by an assistant. It was recognised that the sheer numbers of chickens weighed at each session meant that the assistant was liable to misrecord the weights of some birds and further it was recognised that the most likely form of error was a miscounting of the individual weights on the balance, so that it was possible that the recorded weight of a chicken could be in error by 50, 100, 500 or 1000 grams. It would therefore be entirely reasonable to suppose that any outliers in the data were observations from a normal distribution with a shift in mean (in fact a shift of precisely 50, 100, 500 or 1000 grams) from that of the rest of the sample. It is of interest to note (Thompson (1973)) that this particular set of data was first 'screened' for the presence of recording errors by screening separately the data for each 'family' of chickens (i.e. chickens raised from the same clutch of eggs) for outliers, using as test criterion the studentized extreme deviation from the mean. However this approach failed to detect any outliers in the data and it was not until the sequences of weights of individual chickens were examined, as discussed by Finney (1974), that the recording errors were detected.

This example illustrates a further feature of the difficulties involved in the detection of outliers. A wide range of test criteria can only detect an outlier when it occurs as the 'extreme' of the sample, that is if the value of the test criterion used (in the above example the maximum studentized deviation from the sample mean) achieves its most extreme value for the outlying observation. In samples from a unimodal population this means that outliers can only be detected if they occur as either the maximum or the minimum of the sample. The

reason that outliers were not detected by the first screening, family by family, of Thompson's data was that the intra-family variation was so large that frequently the observations later detected as outlying did not occur as the extremes of their particular families; in cases where they did they were not sufficiently separated from the rest of the sample to allow their declaration as outlying. It was only by examining a sequence of numbers, representing the weights of a chicken at regular intervals in the course of its growth, such as

1·20, 1·60, 1·90, 1·55 2·20 2·25

that the observation 1·55 is readily identified as outlying, particularly when it is remembered that outliers are likely to be in error by 0·5kg. Formal tests of outliers in such time series data are discussed by Fox (1972).

Finally it must be noted that it is only possible to investigate the power of tests for outliers if the distribution from which they arise is specified. Studies of the power of various outliers tests have been made by a number of the authors referred to in later chapters and also, in a paper specifically devoted to the subject, by David and Paulson (1965).

1.5 The Aims and Results of Outlier Detection

Two basic reasons for scrutinizing data for the presence of outliers may be distinguished. The first is to 'clean up' or 'launder' the data before performing any further analysis; the second is that any outliers in the data may be of intrinsic interest, representing, in the context of variety trials say, the 'high yielding' variety.

Historically it was the first of these that was the prime motivation in the development of the many techniques for the detection and rejection of outlying observations. Many of these techniques were developed by experimental scientists (and in particular by astronomers) to handle the spurious observations occurring in their own sets of experimental data. Rider (1933) gives a comprehensive account of this early work, starting from the experimentally motivated work of Peirce (1852), quoted earlier, on a series of measurements of the diameter of Venus to the more statistically based work of Tippett (1925) and Irwin (1925) discussed in Chapter 4. It was only later that similar techniques and methods were applied to samples where the outliers were themselves of intrinsic interest, such as the problems in harmonic analysis considered by Fisher (1929) and in certain geometrical problems considered by Stevens (1939), which are discussed in the following Chapter 2.

The decision of how to proceed when an outlier has been detected in a sample involves not only consideration of the reason for examining the sample for outliers, but also consideration of the aims of the entire analysis. If the sample has been scrutinised for the

presence of outliers with a view to 'cleaning up' the data then the outliers can only be 'rejected' if the very fact of their occurrence is immaterial to the purpose of analysis. To take the original astronomical example considered by Peirce (1852), as illustration; if the aim of the analysis is the estimation of the true diameter of Venus, then the inclusion of any outliers in this estimate can only, as he says so succinctly, "perplex and mislead the inquirer". If, however, the aim of the experiment were an investigation of the reliability of the astronomical equipment used to make the measurements, then the very fact that an outlier had occurred would be of great importance.

Sometimes the decision of whether or not to reject the outliers can only be made after consideration not only of the purpose of the analysis but also of the likely causes of the outliers. Kruskal (1960) discusses a hypothetical example relating to the accuracy with which bombs are dropped on a target. He supposes that a few of the bombs are very wide of the target and that it was observed that the fins on these wild bombs came loose in flight. He points out that if the purpose of the analysis is to investigate the accuracy of the bomb sight then these outliers are irrelevant, while if the aim is to investigate the accuracy of the bombing system as a whole then the occurence of outliers is of vital interest.

The essential point of the example of Kruskal's bombs is that it would only be possible to discount entirely the outliers from the assessment of the bombsight accuracy with the knowledge that the fins had come loose on the bombs involved. If this were not known then the outliers might have had to have been taken to reflect inaccuracies in the bombsighting mechanism. In the example on the measurements of the

diameter of Venus, however, it would not be necessary to know the causes of any outliers in the data before excluding them from the final estimate of the Venusian diameter.

Of course there can be no invariable rule, applicable to all situations, for dealing with outliers once they have been detected. The examples discussed above illustrate that the decision of whether to 'reject' or 'retain' them (or indeed 'correct' them, as would perhaps be possible with Thompson's data discussed in the previous section) can only be made in the light of the particular experimental situation involved.

Chapter 2

Outliers in Gamma Samples

The problems associated with the occurrence of outliers in samples from exponential parent populations have received much attention in recent years. The impetus for this work lies in practical situations, as is indeed the case with the study of all outlier problems. The particular context within much of this work has been concentrated, is that of life testing, that is testing batches of manufactured items (in particular electronic components), whose times to failure may be considered to be exponential random variables. In such situations the experimenter may be confronted with items which survive for a surprisingly long time or which fail suspiciously quickly. The experimenter will then be interested in whether the 'abnormal' items are truly representative of the batch of items under test or whether there is any evidence to indicate that the 'abnormal' items are defective in some way. The statistical problem, given the observed times to failure of the items, is then to decide whether the extreme observations in the sample are consistent with the rest of the data or whether there is evidence that they may be outliers.

Of course life testing is by no means the only situation which gives rise to samples of observations of an exponential random variable. Indeed some of the early distributional results of importance for the detection of outliers in exponential samples were obtained within entirely different contexts. Fisher (1929) considered the problem of testing the largest amplitude in a harmonic analysis, Stevens (1939)

and Fisher (1940) investigated a problem in geometrical probability, and Cochran (1941) examined the problem of testing the homogeneity of a set of estimated variances. Although none of these studies was directed to the specific problem of outlier detection, all developed methods and results which may be used to test as outlying the largest observation in an exponential, or chi-square, or more generally gamma sample. In a later section two examples are discussed one of which involves outliers in a sample of excess cycle times in steel manufacture and the other of which concerns the validation of a pseudo-random number generator in a simulation program. In the next two chapters it is shewn that the problem of detecting outliers in samples from a normal distribution with known mean may be reduced to that of detecting outliers in gamma samples.

While the detection of outliers in exponential, and more generally gamma, samples is applicable to a wide variety of practical situations it is in the field of life testing that it has received perhaps the greatest attention. This is clearly shewn in the work of Epstein (1960(a),(b)), Laurent (1963), Likeš (1966) and Kabe (1970), all of whom consider tests for a single outlier in such situations. Darling (1952), who considers the distribution, under general conditions, of a statistic which in the particular case of a gamma distribution is appropriate for testing for an upper outlier (the statistic $T_{(n)}$ of section 2.2), refers to applications in the broader area of quality control. Other work, following a rather different approach to the problem of outliers, though still within the context of life testing is that of Basu (1968), Joshi (1972), Kale (1974,1975), Kale and Sinha (1971), Sinha (1972,1973(a),(b),(c)), Veale (1975), and Veale and Kale (1972). This work is directed more to the problems of

estimation in the presence of outliers rather than to direct tests for their detection.

The results of this chapter apply to the problems of detection and not directly to estimation. The null distributions of likelihood-based test criteria for single and multiple outliers occurring at either the upper or the lower end of a gamma sample are obtained using a recursive procedure. The methods are extended to consider other criteria for testing for a single outlier and to a criterion appropriate to testing simultaneously the largest and smallest observations in a gamma sample. The first section establishes the terminology and notation.

2.1 Preliminaries

Let $x_1,\ldots,x_n$ be a sample from a gamma population with density function

$$\frac{\lambda^r}{\Gamma(r)} x^{r-1} \exp(-\lambda x) \qquad (x > 0) \qquad (2.1)$$

where r is known but λ is unknown. Denote the ordered sample by $x_{(1)},\ldots,x_{(n)}$ where $x_{(1)} < \ldots < x_{(n)}$. Let $S_n = \sum_{i=1}^{n} x_i$, $\bar{x}_n = S_n/n$ and $T_{n,j} = x_j/S_n$. Also write $T_{n,(j)} = x_{(j)}/S_n$ and in particular $T_{n,(n)} = \max_{j=1}^{n} (T_{n,j}) = x_{(n)}/S_n$; in cases where the sample size is clearly n, $\bar{x}_n$, $T_{n,n}$, $T_{n,(j)}$ and $T_{n,(n)}$ will be abbreviated to $\bar{x}$, T_n, $T_{(j)}$ and $T_{(n)}$ respectively. For any j, $T_{n,j}$ follows a beta distribution with parameters r and r(n-1), i.e. its density function is $\beta_{r,r(n-1)}(.)$ where

$$\beta_{a,b}(u) = \{\Gamma(a+b)/\Gamma(a)\Gamma(b)\}\, u^{a-1}(1-u)^{b-1} \qquad (0 < u < 1) \qquad (2.2)$$

(Here, and elsewhere, a density function whose value is specified only over certain regions is understood to be otherwise zero.)

Two classes of criteria for testing for outliers will be considered, firstly 'studentized' statistics of the form $T_{(n)}$ and secondly those of the form

$$y = \frac{x_{(s)} - x_{(r)}}{x_{(q)} - x_{(p)}} \qquad (1 \leq p \leq r < s \leq q \leq n,\ q - p > s - r) \qquad (2.3)$$

where, for example, the choice $s = q = n$, $r = n - 1$, $p = 1$ is made when testing whether or not the largest observation may be an outlier. Statistics of both forms possess two essential qualifications for consideration as outlier detecting criteria; firstly their value is sensitive to the presence of outliers and secondly they are independent of the scale of measurement used.

2.2 The Statistic $T_{(n)}$

It is shewn first that statistics of the form $T_{(n)}$ may be regarded as likelihood-based in the particular case when the alternative to the null hypothesis, that all members of the sample come from the same gamma population, is that all but one observations come from that gamma population and the remaining one arises from a gamma population with a smaller scale parameter.

Let H_o be the hypothesis that $x_1,\ldots,x_n$ come from a gamma population with density (2.1) and let H_1 be the hypothesis that $x_1,\ldots,x_{n-1}$ come from that population and that one observation, x_n without loss of generality, arises from a population with density function

$$\frac{\mu^r}{\Gamma(r)}\ x^{r-1}\ \exp\ (-\mu x), \qquad (x > 0),$$

where $\mu < \lambda$.

Under H_o the log-likelihood is

$$nr\log\lambda - n\lambda\bar{x}_n + K(x_1,\ldots,x_n)$$

$$(\text{where } K(x_1,\ldots,x_n) = (r-1)\sum_1^n \log x_i - n\log(\Gamma(r)))$$

which is a maximum when $\lambda = r/\bar{x}_n$ giving a maximised log-likelihood under H_o of

$$nr\log(r/\bar{x}_n) - nr + K(x_1,\ldots,x_n).$$

Under H_1 the log-likelihood is

$$(n-1)r\log\lambda-(n-1)\lambda\bar{x}_{n-1}+r\log\mu-\mu x_n+K(x_1,\ldots,x_n)$$

which is a maximum when

$\lambda = r/\bar{x}_{n-1}$ and $\mu = r/x_n$, giving a maximised log-likelihood under H_1

of

$$(n-1)r\log(r/\bar{x}_{n-1})+r\log(r/x_n)-nr+K(x_1,\ldots,x_n).$$

The difference between the maximised log-likelihoods under H_o and H_1 is thus

$$nr\log(\bar{x}_n)-(n-1)r\log(\bar{x}_{n-1})-r\log(x_n),$$

which may be written as

$$nr\ \log\left[\frac{n-1}{n(1-T_n)}\right]+r\ \log\left[\frac{1-T_n}{(n-1)T_n}\right].$$

Thus it is seen that the statistic $T_{(n)}$ has the important property that it is essentially the increase in the maximised log-likelihood consequent upon acceptance of the alternative hypothesis that the queried observation arises from a different distribution.

2.2.1 A Recursive Algorithm for the Distribution of $T_{(n)}$

Suppose $T_{(n)}$ has density function $a_n(.)$ and distribution function $A_n(.)$. Then

$$a_n(u)du = P\,[T_{(n)} \in (u,u+du)]$$

$$= \sum_{j=1}^{n} P\,[T_{n,j} \in (u,u+du), T_{n,(n)} = T_{n,j}]$$

$$= nP\,[T_{n,n} \in (u,u+du), \max_{k=1}^{n-1}\{x_k/(S_n-x_n)\} < x_n/(S_n-x_n)]$$

$$= nP\,[T_{n,n} \in (u,u+du), \max_{k=1}^{n-1}\{x_k/(S_n-x_n)\} < u/(1-u)]$$

$$= nP\,[T_{n,n} \in (u,u+du)]\;\; P\,[\max_{k=1}^{n-1}\{x_k/(S_n-x_n)\} < u/(1-u)]$$

(noting that $T_{n,n} = x_n/S_n$ and $x_k/(S_n-x_n)$ are independent for each $k=1,2,\ldots,n-1$, since the x_i follow a gamma distribution).
Thus

$$a_n(u)du = nP\,[T_{n,n} \in (u,u+du)]\;\; P\,[T_{n-1,(n-1)} < u/(1-u)]\ ,$$

which gives the recurrence relation

$$a_n(u) = n\beta_{r,r(n-1)}(u)A_{n-1}\{u/(1-u)\}\ . \qquad (2.4)$$

Consider first the particular case of an exponential parent population, $(r=1)$. If $\frac{1}{2} \leq u < 1$ then $u/(1-u) \geq 1$, so $A_{n-1}(u/(1-u)) = 1$ and hence from (2.4)

$$a_n(u) = n\beta_{1,n-1}(u) = n(n-1)(1-u)^{n-2},$$

$$A_n(u) = 1-n(1-u)^{n-1}. \qquad (2.5)$$

If $\frac{1}{3} \leq u \leq \frac{1}{2}$ then $\frac{1}{2} \leq u/(1-u) \leq 1$ and so, from (2.5), $A_{n-1}(u/(1-u)) = 1-(n-1)\{(1-2u)/(1-u)\}^{n-2}$, and then, from (2.4),

$$a_n(u) = n(n-1)(1-u)^{n-2}-n(n-1)^2(1-2u)^{n-2},$$

$$A_n(u) = 1-\binom{n}{1}(1-u)^{n-1}+\binom{n}{2}(1-2u)^{n-1};$$

and in general for $1/(q+1) < u \leq \frac{1}{q}$, where $q = \left[\frac{1}{u}\right]$,

$$a_n(u) = n(n-1(1-u)^{n-2}-n(n-1)^2(1-2u)^{n-2}+ \ldots$$

$$-(-)^q \frac{n!(n-1)}{(q-1)!(n-q)!}(1-qu)^{n-2}$$

$$= n(n-1)\sum_{i=0}^{q-1}(-)^{i+1}\binom{n-1}{i}\{1-(i+1)u\}^{n-2},$$

$$A_n(u) = 1-\binom{n}{1}(1-u)^{n-1}+\binom{n}{2}(1-2u)^{n-1}- \ldots +(-)^q\binom{n}{q}(1-qu)^{n-1}$$

$$= 1+\sum_{i=1}^{q}(-)^i\binom{n}{i}(1-iu)^{n-1}. \qquad (2.6)$$

This is of course the well-known result of Fisher (1929,1940), and which has also been obtained by Darling (1952) who used a method employing characteristic functions.

For general r, the same procedure can be applied, using the

recursive formula (2.4) to evaluate the functions $a_n(u), A_n(u)$ in successive intervals $1/(q+1) < u \leq \frac{1}{q}$ $(q=1,\ldots,n-1)$. In particular, since

$$A_{n-1}(u/(1-u)) = 1 \qquad \text{if } u \geq \tfrac{1}{2}$$

and $$A_{n-1}(u/(1-u)) < 1 \qquad \text{if } u < \tfrac{1}{2}$$

it follows that $a_n(u) = n\beta_{r,r(n-1)}(u)$ if $u \geq \frac{1}{2}$

and $$a_n(u) < n\beta_{r,r(n-1)}(u) \text{ if } u < \tfrac{1}{2}$$

so $$P[T_{(n)} > u] = 1-A_n(u) = n\int_u^1 \beta_{r,r(n-1)}(t)dt \quad \text{if } u \geq \tfrac{1}{2}$$

and $$P[T_{(n)} > u] < n\int_u^1 \beta_{r,r(n-1)}(t)dt \qquad \text{if } u < \tfrac{1}{2}.$$

Thus, if r is a positive integer, the upper tail probability $P[T_{(n)} > u]$, which can be used as a significance probability for testing the largest observation as an outlier, satisfies

$$P[T_{(n)} > u] = nP[F_{2r,2r(n-1)} > (n-1)u/(1-u)], \quad \text{if } u \geq \tfrac{1}{2},$$

$$P[T_{(n)} > u] < nP[F_{2r,2r(n-1)} > (n-1)u/(1-u)], \quad \text{if } u < \tfrac{1}{2}, \tag{2.7}$$

where $F_{2r,2r(n-1)}$ denotes a variate following an F-distribution with $(2r,2r(n-1))$ degrees of freedom.

2.3 The Statistic $T_{(1)}$

The statistic $T_{n,(1)}$ is suitable for testing the smallest observation, $x_{(1)}$ as an outlier in a gamma sample of size n; not only is its value deflated if the observation $x_{(1)}$ is an outlier but it may also be regarded as likelihood-based in the sense of section 2.2, where now the alternative hypothesis is that one observation arises from a gamma distribution with a larger scale parameter.

Suppose $T_{n,(1)}$ has density function $b_n(.)$ and distribution function $B_n(.)$. Then, using an argument similar to that of section 2.2.1, we have the recurrence relation

$$b_n(u) = n\beta_{r,r(n-1)}(u) \quad [1-B_{n-1}(u/(1-u))] , \qquad (0 \leq u \leq 1/n). \tag{2.8}$$

Since $B_1(u) = 0$ or 1 according as $u < 1$ or $u \geq 1$ (2.8) gives

$$b_2(u) = 2\beta_{r,r}(u)$$

$$= 2 \frac{\Gamma(2r)}{(\Gamma(r))^2} u^{r-1}(1-u)^{r-1}, \qquad (0 \leq u \leq \tfrac{1}{2}), \tag{2.9}$$

ans successive use of (2.8) gives $b_3(u)$, $b_4(u),\ldots,b_n(u)$.

In the case of an exponential parent population, r=1,

$$b_2(u) = 2, \qquad (0 \leq u \leq \tfrac{1}{2}),$$

and (2.8) becomes

$$b_n(u) = n(n-1)(1-u)^{n-2}\,[1-B_{n-1}(u/(1-u))] \quad (0 \leqslant u \leqslant 1/n)$$

whence $$b_n(u) = n(n-1)(1-nu)^{n-2} \quad (0 \leqslant u \leqslant 1/n). \qquad (2.10)$$

In the case r=2,

$$b_2(u) = 12u(1-u) \quad (0 \leqslant u \leqslant \tfrac{1}{2}),$$

and the recurrence relation (2.8) is

$$b_n(u) = n(2n-1)(2n-2)u(1-u)^{2n-3}\,[1-B_{n-1}(u/(1-u)], \quad (0 \leqslant u \leqslant 1/n)$$

whence recursively

$$b_3(u) = 60u(1-3u)(1-3u^2) \quad (0 \leqslant u \leqslant 1/3)$$

$$b_4(u) = 168u(1-4u)^2(1+3u-12u^2-4u^3) \quad (0 \leqslant u \leqslant 1/4)$$

$$b_5(u) = 360u(1-5u)^3(1+8u-18u^2-80u^3+65u^4) \quad (0 \leqslant u \leqslant 1/5)$$

$$b_6(u) = 660u(1-6u)^4(1+15u-360u^3+864u^5) \quad (0 \leqslant u \leqslant 1/6)$$

$$b_7(u) = 1092u(1-7u)^5(1+24u+75u^2-920u^3-2265u^4+7728u^5+3997u^6)$$
$$(0 \leqslant u \leqslant 1/7) \qquad (2.11)$$

and so on.

In the case r=3,

$$b_2(u) = 60u^2(1-u)^2, \quad (0 < u < \tfrac{1}{2}),$$

and the recurrence relation (2.8) is

$$b_n(u) = \tfrac{1}{2}(3n-1)(3n-2)(3n-3)u^2(1-u)^{3n-4}\,[1-B_{n-1}(u/(1-u)]\ , \quad (0 \leqslant u \leqslant 1/n),$$

so, proceeding recursively

$$b_3(u) = 504u^2(1-3u)(1-2u+4u^2-18u^3+21u^4) \qquad (0 \leqslant u \leqslant 1/3)$$

$$b_4(u) = 1980u^2(1-4u)^2(1-8u+28u^2-224u^3+1540u^4-5266u^5+11032u^6 - 16832u^7+13696u^8) \quad (0 \leqslant u \leqslant 1/4) \quad (2.12)$$

and so on. As r and n increase the polynomial expressions for $b_n(u)$ become increasingly difficult to compute. However the lower tail probability $P\,[T_{(1)} < u]$, which, as in section 2.2.1, may be used as a significance probability for testing the smallest observation as an outlier, satisfies the inequality

$$P\,[T_{(1)} < u] \ < nP\,[F_{2r,2r(n-1)} < (n-1)u/(1-u)]\ , \qquad (u \leqslant 1/n),$$

or equivalently

$$P\,[T_{(1)} < u] < nP\,[F_{2r(n-1),2r} > (1-u)/(n-1)u]\ , \qquad (u \leqslant 1/n). \quad (2.13)$$

2.4 The Joint Distribution of $T_{(1)}$, $T_{(n)}$

In order to consider the null distribution of certain statistics appropriate for testing the largest and smallest observations simultaneously as outliers, it is necessary to find the joint distribution of $T_{(1)}$, $T_{(n)}$, under the null hypothesis that x_1, $x_2,\ldots,x_n$ are all observations from the same gamma population.

Suppose this distribution has joint density function $c_n(.,.)$ and distribution function $C_n(.,.)$. Since $0 \leq T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(n)} \leq 1$ and since $\sum_{j=1}^{n} T_{(j)} = 1$ it follows that

$$(n-1)T_{(1)} + T_{(n)} \leq 1$$

and that

$$T_{(1)} + (n-1)T_{(n)} \geq 1.$$

Thus $c_n(u,v)$ is zero outside the region R_c defined by

$$u \geq 0$$
$$(n-1)u+v \leq 1$$
$$u+(n-1)v \geq 1.$$

For $(u,v) \in R_c$

$$c_n(u,v)\, du\, dv = P\,[T_{(1)} \in (u,u+du), T_{(n)} \in (v,v+dv)]$$

$$= \sum_{\substack{i=1 \\ i\neq j}}^{n} \sum_{j=1}^{n} P\,[T_{n,i} \in (u.u+du), T_{n,j} \in (v,v+dv), T_{n,(1)} = T_{n,i},\ T_{n,(n)} = T_{n,j}]$$

$$= n(n-1)\ P\ [T_{n,n-1} \in (u,u+du),\ T_{n,n} \in (v,v+dv),$$
$$T_{n,(1)} = T_{n,n-1},\ T_{n,(n)} = T_{n,n}]$$

$$= n(n-1)\ P\ [T_{n,n-1} \in (u,u+du),\ T_{n,n} \in (v,v+dv),$$
$$x_{n-1} < x_1 < x_2 < \ldots < x_{n-2} < x_n]$$

$$= n(n-1)\ P\ [T_{n,n-1} \in (u,u+du),\ T_{n,n} \in (v,v+dv),$$
$$u/(1-u-v) < T_{n-2,(1)} < T_{n-2,(2)} < v/(1-u-v)]$$

$$= n(n-1)\ P\ [T_{n,n-1} \in (u,u+du),\ T_{n,n} \in (v,v+dv)]\ \times$$
$$\times\ P\ [u/(1-u-v) < T_{n-2,(1)} < T_{n-2,(n-2)} < v/(1-u-v)] \qquad (2.14)$$

(noting that both $T_{n,n-1} = x_{n-1}/S_n$ and $T_{n,n} = x_n/S_n$ are independent, for each $k=1,2,\ldots,(n-2)$, of $T_{n-2,k} = x_k/(S_n-x_n-x_{n-1})$ since the x_i follow a gamma distribution).

The first of the two probabilities in (2.14) is $\theta_n(u,v)du\ dv$, where $\theta_n(u,v)$ is the value of the joint density of $T_{n,n-1}$ and $T_{n,n}$ at the point (u,v). Now the joint density of $(x_1,x_2,\ldots,x_n)$ is

$$\prod_{i=1}^{n}\{\lambda^r x_i^{r-1}\exp(-\lambda x_i)/\Gamma(r)\}\ \ dx_1dx_2,\ldots,dx_n.$$

The Jacobean of the transformation

$$s_n = \textstyle\sum_1^n x_i,\quad t_2 = x_2/\sum_1^n x_i,\ldots,t_n = x_n/\sum_1^n x_i$$

is s_n^{n-1} so the joint density of $(S_n, T_2, T_2,\ldots,T_n)$ is

$$\lambda^{nr}\{\Gamma(r)\}^{-n}\{\prod_{j=2}^{n}(t_j^{r-1})\}(1-\sum_{2}^{n}t_j)^{r-1}s_n^{nr-1}\exp(-\lambda s_n)ds_n dt_2\ldots dt_n, \qquad (2.15)$$

which, on integrating out the variables $S_n, T_2, \ldots, T_{n-2}$ gives value of the joint density of T_{n-1}, T_n at the point (u,v) as

$$\Gamma(rn)/\{(\Gamma(r))^2\Gamma(r(n-2))\}u^{r-1}v^{r-1}(1-u-v)^{r(n-2)-1}. \qquad (2.16)$$

Thus (2.14) with (2.16) gives

$$c_n(u,v) = n(n-1)\Gamma(rn)/\{(\Gamma(r))^2\Gamma(r(n-2))\}u^{r-1}v^{r-1}(1-u-v)^{r(n-2)-1}\Psi$$

where $\Psi = P\,[u/(1-u-v) < T_{n-2,(1)} < T_{n-2,(n-2)} < v/(1-u-v)]$. (2.17)

When $n \geq 5$, (2.17) can be written

$$\Psi = P[T_{n-2,(n-2)} < v/(1-u-v)] - P\,[T_{n-2,(1)} < u/(1-u-v),\ T_{n-2,(n-2)} < v/(1-u-v)]$$

$$= A_{n-2}\{v/(1-u-v)\} - C_{n-2}\{u/(1-u-v),\ v/(1-u-v)\}.$$

Hence for $n \geq 5$

$$c_n(u,v) = n(n-1)\Gamma(rn)/\{(\Gamma(r))^2\Gamma(r(n-2))\}u^{r-1}v^{r-1}(1-u-v)^{r(n-2)-1} \times$$
$$\times\ [A_{n-2}\{v/(1-u-v)\} - C_{n-2}\{u/(1-u-v),\ v/(1-u-v)\}]. \qquad (2.18)$$

When n=3, the probability Ψ in (2.17) is 1 if $u/(1-u-v) < 1$ and

and $v/(1-u-v) > 1$, and is 0 otherwise; hence

$$c_3(u,v) = 3!\{\Gamma(3r)/(\Gamma(r))^3\}u^{r-1}v^{r-1}(1-u-v)^{r-1},$$

$$(2u+v < 1,\ u+2v > 1,\ u > 0). \quad (2.19)$$

When $n=4$, $T_{2,(1)}+T_{2,(2)} = 1$ and so

$$\Psi = P\,[T_{2,(1)} > \max\{1-(v/(1-u-v)),u/(1-u-v)\}]$$

$$= 2\,\frac{\Gamma(2r)}{(\Gamma(r))^2}\int_w^{\frac{1}{2}} t^{r-1}(1-t)^{r-1}dt \quad \text{(using (2.9))}, \quad (2.20)$$

where $w = (1-u-2v)/(1-u-v)$ for $2u+2v < 1$, $(u,v) \in R_c$

$u/(1-u-v)$ for $2u+2v > 1$, $(u,v) \in R_c$;

hence, in the cases where r is a positive integer,

$$c_4(u,v) = 4!\{\Gamma(4r)/(\Gamma(r))^4\}u^{r-1}v^{r-1}2^{-2r+1} \times$$
$$\times \sum_{k=0}^{r-1} \frac{(-)^k}{2k+1}\binom{r-1}{k}(1-u-v)^{2(r-1-k)}z^{2k+1} \quad (2.21)$$

where $z = u+3v-1$ for $2u+2v < 1$, $u+3v > 1$, $u > 0$

$z = 1+3u-v$ for $2u+2v > 1$, $3u+v < 1$, $u < 0$.

Recursive use of the formula (2.18) now gives expressions for $c_n(u,v)$ for odd n, using (2.19) initially, and for even n, using (2.21) initially.

In the particular case r=1, (the exponential case), this gives

$$c_3(u,v) = 1.2^2.3, \qquad (u+2v > 1,\ 2u+v < 1,\ u > 0),$$

$$c_4(u,v) = 2.3^2.4(u+3v-1), \quad (u+3v > 1,\ 2u+2v < 1,\ u > 0),$$

$$2.3^2.4\{(u+3v-1)-2(2u+2v-1)\}, \quad (2u+2v > 1,\ 3u+v < 1,\ u > 0),$$

$$c_5(u,v) = 3.4^2.5(u+4v-1)^2, \quad (u+4v > 1,\ 2u+3v < 1,\ u > 0),$$

$$3.4^2.5\{(u+4v-1)^2-3(2u+3v-1)^2\}, \quad (2u+3v > 1,\ 3u+2v < 1,\ u > 0),$$

$$3.4^2.5\{(u+4v-1)^2-3(2u+3v-1)^2+3(3u+2v-1)^2\},$$

$$(3u+2v > 1,\ 4u+u < 1,\ u > 0),$$

$$c_6(u,v) = 4.5^2.6\ (u+5v-1)^3, \quad (u+5v > 1,\ 2u+4v < 1,\ u > 0),$$

$$4.5^2.6\{(u+5v-1)^3-4(2u+4v-1)^3\}, \quad (2u+4v > 1,\ 3u+3v < 1,\ u > 0),$$

$$4.5^2.6\{(u+5v-1)^3-4(2u+4v-1)^3+6(3u+3v-1)^3\},$$

$$(3u+3v > 1,\ 4u+2v < 1,\ u > 0),$$

$$4.5^2.6\{(u+5v-1)^3-4(2u+4v-1)^3+6(3u+3v-1)^3-6(4u+2v-1)^3\},$$

$$(4u+2v > 1,\ 5u+v < 1,\ u > 0),$$

and generally

$$c_n(u,v) = k_n(u+(n-1)v-1)^{n-3}, \quad (u+(n-1)v > 1,\ 2u+(n-2)v < 1,\ u > 0),$$

$$k_n\{(u+(n-1)v-1)^{n-3}-(n-1)(2u+(n-2)v-1)^{n-3}\},$$

$$(2u+(n-2)v > 1,\ 3u+(n-3)v < 1,\ u > 0),$$

$$\cdot \quad \cdot \quad \cdot$$

$$\cdot \quad \cdot \quad \cdot$$

$$\cdot \quad \cdot \quad \cdot$$

$$k_n\{(u+(n-1)v-1)^{n-3}-(n-2)(2u+(n-2)v-1)^{n-3}$$

$$+ \binom{n-2}{2}(3u+(n-3)v-1)^{n-3} - \ldots$$

$$+ (-1)^n \binom{n-2}{n-4}((n-3)u+3v-1)^{n-3}$$

$$-(-1)^n \binom{n-2}{n-4}((n-2)u+2v-1)^{n-3}\},$$

$$((n-2)u+2v > 1,\ (n-1)u+v < 1,\ u > 0),$$

where $k_n = (n-2)(n-1)^2 n.$ (2.22)

This may be written more concisely as

$$c_n(u,v) = (n-2)(n-1)^2 n \sum_{i=0}^{q-2} (-1)^i \binom{n-2}{i} \{(i+1)u+(n-i-1)v-1\}^{n-3} - (-1)^{q-2} \binom{n-2}{q-2} \{qu+(n-q)v-1\}^{n-3},$$

for $qu+(n-q)v > 1$, $(q+1)u+(n-q-1)v < 1$, (2.22a)

Again, in the particular case r=2 (2.19) and (2.21) become

$$c_3(u,v) = 720uv(1-u-v), \quad (u+2v > 1,\ 2u+v < 1,\ u > 0)$$

$$c_4(u,v) = 10080uv(u+3v-1)\{(1-u)^2-3v^2\}, (u+3v > 1,\ 2u+2v < 1,\ u > 0),$$
$$10080uv(1-3u-v)\{(1-v)^2-3u^2\}, (2u+2v> 1,\ 3u+v < 1,\ u > 0),$$

(2.23)

whence $c_n(u,v)$ can be obtained recursively for further values of n.

For other positive integral values of r (2.19) and (2.21) give expressions for $c_3(.,.)$ and $c_4(.,.)$ allowing the recursive evaluation of $c_n(.,.)$ by (2.18). In cases when r is non-integral, for example, when the parent population is chi-square with odd degrees of freedom, $c_3(.,.)$ is again given by (2.19) but it will not be possible to express $c_4(.,.)$ in the polynomial form of (2.21). (2.18) can nevertheless be used to derive expressions recursively for $c_n(.,.)$ from (2.19) and (2.20).

2.5 The Null Distribution of $T_{(n)}-T_{(1)}$

The joint density $c_n(.,.)$ of $T_{(1)}$, $T_{(n)}$ may be used to derive expressions for the density function, $d_n(.)$, say, of the statistic $W_{(n)} = T_{(n)}-T_{(1)}$, under the null hypothesis that all the observations are from the same gamma population. The value of the 'studentized' statistic $W_{(n)}$ will be particularly inflated if $x_{(n)}$ is large and $x_{(1)}$ is small, in relation to the values of $x_{(2)},\ldots,x_{(n-1)}$, and so is an appropriate statistic to use for testing simultaneously the largest and smallest observations in a gamma sample.

The transformation $w = v-u$, $z = (n-1)u+v$ gives the joint density of $W_{(n)} = T_{(n)}-T_{(1)}$, $Z_{(n)} = (n-1)T_{(n)}+T_{(n)}$ from $c_n(u,v)$. Integration out of the variable z then gives the density function $d_n(.)$ of $W_{(n)}$. In the exponential case, r=1, this gives

$$d_3(u) = 4u, \quad (0 \leq u \leq 1/2)$$
$$4\{u-(2u-1)\}, \quad (1/2 \leq u \leq 1)$$

$$d_4(u) = 9\{2u^2\}, \quad (0 \leq u \leq 1/3)$$
$$9\{2u^2-(3u-1)^2\}, \quad (1/3 \leq u \leq 1/2)$$
$$9\{2u^2-(3u-1)^2+2(2u-1)^2\}, \quad (1/2 \leq u \leq 1)$$

$$d_5(u) = 16\{6u^3\}, \quad (0 \leq u \leq 1/4)$$
$$16\{6u^3-(4u-1)^3\}, \quad (1/4 \leq u \leq 1/3)$$
$$16\{6u^3-(4u-1)^3+3(3u-1)^3\}, \quad (1/3 \leq u \leq 1/2)$$
$$16\{6u^3-(4u-1)^3+3(3u-1)^3-3(2u-1)^3\}, \quad (1/2 \leq u \leq 1)$$

$$d_6(u) = 25\{12u^4\}, \quad (0 \leq u \leq 1/5)$$

$$25\{12u^4-(5u-1)^4\}, \quad (1/5 \leq u \leq 1/4)$$

$$25\{12u^4-(5u-1)^4+4(4u-1)^4\}, \quad (1/4 \leq u \leq 1/3)$$

$$25\{12u^4-(5u-1)^4+4(4u-1)^4-6(3u-1)^4\}, \quad (1/3 \leq u \leq 1/2)$$

$$25\{12u^4-(5u-1)^4+4(4u-1)^4-6(3u-1)^4+4(2u-1)^4\}, \quad (1/2 \leq u \leq 1)$$

and in general

$$d_n(u) = (n-1)^2\{(n-3)(n-2)u^{n-2}-((n-1)u-1)^{n-2}+\binom{n-2}{1}((n-2)u-1)^{n-2}- \ldots$$
$$+(-1)^{n-q}\binom{n-2}{n-q-1}(qu-1)^{n-2}\}.$$
$$= (n-1)^2\{(n-3)(n-2)u^{n-2}-\sum_{i=1}^{n-q}(-1)^i\binom{n-2}{i-1}\{(n-i)u-1)\}^{n-2}\}$$

where $q = [1/u]$. (2.24)

In the case r=2 the corresponding expressions for $d_3(u)$ and $d_4(u)$ are

$$d_3(u) = 40/9u(2-5u^2), \qquad (0 \leq u \leq 1/2),$$
$$40/9(1-u)^3(1+4u), \qquad (1/2 \leq u \leq 1),$$

$$d_4(u) = 21/32u^2(90-450u^2+182u^4), \qquad (0 \leq u \leq 1/3),$$
$$21/32\{(1-u)^5(13+101u)-2(1-2u)^4(13+104u+100u^2), \qquad (1/3 \leq u \leq 1/2),$$
$$21/32(1-u)^5(13+101u), \qquad (1/2 \leq u \leq 1).$$

(2.25)

2.6 Several Upper Outliers

Consider first the case of two upper outliers. Suppose $g_n(.,.)$ is the joint density of $T_{(n-1)}$ and $T_{(n)}$; since $\sum_1^n T_{(j)} = 1$ and $0 < T_{(1)} < \ldots < T_{(n)} < 1$,

$$T_{(n-1)}+T_{(n)} < 1 \quad \text{and} \quad (n-1)T_{(n-1)}+T_{(n)} > 1.$$

Thus $g_n(u,v)$ will be zero outside the region R_g^2 defined by

$$u + v < 1,$$
$$(n-1)u + v > 1,$$
$$0 < u < v.$$

For $(u,v) \in R_g^2$

$$g_n(u,v)du\,dv = P\,[T_{(n-1)} \in (u,u+du), T_{(n)} \in (v,v+dv)]$$

$$= n(n-1)P\,[T_{n,n-1} \in (u,u+du),\ T_{n,n} \in (v,v+dv),\ T_{n,(n-1)} = T_{n,n-1}, T_{n,(n)} = T_{n,n}]$$

$$= n(n-1)\,P\,[T_{n,n-1} \in (u,u+du),\ T_{n,n} \in (v,v+dv),\ \max_{i=1}^{n-2} x_i < x_{n-1}]$$

$$= n(n-1)\,P\,[T_{n,n-1} \in (u,u+du),\ T_{n,n} \in (v,v+dv)] \times \times\ P\,[T_{n-2,(n-2)} < u/(1-u-v)]$$

(noting that both $T_{n,n-1}$ and $T_{n,n}$ are independent of $T_{n-2,(n-2)}$)

$$= n(n-1)\Gamma(rn/\{(\Gamma(r))^2\Gamma(r(n-2))\}u^{r-1}v^{r-1} \times$$
$$\times (1-u-v)^{r(n-2)r-1}A_{n-2}\{u/(1-u-v)\}dudv$$

(using (2.16)). Thus

$$g_n(u,v) = n(n-1)\Gamma(rn)/\{(\Gamma(r))^2\Gamma(r(n-2))\}u^{r-1}v^{r-1}(1-u-v)^{r(n-2)-1} \times$$
$$\times \{(n-1)\beta_{r,r(n-2)}(u/(1-v))\}^{-1}a_{n-1}(u/(1-v))$$

(using (2.4))

$$= n\Gamma(rn)\{\Gamma(r)\Gamma(r(n-1))\}^{-1}v^{r-1}(1-v)^{r(n-1)-2}a_{n-1}\{u/(1-v)\}. \tag{2.26}$$

In the exponential case (r=1) this becomes

$$g_n(u,v) = n(n-1)(1-v)^{n-3}a_{n-1}\{u/(1-v)\}$$

$$= n(n-1)^2(n-2)\{(1-u-v)^{n-3} - \binom{n-2}{1}(1-2u-v)^{n-3} + \binom{n-2}{2}(1-3u-v)^{n-3}$$
$$- \ldots + (-1)^{q-1}\binom{n-2}{q-1}(1-qu-v)^{n-3}\}$$

$$= n(n-1)^2(n-2)\sum_{i=0}^{q-1}(-1)^i\binom{n-2}{i}\{1-(i+1)u-v\}^{n-3} \tag{2.27}$$

where $q = [(1-v)/u]$.

In the case of three upper outliers, if $g_n(u,v,w)$ is the joint density of $T_{(n-2)}$, $T_{(n-3)}$, $T_{(n)}$ then an immediate extension of the method used above yields, for $(u,v,w) \in R_g^3$, where R_g^3 is the region defined by

$$u + v + w < 1,$$

$$(n-2)u + v + w > 1,$$

$$0 < u < v < w.$$

$$g_n(u,v,w) = n(n-1)(n-2)\Gamma(rn)\{(\Gamma(r))^3\Gamma(r(n-3))\}^{-1}u^{r-1}v^{r-1}w^{r-1} \times$$
$$\times (1-u-v-w)^{r(n-3)-1}A_{n-3}\{u/(1-u-v-w)\}$$

$$= n(n-1)\Gamma(rn)\{(\Gamma(r))^2\Gamma(r(n-2))\}^{-1}v^{r-1}w^{r-1}(1-v-w)^{r(n-2)-2} \times$$
$$\times a_{n-2}\{u/(1-v-w)\},$$

(using (2.4)) (2.28)

which in the exponential case becomes

$$g_n(u,v,w) = n(n-1)^2(n-2)(1-v-w)^{n-4}a_{n-2}\{u/(1-v-w)\}$$

$$= n(n-1)^2(n-2)^2(n-3)\{(1-u-v-w)^{n-4} - \binom{n-3}{1}(1-2u-v-w)^{n-4}$$
$$+\binom{n-3}{2}(1-3u-v-w)^{n-4} - \ldots$$
$$+ (-1)^{q-1}\binom{n-3}{q-1}(1-qu-v-w)^{n-4}\}$$

$$= n(n-1)^2(n-2)^2(n-3)\sum_{i=0}^{q-1}(-1)^i\binom{n-3}{i}\{1-(i+1)u-v-w\}^{n-4}$$

where $q = [(1-v-w)/u]$. (2.29)

For the general case of k upper outliers the same method gives the joint density of $T_{(n-k+1)}, T_{n-k+2)}, \ldots, T_{(n)}$ as

$$g_n(u,v,w,\ldots,z) = (n)_{k-1}\Gamma(rn)\{(\Gamma(r))^{k-1}(r(n-k+1))\}^{-1}v^{r-1}w^{r-1}\ldots z^{r-1}$$
$$(1-v-w-\ldots-z)^{r(n-k+1)-2}a_{n-k+1}\{u/(1-v-w-\ldots-z)\}$$

(where $(n)_k = n!/(n-k)! = \Gamma(n+1)/\Gamma(n-k+1)$) (2.30)

for $(u,v,w, \ldots z) \in R_g^k$, the region defined by

$$u + v + \ldots + z < 1,$$

$$(n-k+1)u+v+w+ \ldots +z > 1,$$

$$0 < u < v < \ldots < z.$$

In the particular case r=1 this gives

$$g_n(u,v,w, \ldots,z) = (n)_{k-1}(n-1)_{k-1}(1-v-w- \ldots -z)^{n-k-1}a_{n-k+1}\{u/(1-u-v-\ldots z)\}$$

$$= (n)_{k+1}(n-1)_{k-1}\{(1-u-v- \ldots -z)^{n-k-1} - \binom{n-k}{1}(1-2u-v-w- \ldots -z)^{n-k-1} + \binom{n-k}{2}(1-3u-v-w- \ldots -z)^{n-k} - \ldots + (-1)^{q-1}\binom{n-k}{q-1}(1-qu-v-w- \ldots -z)^{n-k}$$

$$= (n)_{k+1}(n-1)_{k-1}\sum_{i=0}^{q-1}(-)^i\binom{n-k}{i}\{1-(i+1)u-v-w- \ldots -z)\}^{n-k}$$

where $q = [(1-v-w- \ldots -z)/u]$. (2.31)

A useful statistic for testing for k upper outliers in a gamma sample is $T_{(n-k+1)}+T_{(n-k+2)}+ \ldots +T_{(n)} = Z_{n,(k)}$, say. This statistic is likelihood-based for the alternative hypothesis that k observations arise from a single gamma distribution with a smaller scale parameter than the other (n-k) observations. The density function of $Z_{n,(n)}$ $h_n(.)$ can, in principle, be found from the above joint density. Now since

$$h_n(u) = P\,[Z_{n,(k)} \in (u,u+du)]$$

$$= \binom{n}{k} P\,[T_{n-k+1}+T_{n-k+2}+ \ldots + T_n \in (u,u+du),$$

$$T_{n-k+1}+T_{n-k+2}+ \ldots +T_n = T_{(n-k+1)}+T_{(n-k+2)}+ \ldots +T_{(n)}]$$

$$\leq \binom{n}{k} P\,[T_{n-k+1}+ \ldots +T_{(n)} \in (u,u+du)]$$

$$= \binom{n}{k} \beta_{rk,r(n-k)}(u) \quad .$$

The following inequality therefore for the upper tail probability holds:

$$P\,[Z_{n,(k)} > u] \leq \binom{n}{k} P\,[F_{2rk,2r(n-k)} > ((n-k)/k)(u/(1-u))] \quad . \qquad (2.32)$$

2.7 Several Lower Outliers

Consider first the case of two lower outliers. Suppose $k_n(.,.)$ is the joint density of $T_{(1)}$ and $T_{(2)}$; $k_n(u,v)$ is clearly zero outside the region R_k^2 defined by

$$u + (n-1)v < 1$$
$$0 < u < v.$$

For $(u,v) \in R_k^2$

$$k_n(u,v)du\,dv = P\,[T_{(1)} \in (u,u+du),\ T_{(2)} \in (v,v+dv)]$$

$$= n(n-1)P[T_{n,1} \in (u,u+du),\ T_{n,2} \in (v,v+dv)] \times$$
$$\times P\,[T_{n-2,(1)} > v/(1-u-v)]$$

$$= n(n-1)\Gamma(rn)\{(\Gamma(r))^2\Gamma(r(n-2))\}^{-1}u^{r-1}v^{r-1}(1-u-v)^{(n-2)r-1} \times$$
$$\times \{1-B_{n-2}(v/(1-u-v))\}$$

$$= n\Gamma(rn)\{\Gamma(r)\Gamma(r(n-1))\}^{-1}u^{r-1}(1-u)^{r(n-1)-2}b_{n-1}\{v/(1-u)\}.$$
(using (2.8)) (2.33)

In the exponential case (r=1) this becomes

$$k_n(u,v) = n(n-1)(1-u)^{n-3}b_{n-1}\{v/(1-u)\}$$

$$= n(n-1)^2(n-2)\{1-u-(n-1)v\}^{n-3}\ . \qquad (2.34)$$

In the general case of m lower outliers the same method gives the joint density $k_n(u,v, \ldots, y,z)$ of $T_{(1)}, \ldots, T_{(m)}$, for $(u,v,\ldots,z) \in R_k^m$, the region defined by

$$u + v + \ldots + y + (n-m+1)z < 1,$$

$$0 < u < v < \ldots < z,$$

as

$$k_n(u,v),\ldots,z) = (n)_m \Gamma(rn)\{(\Gamma(r))^m \Gamma(r(n-m))\}^{-1} u^{r-1} v^{r-1} \ldots z^{r-1} \times$$
$$\times (1-u-v-\ldots-z)^{(n-m)r-1}\{1-B_{n-m}(z/(1-u-\ldots-z))\}$$

$$= (n)_{m-1} \Gamma(rn)\{(\Gamma(r))^{m-1} \Gamma(r(n-m+1))\}^{-1} u^{r-1} v^{r-1} \ldots y^{r-1} \times$$
$$\times (1-u-v-\ldots-y)^{(n-m+1)r-2} b_{n-m+1}\{z/(1-u-v-\ldots-y)\}, \tag{2.35}$$

which in the case r=1 gives

$$k_n(u,v,\ldots z) = (n)_{m-1}(n-1)_{n-1}(1-u-v-\ldots-y)^{n-m-1} b_{n-m+1}\{z/(1-u-v-\ldots-y)\}$$

$$= (n)_{m+1}(n-1)_{m-1}\{1-u-v-\ldots-y-(n-m+1)z\}^{n-m-1} . \tag{2.36}$$

An appropriate statistic for testing the significance of m lower outliers in a gamma sample is $T_{(1)}+T_{(2)}+\ldots+T_{(m)} = Y_{n,(m)}$, say. This statistic is likelihood-based for the alternative hypothesis that m observations arise from a gamma distribution with a larger scale parameter than the other n-k observations. The density function, $\ell_n(.)$, of $Y_{n,(m)}$ in principle can be found by transforming the above density. In the case of two lower outliers in an exponential sample this yields

$$\ell_n(t) = n(n-1)^2/(n-2)\{(1-\tfrac{1}{2}nt)^{n-2}-(1-(n-1)t)^{n-2}\} \quad (0 < t < 1/(n-1))$$

$$n(n-1)^2/(n-2)\{(1-\tfrac{1}{2}nt)^{n-2}\} \quad (1/(n-1) < t < 2/n).$$

(2.37)

Even without the exact distribution of $Y_{n,(m)}$ in the general case the method of section 2.6 gives $\ell_n(u) < \binom{n}{m}\beta_{rm,r(n-m)}(u)$, so that

$$P\,[Y_{n,(m)} < u] < \binom{n}{m} P\,[F_{2rm,2r(n-m)} < ((n-m)/m)(u/(1-u))]$$

or equivalently

$$P\,[Y_{n,(m)} < u] < \binom{n}{m} P\,[F_{2r(n-m),2rm} > m(1-u)/(u(n-m))]\ .$$

(2.38)

2.8 Other Test Criteria

The recursive methods of the preceding sections can readily be applied to statistics of the form given in equation (2.3). These were proposed by Dixon (1950, 1951) in the case of testing for outliers in samples from a normal parent population, and have been discussed in the particular context of an exponential parent population by Likeš (1966) and Kabe (1970).

Consider first the statistic

$$y = (x_{(2)}-x_{(1)})/(x_{(n)}-x_{(1)})$$

which can be used for testing the smallest observation as outlying. Let $m_n(.,.,.)$ be the joint density of $T_{(1)},T_{(2)},T_{(n)}$. $m_m(u,v,w)$ will be zero outside the region R_m defined by

$$u+(n-2)v+w < 1,$$
$$u+v+(n-2)w > 1,$$
$$0 < u < v < w.$$

For $(u,v,w) \in R_m$ and $n > 3$

$$m_n(u,v,w)\,dudvdw = (n)_3 P\,[T_1 \in (u,u+du),\ T_2 \in (v,v+dv),\ T_n \in (w,w+dw),\ T_{(1)} = T_1,\ T_{(2)} = T_2,\ T_{(n)} = T_n]$$

$$= (n)_3 P\,[T_1 \in (u,u+du),\ T_2 \in (v,v+dv),\ T_n \in (w,w+dw),\ \min_{i=3}^{n-1} x_i > x_2,\ \max_{i=3}^{n-1} < x_n]$$

$$= (n)_3\Gamma(rn)\{(\Gamma(r))^3\Gamma(r(n-3))\}^{-1}u^{r-1}v^{r-1}w^{r-1}(1-u-v-w)^{r(n-3)-1} \times$$

$$\times P\,[v/(1-u-v-w) < T_{n-3,(1)} < T_{n-3,(n-3)} < w/(1-u-v-w)]$$

$$= n\Gamma(rn)\{\Gamma(r)\Gamma(r(n-1))\}^{-1}u^{r-1}(1-u)^{r(n-1)-3}c_{n-1}\{v/(1-u),w/(1-u)\},$$

(using (2.17))

which, upon making the substitution $y = (v-w)/(w-u)$, gives the joint density function of $T_{(1)}$, y, $T_{(n)}$ as

$$q_n(u,y,w) = n\Gamma(rn)\{\Gamma(r)\Gamma(r(n-1))\}^{-1}u^{r-1}(1-n)^{r(n-1)-3}(w-u) \times$$

$$\times\ c_{n-1}\{(u+(w-u)y)/(1-u),w(1-u)\} \qquad (2.39)$$

for $(u,y,w) \in R_q$, the region defined by

$$(n-2)(u+(w-u)y) + w < 1-u,$$

$$(u + (w-u)y) + (n-2)w > 1-u.$$

$$0 < y < 1,\ u > 0.$$

The distribution of the statistic y can now be derived by integrating out the variables u and w, using the expressions for c_{n-1} given in section 2.4.

For example consider the exponential case (r=1) with n=4. Substituting for c_3 from (2.22) gives

$$q_4(u,y,w) = 4.3!/(2!).(w-u).12 = 144(w-u)$$

for

$$(3+2y)u+(1+2y)w < 1,\ (2-y)u+(2+y)w > 1,\ u > 0,\ 0 < y < 1.$$

Integrating out the variables u and w gives the density function of y as

$$6\{(1+2y)^{-2}-(2+y)^{-2}\} \qquad (0 < y < 1) \tag{2.40}$$

and, integrating, the distribution function as

$$9y/\{(2y+1)(y+2\} \ , \tag{2.41}$$

in agreement with Likeš (1966) and Kabe (1970).

As an example of a non-exponential situation consider the case r=2, n=4. From (2.39) and (2.23)

$$m_4(u,y,w) = 4.7!/(5!)u(1-u)^3(w-u).720(u+(w-u)y)w(1-2u-w-(w-u)y)(1-u)^{-3}$$

$$= 120960uw(w-u)((1-y)u+yw)(1-(2-y)u-(1+y)w),$$

$$(u > 0,\ (3-2y)u+(1+2y)w < 1,\ (2-y)u+(2+y)w > 1).$$

Integrating out the variables u and w gives the density function of y as

$$3/8\{11(1+2y)^{-2}+44(1+2y)^{-3}-69(1+2y)^{-4}+24(1+2y)^{-5}$$
$$-10(2+y)^{-2}-44(2+y)^{-3}-12(2+y)^{-4}+192(2+y)^{-5}\},\ (0 < y < 1), \tag{2.42}$$

and, integrating, the distribution function as

$$3/16\{-11(1+2y)^{-1}-22(1+2y)^{-2}+23(1+2y)^{-3}-6(1+2y)^{-4}$$
$$+20(2+y)^{-1}+44(2+y)^{-2}+8(2+y)^{-3}-96(2+y)^{-4}\} \ . \tag{2.43}$$

Consider next the Dixon type statistic

$$z = (x_{(n)}-x_{(n-1)})/(x_{(n)}-x_{(1)})$$

which can be used for testing the largest observation in the sample as outlying. Let $s_n(.,.,.)$ be the joint density of $T_{(1)}$, $T_{(n-1)}$, $T_{(n)}$. Then $s_n(u,v,w)$ will be zero outside the region R_s defined by

$$(n-2)u+v+w < 1,$$
$$u+(n-2)v+w > 1,$$
$$0 < u < v < w.$$

For $(u,v,w) \in R_s$ and $n \geq 3$ it is easily seen that the application of the method used before gives

$$s_n(u,v,w) = n\Gamma(rn)\{\Gamma(r)\Gamma(r(n-1))\}^{-1}w^{r-1}(1-w)^{r(n-1)-3} \times \\ \times c_{n-1}\{u/(1-w), v/(1-w)\},$$

which, upon making the substitution $z = (w-v)/(w-u)$, gives the joint density function of $T_{(1)}$, z, $T_{(n)}$ as

$$t_n(u,z,w) = n\Gamma(rn)\{\Gamma(r)\Gamma(r(n-1))\}^{-1}w^{r-1}(1-w)^{r(n-1)-3}(w-u) \times \\ \times c_{n-1}\{u/(1-w), (w-(w-u)z)/(1-w)\}, \qquad (2.44)$$

provided $(u,z,w) \in R_t$, the region defined by

$$(n-2)u+(w-(w-u)z) < 1-w,$$

$$u+(n-2)(w-(w-u)z) > 1-w,$$

$$0 < z < 1,\ u > 0.$$

The distribution of z can now be obtained by integrating out the variables u and w.

In the exponential case with n=4 for example,

$$t_4(u,z,w) = 144(w-u)$$

for $(2+z)u+(2-z)w < 1$, $(1+2z)u+(3-2z)w > 1$, $0 < z < 1$, $u > 0$.
Integrating out the variables u and w gives the density function of z as

$$6\{(2-z)^{-2}-(3-2z)^{-2}\} \qquad (0 < z < 1), \tag{2.45}$$

and, integrating, the distribution function as

$$z(z-4z)/\{(2-z)(3-2z)\}, \tag{2.46}$$

in agreement with Likeš (1966) and Kabe (1970).

In the case r=2, n=4, using (2.42) and (2.23)

$$t_4(u,z,w) = 120960uw(w-u)(zu+(1-z)w)(1-(1+z)u-(2-z)w)$$

$$(u > 0,\ (2+z)u+(2-z)w < 1,\ (1+2z)u+(3-2z)w > 1),$$

which, upon integrating out the variables u and w gives the density function of z as

$$3/8\{10(2-z)^{-2}+36(2-z)^{-3}-36(2-z)^{-4}$$
$$-11(3-2z)^{-2}-44(3-2z)^{-3}+69(3-2z)^{-4}-24(3-2z)^{-5}\}$$
$$(0 < z < 1) \qquad (2.47)$$

with distribution function

$$3/16\{20(2-z)^{-1}+36(2-z)^{-2}-24(2-z)^{-3}$$
$$-11(3-2z)^{-1}-22(3-2z)^{-2}+23(3-2z)^{-3}-6(3-2z)^{-4}-32/3\}. \qquad (2.48)$$

2.9 Two Examples

2.9.1 An Example in Steel Manufacture

Table 2.1 shews a sample of 132 excess cycle times in steel manufacture. The two largest observations appear surprisingly large by comparison with other 130 values.

Table 2.1

Excess cycle time	Frequency
1	18
2	12
3	18
4	16
5	10
6	4
7	9
8	9
9	2
10	7
11	6
12	7
13	2
14	1
15	3
21	3
32	2
35	1
92	1
97	1

An examination of the sample moments about the mean of the reduced sample obtained by omitting the observations $x_{(131)} = 92$ and $x_{(132)} = 97$ suggests that the distribution of excess cycle times may be assumed to be exponential. The two values 92 and 97 can be tested as outlying using the results of section 2.6. Calculation of the statistic $Z_{n,(k)}$ of that section gives

$$Z_{132,(2)} = (92+97)/1043 = 0{\cdot}812.$$

Putting n = 132, k=2, r=1 in inequality (3.23) gives

$$P\,[Z_{132,(2)} > 0{\cdot}1812] \leq \binom{132}{2} P\,[F_{4,260} > (130/2)(0{\cdot}1812/0{\cdot}8188)]$$

$$= 8646\ P\,[F_{4,260} > 14\ 385]$$

$$\approx 8646\ P\ [\chi^2_4 > 57{\cdot}54]$$

$$= (8646)29{\cdot}77)\ \exp(-28{\cdot}77)$$

$$< 10^{-7}.$$

Thus there is very strong evidence indeed for regarding the upper two extreme values 92 and 97 as outliers.

2.9.2 An Example in Simulation

In an early version of a computer program designed to simulate the distribution of stock from a warehouse to a chain of retail outlets a library pseudo-random number generator was used to construct a sample of exponentially distributed random numbers which were to represent the loading times of lorries at the warehouse. Table 2.2 gives a sample of twelve such values obtained in a trial run of the program.

Table 2.2

Loading Times

87
62
124
53
343
21
32
4
3
11
323
1067

The programmer was suspicious firstly of the high value 1067, and secondly of the pair of low observations 4 and 3.

Calculation of the statistic $T_{(n)}$ of section 2.2 gives

$$T_{(12)} = 1067/(87+62+ \ldots +1067)$$

$$= 1067/2130$$

$$= 0{\cdot}5009.$$

Equation (2.6) with n=12 gives

$$P\,[T_{(12)} > {\cdot}502] = 1-A_{12}(0{\cdot}5000)$$

$$= 12(0{\cdot}4991)^{11}$$

$$= 0{\cdot}0057.$$

There is thus strong evidence to indicate that the observation

1067 does not belong to the same distribution as the other 11 values. Armed with this knowledge the programmer re-examined his program and discovered that there was an error in the print instructions and that indeed the value 1067 represented not a loading time of a lorry but the 'current time' of the simulation.

The observation 1067 was therefore 'rejected' and an examination made of the compatibility of the lowest two observations with the rest of the sample. Calculation of the statistic $Y_{n,(m)}$ of section 2.7 gives

$$Y_{11,(2)} = (3+4)/1063$$

$$= 0{\cdot}0066.$$

Equation (2.37) with n = 11 gives

$$P\,[Y_{11,(2)} \leq 0{\cdot}0066] = \int_0^{0{\cdot}0066} \ell_{11}(t)dt$$

$$= \int_0^{0\ 0066} 1100/9\{(1-5{\cdot}5t)^9-(1-10t)^9\}dt$$

$$= 1-20/9(1-5{\cdot}5 \times 0{\cdot}0066)^{10}+11/9(1-0{\cdot}066)^{10}$$

$$= 0{\cdot}082.$$

Thus although the values 3 and 4 appear to be suspiciously low the evidence for regarding them as belonging to a different distribution from the rest of the sample is weak and it is not possible to say with any degree of certainty that either the library pseudo-random number generator or the program itself is at fault.

Chapter 3

Single Outliers in Univariate Normal Samples

The detection and the treatment of outliers in data were first studied, as were so many statistical concepts, methods and techniques of wider application, in the context of univariate normal samples. It was not until comparatively recent years that the ideas and methods were developed and extended to consider outliers in samples from other parent populations, such as samples from exponential, and then more generally gamma, populations considered in the previous chapter, or samples from linear models and multivariate populations considered in later chapters. The reasons for this are clear. The question of the treatment of outliers is a practical problem; it arises with the occurrence of aberrant or discordant observations in actual data, especially data which by their very method of generation are subject to contamination or to gross errors of recording or to gross errors of measurement. This is in particular true of repeated observations of some quantity made in the course of scientific experiment, observations which *a priori* could be considered to constitute a univariate normal sample, and it is in this context that the need for coherent and objective methods and techniques for handling spurious observations became pressingly apparent in the latter half of the nineteenth century. It was by workers involved in experimental sciences, notably astronomy, that some of the earliest approaches to the problem were made. Very much the same reasons that led to outliers in the context of univariate

normal samples being the first case to be studied still hold true in the present day and make the further study and refinement of techniques in this seemingly narrow area important, and, from the viewpoint of the practical experimentalist, essential and inevitable. It is with outliers in univariate normal samples that this chapter, and the next are concerned; this chapter primarily with single outliers and the next primarily with multiple outliers. In some cases, however, it is useful to avoid the separation and to consider the two cases concurrently. The particular aspect of the wide general problem that is considered here is the detection of outliers, rather than their accommodation by the use of robust estimates or by trimming or by Winsorisation or by other methods. The problem of how to proceed, as discussed earlier, when an outlier is detected is one which needs consideration in the light of the particular experimental situation under study.

Numerous criteria for the detection, and rejection, of outliers have been proposed. Some are based upon objective considerations, others have only intuitive appeal: intuition which may well be illfounded as with Chauvenet's criterion for the rejection of outliers. The following sections will be concerned in the main with likelihood-based criteria; likelihood-based that is in the face of particular alternatives to the null hypothesis that all the observations come from the same normal population. It is possible that either, or both, of the two parameters specifying the parent normal population may be considered to be known (or at least may be independently estimated), and thus not have to be

estimated internally from the data. These various cases are considered separately in the followingsections; the first section establishes the terminology and notation.

3.1 Preliminaries

Let $x_1,\ldots,x_n$ be a sample from a normal population with density

$$\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\quad(-\infty < x < +\infty). \tag{3.1}$$

Denote the ordered sample by $x_{(1)},\ldots,x_{(n)}$ where $x_{(1)} < \ldots < x_{(n)}$.

Let $\bar{x}_r = \frac{1}{r}\sum_{i=1}^{r} x_i$ and $s_r^2 = \frac{1}{r}\sum_{i=1}^{r}(x_i-\bar{x}_r)^2$, $(1 \leq r \leq n)$.

Most of the criteria that have been considered for detecting outliers fall into three broad classes. These consist firstly of statistics based upon either the studentized or else the standardised range, that is statistics of the form $(x_{(n)}-x_{(1)})/\hat{\sigma}$ where $\hat{\sigma}$ is either an internal or external estimate of σ as considered by David et al (1954), and Hartley (1944), respectively. When σ is known, σ itself is used, this is the case considered by Student (1927)

Secondly there are statistics involving the standardised or studentized extreme deviation from the mean, that is of the form $(x_{(n)}-\hat{\mu})/\hat{\sigma}$ where again $\hat{\mu}$ and $\hat{\sigma}$ are either the known values of μ and σ or are estimated, either internally from the sample itself or externally. Thirdly there are statistics of the Dixon type, that is of the form given in (2.3).

Falling into rather different categories are statistics of the form $(x_{(n)}-x_{(n-1)})/\sigma$ investigated by Irwin (1925) (in a sense this is in a class between standardised range and Dixon type criteria) and other more specialised criteria, specialised in the

sense that they may be designed to have optimal properties for special alternative hypotheses. An example of this latter kind would be the coefficient of skewness as considered by Ferguson (1961). Statistics in all the categories have the two essential qualifications for consideration as outlier detecting criteria; firstly their value is sensitive to the presence of outliers, and secondly they are independent, where these are unknown, of both the scale and the origin of measurement used. It will be shown that statistics in the second category, i.e. ones based upon the studentized extreme deviation from the mean, may in certain cases be regarded as likelihood-based and it is with statistics of this form that the succeeding sections will be concerned.

The tests and criteria proposed and discussed in detail in this chapter are for outliers at the upper end of the sample. The null hypothesis (referred to throughout as H_o) is that all the observations are from the same normal population $N(\mu,\sigma^2)$ and the alternative hypothesis is that all observations but one come from that same normal population whilst the aberrant observation itself arises from a different normal population $N(\mu_1,\sigma_1^2)$, where either $\mu_1 > \mu$ and $\sigma_1 = \sigma$ or $\mu_1 = \mu$ and $\sigma_1 > \sigma$. (The restriction that the aberrant observation arises from a normal distribution may of course be relaxed but this is done at the price of sacrificing the calculation of the likelihood explicitly under the alternative hypothesis). The symmetrical nature of the situation ensures that tests and criteria for outliers at the lower end of the sample will be essentially identical. However the case commonly met in practice is that there is no a priori reason for suspecting that an

outlier may occur at one particular end of the sample and so the observation actually tested as outlying is the more extreme of the largest and the smallest. This is equivalent to formulating the alternative hypothesis (to the null hypothesis referred to above) as that all but either the maximum or the minimum (but not both) observations come from that population, the remaining observation from some other population. In this situation the appropriate test criterion, when, for example, using the studentized extreme deviation from the mean, should be $\max\{(x_{(n)}-\hat{\mu})/\hat{\sigma},\ (\hat{\mu}-x_{(1)})/\hat{\sigma}\}$ rather than either of the two statistics based upon $x_{(n)}$ and $x_{(1)}$ separately. This point will be returned to in a later section.

3.2 The case both μ and σ known

This very simple case is of less practical importance than those considered later but is presented here for completeness.

Under the hypothesis H_o the log-likelihood is

$$-\tfrac{1}{2}n\log(2\pi\sigma^2)-\tfrac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2$$

and under the alternative hypothesis H_1, that all the observations except one, x_n without loss of generality, arise from the same normal population $N(\mu,\sigma^2)$ and x_n arises from the normal population $N(\mu_1,\sigma^2)$, where $\mu_1 > \mu$, the maximised log-likelihood is, on substituting $\hat{\mu}_1 = x_n$,

$$-\tfrac{1}{2}n\log(2\pi\sigma^2)-\tfrac{1}{2}\sum_{i=1}^{n-1}(x_i-\mu)^2/\sigma^2.$$

The difference in maximised log-likelihoods under H_o and H_1 is thus $\frac{1}{2}(x_n-\mu)^2/\sigma^2$.

If $z_{(n)}=(x_{(n)}-\mu)/\sigma$ then $z_{(n)}$ will be a likelihood-based criterion appropriate for testing the largest observation as outlying. The null distribution of $z_{(n)}$ is well known (Tippett 1925) (it is the maximum of n observations of a standard N(0,1) normal distribution) and is extensively tabulated. (Biometrika Tables for Statisticians Vol. I, Table 24)

3.3 The case μ unknown σ known

Let $u_{r,i} = (x_i - \bar{x}_r)/\sigma$ and $u_{r,(i)} = (x_{(i)} - \bar{x}_r)/\sigma$, $(1 \leqslant i \leqslant r)$. $u_{n,n}$ and $u_{n,(n)}$ will be abbreviated to u_n and $u_{(n)}$ when the sample size is clearly n.

Under the hypothesis H_o the log-likelihood of the sample $x_1,\ldots,x_n$ is

$$-\tfrac{1}{2}n\log(2\pi\sigma^2)-\tfrac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2,$$

which is maximised when $\mu = \bar{x}_n$, (assuming that the true value of σ is known), giving a maximised log-likelihood under H_o of

$$-\tfrac{1}{2}n\log(2\pi\sigma^2)-\tfrac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2/\sigma^2.$$

Under the alternative hypothesis H_1, that x_n arises from a normal population $N(\mu_1,\sigma^2)$, where $\mu_1 > \mu$, and all the other observations are from the same distribution $N(\mu,\sigma^2)$, the log-likelihood of the sample is

$$-\tfrac{1}{2}n\log(2\pi\sigma^2)-\tfrac{1}{2}\sum_{i=1}^{n-1}(x_i-\mu)^2/\sigma^2-\tfrac{1}{2}(x_n-\mu_1)^2/\sigma^2,$$

which is maximised when $\mu = \bar{x}_{n-1}$ and $\mu_1 = x_n$, giving a maximised log-likelihood of

$$-\tfrac{1}{2}n\log(2\pi\sigma^2)-\tfrac{1}{2}\sum_{i=1}^{n-1}(x_i-\bar{x}_{n-1})^2/\sigma^2.$$

The difference in maximised log-likelihoods under H_o and H_1 is thus

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

which may be written as

$$\frac{n}{2(n-1)}\,\frac{(x_n-\bar{x}_n)^2}{\sigma^2} = \frac{n}{2(n-1)}\,u_n^2 .$$

It follows therefore that the statistic $u_{(n)}$ is essentially the increase in log-likelihood consequent upon acceptance of the alternative hypothesis that the queried observation arises from a normal population with a different mean.

The null distribution of $u_{(n)}$ is well known, having been found independently by McKay (1935), Nair (1948) and Grubbs (1950). However the following derivation, analagous to the one in 2.2.1., is presented for its simplicity and because the method may be extended to the more general cases when σ is unknown.

3.3.1 Recursive Algorithm for the Distribution of $u_{(n)}$

Suppose $u_{(n)}$ has density function $a_n(.)$ and distribution function $A_n(.)$. Then

$$a_n(y)dy = P\,[\,u_{(n)} \in (y,y+dy)]$$

$$= \sum_{j=1}^{n} P\,[\,u_{n,j} \in (y,y+dy), u_{n,(n)} = u_{n,j}\,]$$

$$= n\;P\,[\,u_{n,n} \in (y,y+dy), \max_{j=1}^{n-1} u_{n,j} < u_{n,n}\,]$$

$$= n\;P\,[\,u_{n,n} \in (y,y+dy), \max_{j=1}^{n-1} x_j < x_n\,]$$

$$= n \; P\left[u_{n,n} \in (y,y+dy),\ \max_{j=1}^{n-1} u_{n-1,j} < (x_n - \bar{x}_{n-1})/\sigma \right]$$

$$= n \; P\left[u_{n,n} \in (y;y+dy),\ u_{n-1,(n-1)} < \frac{n}{n-1} u_n \right]$$

$$\left(\text{since } x_n - \bar{x}_{n-1} = \frac{n}{n-1}(x_n - \bar{x}_n)\right)$$

$$= n \; P\left[u_{n,n} \in (y,y+dy)\right]\left[P \;\; u_{n-1,(n-1)} < \frac{n}{n-1} y \right],$$

(noting that $(x_j - \bar{x}_{n-1})$ and $(x_n - \bar{x}_n)$ have zero covariance and are therefore independent for each $j=1,2,\ldots,n-1$).

Thus

$$a_n(y) = n \sqrt{\frac{n}{n-1}} \cdot \frac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}ny^2/(n-1)\} A_{n-1}\{ny/(n-1)\} \qquad (3.2)$$

(since for arbitrary j, $u_{n,j}$ follows a normal distribution $N(0,(n-1)/n)$), which is equation (29) of McKay (1935). The density functions $a_n(.)$ may now be found for successive values of n, for when $n=2$ $\quad u_{(2)} = (x_{(2)} - x_{(1)})/2\sigma$, and then

$$a_2(y) = \frac{2}{\sqrt{\pi}} \exp\{-y^2\} \qquad (y \geq 0).$$

Percentage points of the distribution have been tabulated by Nair (1948b), and are given in Table 25, Biometrika Tables for Statisticians, Vol I, Table 25.

3.4 The case μ known and σ unknown

Under the null hypothesis H_o the log-likelihood of the sample $x,\ldots,x_n$ is

$$-n\log\sigma-\tfrac{1}{2}n\log(2\pi)-\tfrac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2$$

which is maximised when $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$, assuming that the value of μ is known, giving a maximised log-likelihood of

$$-\tfrac{1}{2}n\log(\sum_{i=1}^{n}x_i-\mu)^2)+\tfrac{1}{2}n\log(n)-\tfrac{1}{2}n\log(2\pi)-\tfrac{1}{2}n.$$

Under the alternative hypothesis that all the observations other than x_n arise from the same normal distribution $N(\mu,\sigma^2)$ and x_n arises from a normal distribution $N(\mu,\sigma_1^2)$ where $\sigma_1 > \sigma$, the likelihood of the sample is

$$-(n-1)\log\sigma-\log\sigma_1-\tfrac{1}{2}n\log(2\pi)-\tfrac{1}{2}\sum_{i=1}^{n-1}(x_i-\mu)^2/\sigma^2-\tfrac{1}{2}(x_n-\mu)^2/\sigma_1^2$$

which is maximised when $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i-\mu)^2$ and $\sigma_1^2 = (x_n-\mu)^2$, giving a maximised log-likelihood of

$$-\tfrac{1}{2}(n-1)\log(\sum_{i=1}^{n-1}(x_i-\mu)^2)-\tfrac{1}{2}\log(x_n-\mu)^2-\tfrac{1}{2}n\log(2\pi) + \tfrac{1}{2}(n-1)\log(n-1)-\tfrac{1}{2}n.$$

The difference in maximised log-likelihoods under the two hypotheses is therefore

$$\tfrac{1}{2}n\log\{\sum_{i=1}^{n}(x_i-\mu)^2/\sum_{i=1}^{n-1}(x_i-\mu)^2\}-\tfrac{1}{2}\log\{(x_n-\mu)^2/\sum_{i=1}^{n-1}(x_i-\mu)^2\} + \tfrac{1}{2}(n-1)\log(n-1)-\tfrac{1}{2}n\log(n),$$

which may be written as

$$-\tfrac{1}{2}n\log(1-T_n)-\tfrac{1}{2}\log\{T_n/(1-T_n)\} + \tfrac{1}{2}(n-1)\log(n-1)-\tfrac{1}{2}n\log(n),$$

(where $T_n = (x_n-\mu)^2/\sum_{i=1}^{n}(x_i-\mu)^2$).

It follows that an appropriate likelihood based statistic for testing the most extreme observation (either $x_{(1)}$ or $x_{(n)}$, whichever is the greater absolute distance from μ) as outlying is

$$T_{(n)} = \max\{(x_{(1)}-\mu)^2,(x_{(n)}-\mu)^2\}/\sum_{i=1}^{n}(x_i-\mu)^2,$$

which, under the null hypothesis H_o, is the ratio of the largest of a set of n independent variates, each distributed as χ^2 with one degree of freedom, to the sum of those variates. Thus the problem of detecting an outlier in a random sample from a normal population with known mean is essentially equivalent to the simpler problem of detecting an upper outlier in a random sample from a χ^2_1 population. This problem was considered in the more general case of a gamma population in section 2.2. Putting $r = \frac{1}{2}$ in (2.7) gives

$$P[\,T_{(n)} > u\,] = nP[\,F_{1,n-1} > (n-1)u/(1-u)\,] \quad u \geq \tfrac{1}{2}$$

$$P[\,T_{(n)} > u\,] < nP[\,F_{1,n-1} > (n-1)u/(1-u)] \quad u < \tfrac{1}{2} \qquad (3.3)$$

where $F_{1,n-1}$ denotes a variate following an F-distribution with (1,n-1) degrees of freedom. Upper 5% and 1% points of $T_{(n)}$ for n = 2(1)10,12,15,20 have been calculated by Eisenhart, Hastay and Wallis (1947) and are given in Table 31(a) of Biometrika Tables for

Statisticians Vol. I.

It is of interest to note that not only is $T_{(n)}$ a likelihood-based statistic in the sense described above, but it can also be regarded as likelihood-based when the alternative hypothesis is that the queried observation arises from a normal distribution with the same variance as, but a different mean from, the parent distribution of the rest of the sample; that is that x_n arises from a normal distribution $N(\mu_1,\sigma^2)$ with $\mu_1 \neq \mu$ (rather than $N(\mu,\sigma_1^2)$ with $\sigma_1 > \sigma$ as considered above).

This follows since under this alternative hypothesis the log-likelihood is

$$-n\log\sigma-\tfrac{1}{2}n\log(2\pi)-\tfrac{1}{2}\sum_{i=1}^{n-1}(x_i-\mu)^2/\sigma^2-\tfrac{1}{2}(x_n-\mu_1)^2/\sigma^2,$$

which is maximised when

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n-1}(x_i-\mu)^2 \qquad \text{and} \qquad \mu_1 = x_n$$

giving a maximised log-likelihhod of

$$-\tfrac{1}{2}n\log\sum_{i=1}^{n-1}(x_i-\mu)^2-\tfrac{1}{2}n\log 2\pi + \tfrac{1}{2}n\log(n)-\tfrac{1}{2}n,$$

so that the increase in maximised log-likelihhods in this case is

$$-\tfrac{1}{2}n\log(1-T_n) .$$

The problem of detecting outliers in normal samples with known

mean would seem to be one with practical applications. One context which suggests itself is the control of a manufacturing process by a quality control chart, particularly one based upon a continuous variable. The detection of a sudden or abnormal change in the operating conditions from the control chart plots could be thought of perhaps as a problem in outlier detection. This would apply whether the change was representable as a shift in mean or an increase in variance.

3.5 The case when both μ and σ are unknown

The study of outliers in samples drawn from a normal population with unknown mean and variance is of considerable practical importance and has received much attention in recent years. Broadly the work falls into two categories; firstly the detection of outliers and secondly their accommodation by various techniques of robust estimation and analysis, the latter category would also include the various aspects of the Bayesian approach to the analysis of outliers (such as that of de Finetti (1961), Box & Tiao (1968), Dempster & Rosner (1971)). In some cases the Bayesian methods, though directed to the accommodation of outliers, are equivalent to their detection. The detection of outliers in this situation has been studied by a variety of authors, who have proposed a number of different criteria and test statistics, usually on intuitive grounds. In some cases the exact null distribution of the statistic has been calculated and tables of upper percentage points have been derived; in other cases the exact null distribution has not been found and tables of upper and lower bounds for the upper percentage points are given, these are obtained by iterative procedures based upon Bonferroni's inequalities.

This section is concerned with statistics that may be regarded as likelihood-based under certain conditions, and with closely related criteria. Two cases are to be distinguished; the first is when the actual sample under inspection contains all the available information relating to the parent normal population, the second is when an external estimate of the parent population variance is available (independent of the sample) In the first case

the null distribution of a likelihood-based statistic will be obtained by the recursive methods described earlier, in the second case it will be shewn that the natural modification, using the extra information available, of the likelihood-based criteria results in a statistic whose null distribution can also be obtained recursively.

3.5.1 Studentized Criteria Based upon an Internal Estimate of σ

Under the null hypothesis H_o the log-likelihood of the sample $x_1,\ldots,x_n$, is

$$-\tfrac{1}{2}n\log(2\pi)-n\log\sigma-\tfrac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2,$$

where both μ and σ may vary.

This is maximised when $\mu = \bar{x}_n$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$, giving a maximised log-likelihood of

$$-\tfrac{1}{2}n\log(2\pi)-\tfrac{1}{2}n\log\{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\}+\tfrac{1}{2}n\log(n)-\tfrac{1}{2}n.$$

Under the alternative hypothesis, H_1, that all the observations with the exception of x_n are from the same distribution $N(\mu,\sigma^2)$ and x_n arises from a normal distribution $N(\mu_1,\sigma^2)$, with $\mu_1 > \mu$, the log-likelihood of the sample is

$$-\tfrac{1}{2}n\log(2\pi)-n\log\sigma-\tfrac{1}{2}\sum_{i=1}^{n-1}(x_i-\mu)^2/\sigma^2-\tfrac{1}{2}(x_n-\mu_1)^2/\sigma^2.$$

This is maximised when $\mu = \bar{x}_{n-1}$, $\mu_1 = x_n$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n-1}(x_i-\bar{x}_{n-1})^2$,

giving a maximised log-likelihood of

$$-\tfrac{1}{2}n\log(2\pi)-\tfrac{1}{2}n\log\{\sum_{i=1}^{n-1}(x_i-\bar{x}_{n-1})^2\}+\tfrac{1}{2}n\log(n)-\tfrac{1}{2}n.$$

The difference in maximised log-likelihoods is thus

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

which may be written as

$$\tfrac{1}{2}(n\log(n-1)-n\log\{(n-1)-nU_n^2\})$$

where

$$U_n^2 = (x_n-\bar{x}_n)^2/\{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\},$$

upon application of the identity

$$\sum_{i=1}^{n}(x_n-\bar{x}_n)^2 = \sum_{i=1}^{n-1}(x_i-\bar{x}_{n-1})^2+\frac{n}{n-1}(x_n-\bar{x}_n)^2. \qquad (3.4)$$

It follows that an appropriate statistic for testing the largest observation $x_{(n)}$ as outlying is one of the form

$$U_{(n)} = (x_{(n)}-\bar{x}_n)/\{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\}^{\frac{1}{2}}.$$

This statistic is of course equivalent to the studentized extreme deviation from the mean considered by many authors, in particular Pearson and Chandra Sekar (1936) and Grubbs (1950).

It should be noted that in general $U_{(n)}$ cannot be regarded as likelihood-based when the alternative hypothesis is that the queried observation arises from a normal population $N(\mu,\sigma_1^2)$, that is from a population with a shift in variance rather than a shift in location.

With the following notation the null distribution of $U_{(n)}$ will be obtained by a recursive procedure similar to that of earlier sections.

Let

$$S_r^2 = \sum_{i=1}^{n}(x_i-\bar{x}_r)^2,$$

$$U_{r,i} = (x_i-\bar{x}_r)/S_r.$$

In particular

$$U_{n,n} = (x_n-\bar{x}_n)/S_n$$

and

$$U_{n,(n)} = \max_{i=1}^{n}\{U_{n,i}\} = (x_{(n)}-\bar{x}_n)/S_n$$

which will be abbreviated to U_n and $U_{(n)}$ when the sample size is clearly n.

Now for any i, $1 \leq i \leq r$, S_r^2 may be written as the sum of two independent variates, X and Y say, where $X = \frac{r}{r-1}(x_i-\bar{x}_r)^2$ which is distributed as $\sigma^2\chi^2$ with one degree of freedom and where

Y is distributed as $\sigma^2\chi^2$ with r-2 degrees of freedom. Then $U_{r,i}$ may be written as $(\frac{r}{r-1}+Y/(x_i-\bar{x}_r)^2)^{-\frac{1}{2}}$ or as $(\frac{r}{r-1})^{-\frac{1}{2}}(1+\frac{r-2}{Z^2})^{-\frac{1}{2}}$ where Z is distributed as Student's t with n-2 degrees of freedom. It follows that for arbitrary i $U_{r,i}$ has density function $\phi_r(.)$ given by

$$\phi_{(u)} = \sqrt{\frac{r}{\pi(r-1)}} \cdot \frac{\Gamma\left(\frac{r-1}{2}\right)}{\Gamma\left(\frac{r-2}{2}\right)} \{1-ru^2/(r-1)\}^{\frac{1}{2}(r-4)}, \quad 0 \leq u \leq \sqrt{\{(r-1)/r\}}. \qquad (3.5)$$

Now suppose that $U_{(n)}$ has density function $b_n(.)$ and distribution function $B_n(.)$. Then

$$b_n(u)du = P[\, U_{(n)} \in (u,u+du)\,]$$

$$= \sum_{j=1}^{n} P[\, U_{n,j} \in (u,u+du),\ U_{n,(n)} = U_{n,j}\,]$$

$$= nP[\, U_{n,n} \in (u,u+du),\ x_{(n-1)} < x_n\,]$$

$$= nP[\, U_{n,n} \in (u,u+du),\ (x_{(n-1)}-\bar{x}_{n-1})/S_{n-1} < (x_n-\bar{x}_{n-1})/S_n\,]$$

$$= nP[\, U_{n,n} \in (u,u+du),\ (x_{(n-1)}-\bar{x}_{n-1})/S_{n-1} < \{n^2u^2/((n-1)^2-n(n-1)u^2)\}^{\frac{1}{2}}\,]$$

(upon application of identity (3.4)).

Now $\bar{x}_n$ and S_n^2 are complete sufficient statistics for μ and σ^2, also for each j, $1 \leq j \leq n-1$, the distribution of $(x_j-\bar{x}_{n-1})/S_{n-1}=U_{n-1,j}$ is independent of both μ and σ^2 and so it follows (Basu (1955)) that the joint distribution of $\bar{x}_n$ and S_n^2 is independent of that of $U_{n-1,j}$

for each j, $1 \leq j \leq n-1$. Further $U_{n-1,j}$ is independent of x_n, (for $1 \leq j \leq n-1$) and so $U_{n,n}$ and $U_{n-1,j}$ are independent for each j, $1 \leq j \leq n-1$. Thus $U_{n,n}$ and $U_{n-1,(n-1)}$ are independent and then

$$b_n(u)du = nP\left[U_{n,n} \in u,u+du\right] P\left[U_{n-1,(n-1)} < \{n^2u^2/((n-1)^2-n(n-1)u^2)\}^{\frac{1}{2}}\right],$$

yielding the recurrence relation

$$b_n(u) = n\phi_n(u)\, B_{n-1}\left[\{n^2u^2/((n-1)^2-n(n-1)u^2)\}^{\frac{1}{2}}\right], \quad 0 < u < \sqrt{\{(n-1)/n\}}. \tag{3.6}$$

Now U_n^2 may be written as

$$U_n^2 = \frac{n-1}{n} - \frac{(n-1)}{nS_n^2}\left\{\sum_{i=1}^{n-1}(y_i-\bar{y}_{n-1})^2\right\}$$

where $y_i = x_i - \bar{x}_n$ and $\bar{y}_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1} y_i = -y_n/(n-1)$.

Consequently $U_n^2 \leq (n-1)/n$ for all samples $x_1,\ldots,x_n$. (Note that equality is achieved if all the observations except one are equal). Thus if $u \geq \{(n-1)/n\}^{\frac{1}{2}}$ then $B_n(u) = 1$. If

$$\{\tfrac{1}{2}(n-2)/n\}^{\frac{1}{2}} \leq u \leq \{(n-1)/n\}^{\frac{1}{2}}$$

then

$$\{n^2u^2/((n-1)^2-n(n-1)u^2)\}^{\frac{1}{2}} \geq \{(n-2)/(n-1)\}^{\frac{1}{2}}$$

and (3.6) becomes

$$b_n(u) = n\phi_n(u)$$

$$= n\sqrt{\frac{n}{\pi(n-1)}}\ \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}\{1-nu^2/(n-1)\}^{\frac{1}{2}(n-4)}. \qquad (3.7)$$

If $\{\frac{1}{3}(n-3)/n\}^{\frac{1}{2}} \leq u \leq \{\frac{1}{2}(n-2)/n\}^{\frac{1}{2}}$ then

$$\{\tfrac{1}{2}(n-3)/(n-1)\}^{\frac{1}{2}} \leq \{n^2u^2/(n-1)^2-n(n-1)u^2)\}^{\frac{1}{2}} \leq \{(n-2)/(n-1)\}^{\frac{1}{2}}$$

and so from (3.7)

$$B_{n-1}(t_n) = 1 - \int_{t_n}^{\sqrt{\{(n-2)/(n-1)\}}} (n-1)\phi_{n-1}(t)dt,$$

(writing t_n for $\{n^2u^2/(n-1)^2-n(n-1)u^2)\}^{\frac{1}{2}}$),

and then $b_n(u)$ may be obtained from (3.6).

Proceeding in this fashion it can be seen that $b_n(u)$ may be obtained for successive values of n by evaluating it in each of the (n-1) intervals $(\{(n-r-1)/(r+1)n\}^{\frac{1}{2}}, \{(n-r)/(rn\}^{\frac{1}{2}})$ for $r = 1,2,\ldots,(n-2)$, using the recursive relation (3.6).

Now $U_{(2)} = \sqrt{\frac{1}{2}}$ for all samples x_1, x_2, so $B_2(u) = 0$ or 1 according as $u < \sqrt{\frac{1}{2}}$ or $u \geq \sqrt{\frac{1}{2}}$. Thus

$$b_3(u) = 3\phi_3(u) \qquad u \geq \sqrt{\frac{1}{6}}$$

$$0 \qquad u < \sqrt{\frac{1}{6}}. \qquad (3.8)$$

$b_n(u)$ may now be obtained for $n = 4,5,\ldots$ using (3.6)

recursively with (3.7) initially. It may be noted that

$$b_n(u) = 0 \text{ if } 0 \leq u \leq \{n(n-1)\}^{-\frac{1}{2}}.$$

Upper percentage points of $U_{(n)}$, or of a monotonic function of $U_{(n)}$, have been tabulated by Quesenberry and David (1961) and Grubbs (1950); Biometrika Tables for Statisticians, Vol 1, Table 26(a) give the upper 5% and 1% points of $U_{(n)}$ for sample sizes 3(1)10,12,15,20. Grubbs and Beck (1972) give 10%,5%,2·5% 1%,·5%,·1% points of the statistic $(n-1)^{\frac{1}{2}} U_{(n)}$ for sample sizes 3(1)147.

3.5.2 Studentized Criteria Incorporating an External estimate of σ

It may happen that in addition to the sample $x_1, x_2, \ldots x_n$ from a Normal population $N(\mu,\sigma^2)$ there is available an unbiased external estimate of σ^2, s_ν^2 say, such that νs_ν^2 is distributed as χ^2 with ν degrees of freedom independently of the sample under study. This may occur, for example, when repeated sets of observations are made on the same population, in addition to the sample under study, and the estimate of σ^2 is obtained from these additional sets of observations. Another common situation when an independent estimate of σ^2 is available is in the analysis of an orthogonally designed experiment, when, in the context of variety trials, one would be interested in the variety with the highest or lowest yield. An interesting example in this latter category is discussed by Pearson and Hartley (Biometrika Tables for Statisticians, Vol 1, Example 25). They quote data relating to the yields of four strains of wheat replicated in five blocks, and show that the

minimum yielding strain is highly significantly low. Further they show that had the residual error been rather larger, so that an overall F-test would have failed to detect any difference between strains in general, the 'outlier' test would nevertheless still have detected the low-yielding strain.

In cases such as these it would seem advantageous on intuitive grounds to utilize the extra available information when testing for outliers and modify the test statistic $U_{(n)}$ of 3.5.1 to take account of this. The natural modification to make is to replace the divisor S_n, which is essentially an estimate of the standard deviation σ, by a pooled estimate of the standard deviation, R_n say, where $R_n^2 = S_n^2 + \nu s_\nu^2$. Then a test criterion for testing the largest observation $x_{(n)}$ as outlying is

$$V_{(n)} = (x_{(n)} - \bar{x}_n)/R_n.$$

Although this statistic is not likelihood-based for any immediate alternative hypothesis which allows the occurrence of an outlier in the sample, it is possible to show that under certain restrictions a test based upon this statistic maximises the probability of detecting a single outlier, (Kudo (1956)). The distribution of statistics of this form has been considered by Quesenberry and David (1961) who obtained approximate percentage points for the statistic using iterative procedures, based upon Bonferroni's inequalities. They tabulate the upper 5% and 1% points of $V_{(n)}$ for n = 3(1),10,12,15,20 and ν = 0(1),10,12,15,20, 24,30,40,50. The tables are reproduced in the Biometrika Tables for Statisticians, Table 26(a).

The exact null distribution of $V_{(n)}$ can in principle be obtained by the recursive method of the earlier sections. With notation analogous to that of 3.5.1, it can readily be shown by an extension of the method of 3.5.1, that for arbitrary i, $V_{i,r}$ has density function $\psi_r(.)$ given by

$$\psi_r(u) = \left(\frac{r}{\pi(r-1)}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(r+\nu-1))}{\Gamma(\frac{1}{2}(r+\nu-2))} \{1-ru^2/(r-1)\}^{\frac{1}{2}(r+\nu-4)},$$

$$|u| \leq \sqrt{\{(r-1)/r\}}. \qquad (3.9)$$

Suppose that $V_{(n)}$ has density function $c_n(.)$ and distribution function $C_n(.)$ then it is easily seen that

$$c_n(u) = n\psi_n(u)C_{n-1}[\{n^2u^2/((n-1)^2-n(n-1)u^2)\}^{\frac{1}{2}}]. \qquad (3.10)$$

This recurrence relation is essentially identical to (3.6) with $\phi_n(.)$ replaced by $\psi(.)$. Further $V_n^2 \leq (n-1)/n$ for all samples $x_1,\ldots,x_n$ and it follows that the null distribution of $V_{(n)}$ may be evaluated in successive intervals

$$(\{(n-r-1)/(r+1)n\}^{\frac{1}{2}}, \{(n-r)/rn\}^{\frac{1}{2}})$$

for $r = 1,2,\ldots,(n-1)$, for each $n = 3,4,\ldots$, in a manner which parallels the derivation of the distribution of $U_{(n)}$. The principal distinction is that whereas $U_{(2)} = \sqrt{\frac{1}{2}}$, a constant,

$$V_{(2)} = \tfrac{1}{2}(x_{(2)}-x_{(1)})/\{\tfrac{1}{2}(x_{(2)}-x_{(1)})^2+\nu s_\nu^2\}^{\frac{1}{2}},$$

which is a non-degenerate random variable. The distribution of

$V_{(2)}$ may be obtained by noting that

$$V_{(2)} = \frac{\sqrt{\frac{1}{2}}|t_\nu|}{\{t_\nu^2+\nu s_\nu^2\}^{\frac{1}{2}}} ,$$

where t_ν denotes a variate following Student's t-distribution with ν degrees of freedom.

Thus

$$c_2(u) = 2\sqrt{\frac{2}{\pi}}\ \frac{\Gamma(\frac{1}{2}(\nu+1))}{\Gamma(\frac{1}{2}\nu)}\ \{1-2u^2\}^{\frac{1}{2}(\nu-2)}, \quad 0 \leq u \leq \sqrt{\tfrac{1}{2}} \quad . \qquad (3.11)$$

$c_n(u)$ may now be obtained for further values of n using the recurrence relation (3.10).

There are situations when it is desirable to use a test statistic which does not involve an internal estimate of the population variance. If for example the sample contains more than one outlier then the estimate of the population variance obtained internally from the sample will be seriously inflated, leading to a small value of the test statistic $V_{(n)}$, and indeed $U_{(n)}$, so lessening the chance of detecting a single outlier using either of these statistics. It should be noted that although the estimate of the population variance will be inflated by the presence of a single outlier, this inflation is of no consequence when testing for that outlier since the test statistics $V_{(n)}$, and $U_{(n)}$, are essentially equivalent to statistics based upon an estimate of the population variance obtained by omitting the suspected outlier. For example, if

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

then it is easy to show, using identity (3.4), that

$$U_n' = \{n/(n-1)\}U_n/\{1-(n/(n-1))U_n^2\}^{\frac{1}{2}}.$$

It is only when there are two or more outliers that the inflation of the estimated population variance may result in failure to identify a single outlier. This is the problem of "masking" discussed in the next chapter.

If it is suspected that the sample may contain more than one outlier the experimenter may wish to test for just one outlier using a statistic which will afford some protection if there is in fact more than one outlier. This procedure is an expedient; if it is thought that there may be multiple outliers in the sample then this should be tested directly. This situation is considered in the next chapter. The natural modification of the likelihood-based test statistic to use in such cases is the externally studentized extreme deviation from the mean or equivalently the statistic

$$W_{(n)} = (x_{(n)}-\bar{x})/\nu^{\frac{1}{2}}s_\nu.$$

Approximate percentage points of the statistic $\sqrt{\nu}W_{(n)}$ have been extensively tabulated; David (1956(b)) gives upper 10%, 5%, 2·5%, 1%, 0·5% and 0·1% points for n = 3(1)10,12 and ν = 10(1)20, 24,30,40,60,120, Pillai (1959) gives upper 5% and 1% points for n = 2(1),10,12 and ν = 1(1)10, Nair (1948) gives lower 5% and 1% points for n = 3(1)9 and ν = 10,15,30. These last have no immediate application to the problem of detecting outliers.

It is of interest to note that the distribution of $W_{(n)}$ may be derived by the recursive methods described earlier. With notation analagous to that used earlier, for arbitrary i $W_{r,i}$ has density function $\theta_r(.)$ given by

$$\theta_r(u) = \sqrt{\frac{n\nu}{n-1}}\ t_\nu\left(\sqrt{\frac{n\nu}{n-1}}\ u\right) ,$$

where $t_\nu(.)$ is the density of a variate following Student's t-distributions with ν degrees of freedom, that is

$$\theta_r(u) = \sqrt{\frac{n}{\pi(n-1)}}\ \frac{\Gamma(\frac{1}{2}(\nu+1))}{\Gamma(\frac{1}{2}\nu)}\ \{1+nu^2/(n-1)\}^{-\frac{1}{2}(\nu+1)} .$$

If $d_n(.)$ and $D_n(.)$ are the density and distribution functions of $W_{(n)}$ then the methods of 3.5.1 and 3.3.1 give

$$d_n(u) = n\theta_n(u)D_{n-1}\{nu/(n-1)\} . \qquad (3.12)$$

When n = 2 $\quad W_{(n)} = \frac{1}{2}(x_{(2)}-x_{(1)})/\nu^{\frac{1}{2}}s_\nu$, so

$$d_2(u) = 2\sqrt{\frac{2}{\pi}}\ \frac{\Gamma(\frac{1}{2}(\nu+1))}{\Gamma(\frac{1}{2}\nu)}\ \{1+2u^2\}^{-\frac{1}{2}(\nu+1)} \qquad (u \geq 0),$$

and then $d_n(u)$ may be found for successive values of n. The similarity of (3.12) with (3.10) is clear.

3.6 Two-sided criteria

The test criteria discussed in the preceding sections (with the exception of 3.4) are designed to test the null hypothesis H_o against the alternative that the observation at a specified end of the sample is an outlier; without loss of any essential generality it was assumed that the possible outlier was the maximum of the sample. These criteria would be employed in situations where for some reason one was interested only in detecting outliers which occurred at the upper end, or in situations where there were *a priori* reasons for believing that if an outlier should occur in the sample then it could only manifest itself as the maximum of the sample. The latter case corresponds to specifying the one-sided alternative hypothesis that the maximum observation is from a normal population $N(\mu_1,\sigma^2)$ with $\mu_1 > \mu$.

In many situations there are no *a priori* reasons for suspecting which end of the sample a possible single outlier may occur and one may be equally interested in either an upper or a lower single outlier. The alternative hypothesis in such situations may be formulated either as that the parent population of the most extreme observation (whether it be the maximum or minimum of the sample), is normal with a shift in mean (either increase or decrease corresponding to whether the maximum or minimum is the more extreme), or that this parent population is normal with an increase in variance, as considered in 3.4.

In these cases where the alternative hypothesis is two-sided, an appropriate two-sided test criterion is the larger of the two one-sided criteria appropriate for testing the maximum and the minimum

observations individually as outlying. That is, for example, in the case considered in 3.5.1, the appropriate statistic would be

$$U^* = \max \{-U_{(1)}, U_{(n)}\} = \max_{i=1}^{n} \{|x_i - \bar{x}_n| / S_n\} ,$$

with corresponding definitions of V* and W* of 3.5.2. In principle the null distribution of U* can be obtained by integration of the joint density of $U_{(1)}$ and $U_{(n)}$. This problem will be considered in the following chapter. However it is easy to obtain upper bounds for the percentiles of the distribution of U*; since

$$P[U^* > u] = P[U_{(n)} > u] + P[U_{(1)} < -u] - P[U_{(n)} > u,\ U_{(n)} < -u]$$

$$\leq 2P[U_{(n)} > u],$$

an upper bound for the α% point of the distribution of U* is provided by the upper $\frac{1}{2}\alpha$% point of $U_{(n)}$. Lower bounds for the percentage points of U* may be obtained from the Bonferroni inequality

$$P[U^* > u] \geq nP[U_{j,n}| > u] - \binom{n}{2} P[|U_{j,n}| > u, |U_{i,n}| > u].$$

For small samples and, in the case of the statistics V* and W*, for small values of ν the upper and lower bounds agree closely (Quesenberry and David (1961), Halperin et al. (1955).

Upper and lower bounds for the upper 5% and 1% of U* and V* are tabulated by Quesenberry and David (1961), and are reproduced in Biometrika Tables for Statisticians, Table 26(b), and of $\nu^{\frac{1}{2}}W^*$ by

Halperin et al (1955), who also tabulate those of the statistic u* for use in cases where the population variance is known, as in section 3.3.

Tietjen and Moore (1972) give approximate percentage points, obtained by Monte Carlo methods, for the statistic

$$\min\{\sum_{i=1}^{n-1}(x_{(i)}-\bar{x}_{n-1})^2, \sum_{i=2}^{n}(x_{(n)}-\bar{x}_{2,n})^2\}/S_n^2 \qquad (3.13)$$

$$(\text{where } \bar{x}_{2,n} = \frac{1}{n-1}\sum_{i=2}^{n} x_i)$$

which is equivalent to U*, but they erroneously state that their empirical 2α% points may be compared for accuracy with the α% points of the statistic

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

which were tabulated by Grubbs (1950). These points are of course lower bounds for the 2α% points of the statistic (3.13), and equiality is achieved only for small samples.

Chapter 4

Multiple Outliers in Univariate Normal Samples

There are two distinct approaches to the problem of detecting the presence of several outliers occurring simultaneously in a normal sample. The first approach is to use statistics specifically designed to test simultaneously several observations as outlying, that is, statistics whose value is markedly inflated when all the suspected observations are outlying.

The second approach, which has immediate intuitive appeal, is the so-called 'sequential' method; the most extreme observation is tested as outlying and, if declared an outlier, is rejected and the test repeated on the reduced sample, the process continues until the sample is reduced to one where no outliers can be detected. The test used at each stage can be based upon any of the criteria discussed in the previous chapter. McMillan and David (1971) and McMillan (1971) consider the case of detecting two outliers (assumed to be at the same end of the sample) using likelihood-based criteria and their natural modifications based upon external estimates of the population variance. Dixon (1953) and Ferguson (1961) suggest procedures based upon Dixon type criteria and the coefficients of skewness and kurtosis respectively. The problems associated with such a procedure are complex. Firstly there is the possibility that in cases where the parent population is mistakenly assumed to be normal instead of fatter tailed an unreasonably large number of observations may be declared outliers. Indeed it is easy to construct artificial samples of any specified size where all but two of the observations would be declared outlying by such a

'sequential' procedure. Secondly, and of more practical importance, is the problem of "masking"; the presence of several outliers, as detected by a simultaneous test, could inflate the internal estimate of the variance so greatly that even the most extreme observation would not be detected as outlying by a test of that observation singly and the procedure would stop. Such cases are discussed by Pearson and Chandra Sekar (1936) and Tietjen and Moore (1972).

This problem may be overcome by using an external estimate of the variance (see 3.5.2) if this is available, or by employing one of the many Dixon type criteria, although deciding which of the criteria was the most appropriate in any particular case would involve careful preliminary inspection of the data. A third difficulty with the 'sequential' procedure is the calculation of the significance probabilities of a group of several outliers; while the probability of rejecting the null hypothesis that all observations are from the same normal populations depends only upon the level of the first test in the sequence used, this will not be the same as the significance level attached to the group of observations eventually declared as outlying. This fact appears to have been overlooked by Tietjen and Moore (1972). McMillan and David (1971) and McMillan (1971) have considered the calculation of these probabilities for the case of two outliers at the same end of the sample, when using the likelihood-based criteria $U_{(n)}$, but see the correction in Moran and McMillan (1973). Hawkins (1973) considers the repeated use of the partially externally studentized criteria $V_{(n)}$ in samples containing a maximum of two upper outliers and calculates the loss of power due to 'masking'.

The approach to the detection of multiple outliers discussed

in later sections of this chapter is the non-sequential method described in the opening paragraph; various test criteria will be considered for testing several possible outliers simultaneously. There are of course many problems associated with this approach also. It is implicit that the procedure is to perform only one test, for k outliers say, and that no test for k-1 or k+1 outliers, depending upon the result of the test for k outliers, will be made. In many cases it may be unreasonable to expect the experimenter to specify in advance how many observations he wishes to test as outlying and some decision has to be made as to which value of k to use. This problem is considered by Daniel (1959) who suggests the use of a sequence of half-normal plots of the data and by Dempster and Rosner (1972) who consider a "semi-Bayesian" approach. A further complication is that the choice of k may determine which set of observations to test, and the set determined by the choice k may not be a subset of the set determined by the choice k+1. Complementary to the phenomenon of "masking" in the 'sequential' procedure is the phenomenon of "swamping"; if the sample contains only one outlier which is sufficiently far removed from the rest of the data then a test of the two or more most extreme observations as outlying may erroneously detect two or more outliers, only one of which is a 'true' outlier.

In the succeeding sections it will be assumed that the choice of k, the number of observations to be tested as outlying, has been made, whether by external a priori reasons or by one of the procedures referred to earlier. The tests and criteria discussed will be for a specified number of outliers at specified ends of

the sample, these criteria correspond to the "one-sided" criteria discussed in the previous chapter; the extension of the methods to situations corresponding to those requiring "two-sided" criteria increases in complexity rapidly with the number of possible outliers tested.

The notation of chapter 3 will be used, with any extensions specified as the occasion arises. Throughout, the null hypothesis H_o will be that all the observations arise from a single normal population $N(\mu,\sigma^2)$, where both μ and σ may be unknown, or where either or both of μ and σ may be known. This will be tested against various alternatives of the general form that a given number $k > 1$ of the observations are from a normal population with a shift in mean or an increase in variance. The various cases of μ and σ known and unknown, and the case when an external estimate of σ is available will be considered separately.

4.1 Several Upper Outliers

Consider first the case of two upper outliers.

If both μ and σ are known it is easily seen that an appropriate likelihood-based criterion, for the alternative hypothesis that the two largest observations are from a normal distribution with an increased mean, is

$$(x_{(n)}+x_{(n-1)}-2\mu)/\sigma.$$

The null distribution of this statistic may be derived from the joint distribution of $x_{(n-1)}$ and $x_{(n)}$. The case is of little practical importance and the details are omitted. Upper and lower bounds for the percentiles of the distribution may be obtained by noticing that

$$P\,[2Z_{(n-1)} > y] \;<\; P\,[Z_{(n-1)}+Z_{(n)} > y] \;<\; P\,[2Z_{(n)} > y]\,,$$

(writing $Z_{(r)}$ for $(x_{(r)}-\mu)/\sigma$).

It follows that if $k_{(n-1),\alpha}$, $k_{(n),\alpha}$ and k_{α} are the α percentiles of the distributions of $Z_{(n-1)}$, $Z_{(n)}$ and $Z_{(n-1)}+Z_{(n)}$ respectively then the following inequality holds:

$$2k_{(n-1),\alpha} < k_{\alpha} < 2k_{(n),\alpha}\,. \tag{4.1}$$

Consider now the case when σ is known but μ is unknown. It is not difficult to shew by essentially the same argument as used

in section 3.3 that for the alternative hypothesis that the two largest observations arise from a normal population with the same variance as the remainder of the sample, but with an increase in mean, the likelihood-based statistic for testing the two largest observations as outlying is

$$(x_{(n-1)}+x_{(n)}-2\bar{x}_n)/\sigma = u_{n,(n-1)}+u_{n,(n)}.$$

McMillan and David (1971) obtain approximate upper 5% and 1% points, for sample sizes 4(1)27, of this statistic by approximating the integral, over an appropriate region, of the joint density of $(x_{(n-1)}-\bar{x}_{n-2})$ and $\{(n-1)/n\}^{\frac{1}{2}}(x_{(n)}-\bar{x}_{n-1})$ which they obtain by several successive transformations and integrations of the joint density of the n ordered values $x_{(1)},\ldots,x_{(n)}$.

The distribution of $u_{n,(n-1)}+u_{n,(n)}$ may be derived from the joint density of $u_{n,(n-1)}$ and $u_{n,(n)}$, $g_n(.,.)$ say, a recursive formula for which may be obtained by the following probabilistic argument.

$$g_n(u,v)\,du\,dv = P\,[u_{n,(n-1)} \in (u,u+du), u_{n,(n)} \in (v,v+dv)]$$

$$= \sum_{i<j} P\,[u_{n,i} \in (u,u+du), u_{n,j} \in (v,v+dv),$$

$$u_{n,(n-1)} = u_{n,i}, u_{n,(n)} = u_{n,j}]$$

$$= n(n-1)\,P\,[u_{n,n-1} \in (u,u+du), u_{n,n} \in (v,v+dv),$$

$$\max_{i=1}^{n-2} x_i < x_{n-1}]$$

$$= n(n-1)\ P\left[u_{n,n-1} \in (u,u+du), u_{n,n} \in (v,v+dv),\right.$$

$$\left. u_{n-2,(n-2)} < (x_{n-1}-\bar{x}_{n-2})/\sigma\right]$$

$$= n(n-1)\ P\left[u_{n,n-1} \in (u,u+du), u_{n,n} \in (v,v+dv),\right.$$

$$\left. u_{n-2,(n-2)} < ((n-1)u+v)/(n-2)\right]$$

$$= n(n-1)\ P\left[u_{n,n-1} \in (u,u+du), u_{n,n} \in (v,v+dv)\right] \times$$

$$\times\ P\left[u_{n-2,(n-2)} < ((n-1)u+v)/(n-2)\right]$$

(4.2)

(noting that both $u_{n,n-1}$ and $u_{n,n}$ have zero covariance with, and are therefore independent of, $u_{n-2,j}$ for each $j=1,\ldots,n-2$). The first of the two probabilities in (4.2) is $\zeta_n(u,v)du\,dv$, where $\zeta_n(u,v)$ is the joint density of $u_{n,n-1}$ and $u_{n,n}$ at the point (u,v). Now $u_{n,n-1}$ and $u_{n,n}$ have means zero, variances $(n-1)/n$ and covariance $-1/n$, and so have joint density at the point (u,v)

$$(2\pi)^{-1}/\{n/(n-2)\}\exp\left[-\tfrac{1}{2}\{(n-1)(u^2+v^2)+2uv\}/(n-2)\right].$$

The second of the two probabilities in (4.2) is the distribution function of $u_{n-2,(n-2)}$ evaluated at $((n-1)u+v)/(n-2)$. Thus

$$g_n(u,v) = \frac{n(n-1)}{2\pi\sqrt{\{(n-2)/n\}}}\ \exp\left[-\frac{(n-1)u^2+2uv+(n-1)v^2}{2(n-2)}\right] \times$$

$$\times\ A_{n-2}\{((n-1)u+v)/(n-2)\},\quad (4.3)$$

and $g_n(u,v)$ may be evaluated for successive values of n using the results for $A_n(.)$. The results of McMillan and David may be readily derived from (4.3).

Suppose now that both μ and σ are unknown. For the alternative hypothesis, H_1 say, that two observations arise from a common normal distribution with the same variance as the remainder of the sample, but with an increase in mean, the likelihood based statistic for testing the two largest observations as outlying is easily seen to be

$$(x_{(n-1)}+x_{(n)}-2\bar{x}_n)/S_n,$$

which can be written as $U_{(n-1)}+U_{(n)}$. This statistic is essentially different from the statistic

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

proposed by Grubbs (1950). Grubbs (1950) investigated the distribution of this latter statistic and calculated lower 1%, 2·5%, 5% and 10% points for sample sizes 4(1)20. Tietjen and Moore (1972), using Monte Carlo procedures, obtained approximate lower percentage points for the same levels for sample sizes 4(1)20(5)40. Grubbs and Beck (1972) give lower ·1%, ·5%, 1%, 2·5%, 5% and 10% points for sample sizes 4(1)149. It is easy to show that Grubbs' statistic is likelihood-based for the alternative hypothesis, H_2 say, that two observations arise from two separate normal distributions both with the same variance as the remainder of the sample, but with larger means. Grubbs' statistic is

essentially equivalent to

$$U^2_{(n-1)}+2U_{(n-1)}U_{(n)}/(n-1)+U^2_{(n)}.$$

If there is available an unbiased external estimate of σ^2, s^2_ν say, such that νs^2_ν is distributed as χ^2 with ν degrees of freedom independently of the sample under study, then the natural modifications of the likelihood-based criteria are

$$(x_{(n-1)}+x_{(n)}-2\bar{x}_n)/R_n,$$

where $R^2_n = S^2_n+\nu s^2_\nu$, when the alternative hypothesis is H_1, and

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

when it is H_2.

The first of these criteria may be written as

$$V_{(n-1)}+V_{(n)}$$

and the second is essentially equivalent to

$$V^2_{(n-1)}+2V_{(n-1)}V_{(n)}/(n-1)+V^2_{(n)}.$$

The distributions of both of these statistics may be obtained from the joint distribution of

$$V_{(n-1)} = (x_{(n-1)}-\bar{x}_n)/R_n \quad \text{and} \quad V_{(n)} = (x_{(n)}-\bar{x}_n)/R_n,$$

$h_n(.,.)$ say.

Now if $\nu > 0$, $h_n(u,v)$ is clearly zero outside the region R_h^2 defined by

$$0 < u < v < \{(n-1)/n\}^{\frac{1}{2}} \qquad \text{and}$$

$$u^2+2uv/(n-1)+v^2 < (n-2)/(n-1).$$

For $(u,v) \in R_h^2$

$$h_n(u,v)\,du\,dv = P\,[V_{n,(n-1)} \in (u,u+du), V_{n,(n)} \in (v,v+dv)]$$

$$= n(n-1)\,P\,[V_{n,n-1} \in (u,u+du), V_{n,n} \in (v,v+dv),$$

$$V_{n-2,(n-2)} < (x_{n-1}-\bar{x}_{n-2})/R_{n-2}]$$

$$= n(n-1)\,P\,[V_{n,n-1} \in (u,u+du), V_{n,n} \in (v,v+dv),$$

$$V_{n-2,(n-2)} < \{(n-1)u+v\}/\{(n-2)((n-2)-(n-1)u^2-2uv-(n-1)v^2)\}^{\frac{1}{2}}]$$

$$= n(n-1)\,P[V_{n,n-1} \in (u,u+du), V_{n,n} \in (v,v+dv)] \times$$

$$\times\, P[V_{n-2,(n-2)} < \{(n-1)u+v\}/\{(n-2)((n-2)-(n-1)u^2-2uv-(n-1)v^2)\}^{\frac{1}{2}}]. \qquad (4.4)$$

(noting that both $V_{n,n-1}$ and $V_{n,n}$ are independent of $V_{n-2,(n-2)}$).

The first of these two probabilities is $\mu_n(u,v)\,du\,dv$, where $\mu_n(.,.)$ is the joint density of $V_{n,n-1}$ and $V_{n,n}$. This is (Quesenberry and David (1961))

$$\left(\frac{n}{n-2}\right)^{\frac{1}{2}} \frac{n+\nu-3}{2\pi} \left(1 - \frac{n-1}{n-2}u^2 - \frac{2uv}{n-2} - \frac{n-1}{n-2}v^2\right)^{\frac{1}{2}(n+\nu-5)}$$

for $(n-1)u^2+2uv+(n-1)v^2 \leq (n-2)$,

(note the error of sign in the region of support of this distribution given by Quesenberry and David (1961) in their equation (3.3)).

The second of the probabilities in (4.4) is the distribution function of $V_{n-2,(n-2)}, C_{n-2}(.)$.

Thus

$$h_n(u,v) = n(n-1)\{n/(n-2)\}^{\frac{1}{2}}(n+\nu-3)/2\pi\{1-(n-1)u^2/(n-2)-2uv/(n-2)-$$

$$(n-1)v^2/(n-2)\}^{\frac{1}{2}(n+\nu-5)} \times$$

$$\times C_{n-2}\{\{(n-1)u+v\}/\{(n-2)((n-2)-(n-1)u^2-uv-(n-1)v^2)\}^{\frac{1}{2}}\}. \tag{4.5}$$

Now $C_r(x) = 1$ if $x > \sqrt{\{(r-1)/r\}}$, and if

$$2(n-1)u^2+4uv+(n-2)v^2 > n-3, \tag{4.6}$$

then

$$\{(n-1)u+v\}/\{(n-2)((n-2)-(n-1)u^2-2uv-(n-1)v^2)\}^{\frac{1}{2}} > \sqrt{\{(n-3)/(n-2)\}}$$

and so (4.5) becomes

$$h_n(u,v) = n(n-1)\mu_n(u,v) \text{ for } (u,v) \text{ satisfying (4.6).} \qquad (4.7)$$

$$\text{Further } C_r(x) = 1 - \int_x^{\sqrt{\{(r-1)/n\}}} r\psi_r(t)dt$$

if $\sqrt{\{\frac{1}{2}(r-2)/r\}} < x < \sqrt{\{(r-1)r\}}$,

(where $\psi_r(t)$ is as defined in (3.9). If

$$3(n-1)u^2+6uv+(n-3)v^2 > (n-4)$$

and (4.8)

$$2(n-1)u^2+4uv+(n-2)v^2 < n-3$$

then

$$\sqrt{\{\tfrac{1}{2}(n-4)/(n-2)\}} < \{(n-1)u+v\}/\{(n-2)((n-2)-(n-1)u^2-2uv-(n-1)v^2)\}^{\frac{1}{2}} <$$

$$\sqrt{\{(n-3)/(n-2)\}},$$

$$\text{and so } C_{n-2}(t_n) = 1 - \int_{t_n}^{\sqrt{\{(n-3)(n-2)}} \psi_{n-2}(t)dt,$$

(writing t_n for $\{(n-1)u+v\}/\{(n-2)((n-2)-(n-1)u^2-2uv-(n-1)v^2)\}^{\frac{1}{2}}$).

Thus (4.5) becomes

$$h_n(u,v) = n(n-1)\mu_n(u,v)\{1 - \int_{t_n}^{\sqrt{\{(n-3)/(n-2)\}}} \psi_{n-2}(t)dt\}$$

for (u,v) in the region defined by the inequalities (4.8).

It is readily seen, extending the arguments above, that the evaluation of $C_{n-2}(t_n)$ in each of the intervals

$$(\sqrt{\{(n-r-3)/(r+1)(n-2)\}}\ ,\ \sqrt{\{(n-r-2)/r(n-2)\}}) \text{ for } r = 0,1,2,3,\ldots,(n-3)$$

allows the calculation of $h_n(u,v)$, using (4.5), in each of the $(n-2)$ subregions of R_h^2 defined by the pair of inequalities

$$(r+2)(n-1)u^2+2(r+2)uv+(n-r-2)v^2 \geq n-r-3$$

and (4.9)

$$(r+1)(n-1)u^2+2(r+1)uv+(n-r-1)v^2 \leq n-r-2, \text{ for } r = 0,1,2,\ldots,(n-3).$$

It may be noted that if $\nu = 0$, (the case when no external estimate of σ^2 is available), then $C_r(.)$ will reduce to $B_r(.)$ and in particular $h_n(u,v)$ will be zero for $0 < u < \{n(n-1)\}^{-\frac{1}{2}}$ and in the region defined by the last of inequalities (4.9).

The null distribution of $V_{(n-1)}+V_{(n)}$ and $V_{(n-1)}^2+2V_{(n-1)}V_{(n)}/(n-1)+V_{(n)}^2$ may in theory be derived by integration of $h_n(u,v)$ over the appropriate regions. It is easy however to obtain the upper tail of the distribution of $V_{(n-1)}+V_{(n)}$ directly by noting that if $u+v > \sqrt{\{(3n-8)/2n\}}$ then

$$2(n-1)u^2+4uv+(n-2)v^2 > (n-3) \text{ and } (n-2)u^2+4uv+2(n-1)v^2 > n-3,$$

so that if $V_{n-1}+V_n > \sqrt{\{(3n-8)/2n\}}$ then V_{n-1} and V_n must be the largest pair $V_{(n-1)}$ and $V_{(n)}$ (in some order). Hence

$$P\,[V_{(n-1)}+V_{(n)} > K] = \binom{n}{2} P\,[V_{n-1}+V_n > K] \quad \text{if } K > \sqrt{\{(3n-8)/2n\}}.$$

Now R_n^2 may be written as

$$R_n^2 = \{(n-2)/2n\}(x_n+x_{n-1}-2\bar{x}_{n-2})^2+Y+\nu s_\nu^2$$

where Y is a $\sigma^2\chi^2_{n-2}$ variate distributed independently of s_n, x_{n-1} and $\bar{x}_{n-2}$.

Also $x_n+x_{n-1}-2x_n = \frac{n-2}{n}(x_n+x_{n-1}-2\bar{x}_{n-2})$ and so

$$(V_n+V_{n-1})^2 = ((n-2)/n)^2(x_n+x_{n-1}-2\bar{x}_{n-2})^2/\{\{(n-2)/2n\}(x_n+x_{n-1}-2\bar{x}_{n-2})^2+ Y+\nu s_\nu^2\}$$

$$= (2(n-2)/n)t^2_{n+\nu-2}/\{t^2_{n+\nu-2}+n+\nu-2\},$$

where $t_{n+\nu-2}$ denotes a variate following Student's t-distribution with $n+\nu-2$ degrees of freedom.

Thus for $K > \{(3n-8)/2n\}^{\frac{1}{2}}$

$$P\,[V_{(n-1)}+V_{(n)} > K] = \binom{n}{2} P\,[t_{n+\nu-2} > \sqrt{\{K^2n(n+\nu-2)/\{2(n-2)-nK^2\}\}}] \tag{4.10}$$

This is an extension of the result of McMillan (1971).

The extension of the above methods and results to the case of

three, and more, upper outliers, is straightforward. In the case when σ is known and μ unknown the likelihood-based statistic for testing the k largest observations as outlying, for the alternative hypothesis that k observations in the sample arise from a normal distribution with variance σ^2 and mean $\mu_1 > \mu$, is

$$\sum_{i=1}^{k} u_{n',(n-i+1)}. \tag{4.11}$$

The distribution of this statistic may be derived from the joint distribution of $(u_{n,(n-k+1)}, u_{n,(n-k+2)}, \ldots, u_{n,(n)})$, $g_{n,k}(.,.,\ldots,.)$ say. A probabilistic argument essentially identical to that described in the case of two outliers gives

$$g_{n,k}(u_1,\ldots,u_k) = (n)_k \phi_{n,k}(u_1,\ldots,u_k) A_{n-k}\{((n-k+1)u_1+u_2+\ldots+u_k)/(n-k)\} \tag{4.12}$$

where $\phi_{n,k}(u_1,\ldots,u_k)$ is the density function at the point $(u_1,\ldots,u_k)$ of a k-dimensional multivariate normal random variable with mean zero and variance-covariance matrix $\{I_k - n^{-1} 1_k\}$, here 1_k denotes the $k \times k$ matrix all of whose elements are 1.

Suppose now that both μ and σ are unknown. For the alternative hypothesis, H_1 say, that k observations arise from a normal distribution $N(\mu_1,\sigma^2)$ where $\mu_1 > \mu$ the likelihood-based statistic for testing simultaneously $x_{(n-k+1)},\ldots,x_{(n)}$ is $\sum_{i=1}^{k} U_{(n-i+1)}$. For the alternative hypothesis, H_2 say, that k observations arise from k separate normal distributions each with variance σ^2 and means greater than μ the likelihood-based test criterion is the Grubbs type statistic

$$\sum_{i=1}^{n-k}(x_i-\bar{x}_{n-k})^2/S_n^2.$$

Approximate upper 1%, 2·5%, 5% and 10% points of this latter statistic were obtained by Tietjen and Moore (1972), using Monte Carlo procedures, for k = 1(1)10 and n = 2k(1)20(5)40. In situations where an external estimate based on ν degrees of freedom of σ^2 is available, s_ν^2 say, the modifications of these two statistics, utilising the extra information, are $\sum_{i=1}^{k}V_{(n-i+1)}$ and $\sum_{i=1}^{k}(x_{(i)}-\bar{x}_{n-k})^2/R_n^2$ respectively. The null distributions of both of these statistics may be derived from the joint density function of $(V_{(n-k+1)},\ldots,V_{(n)})$, $h_{n,k}(.,.,\ldots,.)$ say.

Applications of the two identities

$$x_{n-k+1}-\bar{x}_{n-k} = (n-k)^{-1}\{(n-k)(x_{n-k+1}-\bar{x}_n) + \sum_{i=1}^{k}(x_{n-i+1}-\bar{x}_n)\}$$

and (4.13)

$$R_{n-k}^2 = R_n^2 - \sum_{i=1}^{k}(x_{n-i+1}-\bar{x}_n)^2-(n-k)^{-1}\{\sum_{i=1}^{k}(x_{n-i+1}-\bar{x}_n)\}^2$$

(the latter is a generalisation of identity (3.4)) and an extension of the probabilistic argument, described earlier in the case of two outliers, give an expression for the joint density $h_{n,k}(.,.,\ldots,.)$. For $(u_1,\ldots,u_k) \in R_h^k$, the region defined by

$$0 < u_1 < u_2 < \ldots < u_k < \{(n-1)/n\}^{\frac{1}{2}} \quad \text{and}$$

$$(n-k)\textstyle\sum_1^k u_i + \{\sum_1^k u_i\}^2 < (n-k),$$

$$h_{n,k}(u_1,\ldots,u_k) = (n)_k \mu_{n,k}(u_1,\ldots,u_k) \times$$

$$\times C_{n-k}\{\{(n-k)u_1+\sum_{i=1}^{k}u_i\}/\{(n-k)^2(1-\sum_{i=1}^{k}u^2-(n-k)^{-1}\{\sum_{i=1}^{k}u_i\}^2)\}^{\frac{1}{2}}\}$$

(4.14)

where $\mu_{n,k}(.,\ldots,.)$ is the joint density of $(V_{n,n-k+1},\ldots,V_{n,n})$ and $C_{n-k}\{.\}$ is the distribution function of $V_{n-k,(n-k)}$.

The joint density $\mu_{n,k}(.,\ldots,.)$ of $(V_{n,n-k+1},\ldots,V_{n,n})$ may be derived by a generalisation to k variables of the method used by Quesenberry and David (1961) in the case of two variables. Consider first the joint density, $\mu^*_{n,k}(.,\ldots,.)$ say, of $(V_{n-k+1},\ldots,V_{n,n})$. The density function of $V_{n-k+j,n-k+j}$ is, putting r = n-k+j in (3.9),

$$\psi_{n-k+j}(u) = \left(\frac{n-k+j}{\pi(n-k+j-1)}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n+\nu-k+j-1))}{\Gamma(\frac{1}{2}(n+\nu-k+j-2))} \times$$

$$\times \{1-(n-k+j)u^2/(n-k+j-1)\}^{\frac{1}{2}(n+\nu-k+j-4)},$$

$$\text{for } |u| \leq \sqrt{\{(n-k+j-1)/(n-k+j)\}}.$$

Since the $V_{n-k+j,n-k+j}$, (j=1,2,...,k) are mutually independent (Quesenberry and David (1961))

$$\mu^*_{n,k}(u_1,\ldots,u_k) = \prod_{j=1}^{k}\psi_{n-k+j}(u_j)$$

$$= \pi^{-\frac{1}{2}k} \left(\frac{n}{n-k}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n+\nu-1))}{\Gamma(\frac{1}{2}(n+\nu-k-1))} \times$$

$$\times \prod_{j=1}^{k} \{1-(n-k+j)u_j^2((n-k+j-1)\}^{\frac{1}{2}(n+\nu-k+j-4)},$$

$$\text{for } |u| \leq \sqrt{\{(n-k+j-1)/(n-k+j)\}}, j=1,\ldots,k. \qquad (4.15).$$

$$\text{Now } V_{n-k+j,r} = \frac{\{V_{n-k+k+1,r} + V_{n-k+j+1,n-k+j+1}/(n-k+j)\}}{\{1-(n-k+j+1)V^2_{n-k+j+1,n-k+j+1}/(n-k+j)\}^{\frac{1}{2}}}$$

$$\text{for } r \leq n-k+j, j=1,2,\ldots,(k-1).$$

Repeated application of this transformation to (4.15) for each r in the range $n-k+1 \leq r \leq n-k+j$ for each $j=1,2,\ldots,(k-1)$ gives

$$\mu_{n,k}(u_1,\ldots,u_k) = \pi^{-\frac{1}{2}k} \left(\frac{n}{n-k}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}(n+\nu-1))}{\Gamma(\frac{1}{2}(n+\nu-k-1))} \times$$

$$\times \{1-\sum_{1}^{k} u_i^2-(n-k)^{-1}\{\sum_{1}^{k} u_i\}^2\}^{\frac{1}{2}(n+\nu-k-3)}$$

over the region $\sum_{i=1}^{k} u_i^2+(n-k)^{-1}\{\sum_{i=1}^{k} u_i\}^2 \leq 1.$ (4.16)

Evaluation of $C_{n-k}\{.\}$ in each of the intervals

$$(\sqrt{\{(n-k-r-1)/(r+1)(n-k)\}} \ , \ \sqrt{\{(n-k-r)/r(n-k)\}}) \text{ for } r=0,1,2,.,n-k$$

allows the calculation of $h_{n,k}(u_1,\ldots,u_k)$ using (4.14) in each of the (n-k) subregions of R_h^k defined by the pair of inequalities

$$(r+1)(n-k)u_1^2+2(r+1)u_1\sum_1^k u_i+(n-k-r-1)\sum_1^k u_i+\{\sum_1^k u_i\}^2 \gtrless (n-k-r-1)$$

and (4.17)

$$r(n-k)u_1^2+2ru_1\sum_1^k u_i+(n-k-r)\sum_1^k u_i^2+\{\sum_1^k u_i\}^2 \lesseqgtr (n-k-r).$$

It may be noted that in the case when no external estimate of σ^2 is available $\nu=0$ and $h_{n,k}(u_1,\ldots,u_k)$ will be zero for $0 < u_1 < \{n(n-1)\}^{-\frac{1}{2}}$ and in the region defined by the last of inequalities (4.17).

Of particular interest is the subregion of R_h^k defined by the first of the inequalities (4.17)). If

$$(n-k)u_1^2+2u_1\sum_1^k u_i+(n-k-1)\sum_1^k u_i+\{\sum_1^k u_i\}^2 > n-k-1. \qquad (4.18)$$

then $t_{n,k} > \sqrt{\{(n-k-1)/(n-k)\}}$, where

$$t_{n,k}\ \{(n-k)u_1+\sum_1^k u_i\}/\{(n-k)^2(1-\sum_1^k u_i^2-(n-k)^{-1}\{\sum_1^k u_1\}^2)\}^{\frac{1}{2}},$$

and so $C_{n-k}\{t_{n,k}\} = 1$. Thus (4.14) becomes

$$h_{n,k}(u_1,\ldots,u_k) = (n)_k \mu_{n,k}(u_1,\ldots,u_k)$$

for $(u_1,\ldots,u_k)$ satisfying (4.18). The upper tail of the distribution of $V_{(n-k+1)}+V_{(n-k+2)}+,\ldots,+V_{(n)}$ may be derived directly from this result, but it is easier to note that if

$$(n-k)V_{n,n-k+j}+2V_{n,n-k+j}\sum_{i=1}^k V_{n,n-k+i}+(n-k-1)\sum_{i=1}^k V_{n,n-k+i}^2$$

$$+\{\sum_{i=1}^k V_{n,n-k+i}\}^2 \gtrless (n-k-1) \text{ for each } j=1,2,\ldots,k, \qquad (4.19)$$

then $(V_{n,n-k+1},\ldots,V_{n,n})$ must be the set consisting of the k largest $V_{n,r}$, $(V_{n,(n-k+1)},\ldots,V_{n,(n)})$, in some order. Further, since the hyperplane

$$\sum_{i=1}^{k} u_i = \sqrt{\{((2k-1)n-2k^2)/2n\}}$$

is tangential to each of the surfaces

$$(n-k)u_j+2u_j\sum_{i=1}^{k}u_i+(n-k-1)\sum_{i=1}^{k}u_i^2+\{\sum_{i=1}^{k}u_i\}^2 = n-k-1, j=1,2,\ldots,k,$$

each of which encloses a convex region, it follows that if

$$V_{n,n-k+1}+\ldots+V_{n,n} > \sqrt{\{((2k-1)n-2k^2)/2n\}}$$

then each of the inequalities (4.10) will be satisfied.

Thus if

$$K \geq \sqrt{\{((2k-1)n-2k^2)/2n\}}, \quad \text{then}$$

$$P\,[V_{(n-k+1)}+\ldots+V_{(n)} > K] = \binom{n}{k} P\,[V_{n,n-k+1}+\ldots+V_{n,n} > K].$$

Now R_n^2 may be written as

$$R_n^2 = \{(k-1)(n-k)/nk\}\{\sum_{i=1}^{k}(x_{n-k+i}-\bar{x}_{n-k})\}^2+Y+\nu s_\nu^2,$$

where Y is a $\sigma^2\chi^2_{n-k}$ variate distributed independently of $x_{n-k+1},\ldots,x_n$ and $\bar{x}_{n-k}$. Also

$$\sum_{i=1}^{k}(x_{n-i+1}-\bar{x}_n) = \{(n-k)/n\}\sum_{i=1}^{k}(x_{n-i+1}-\bar{x}_{n-k})$$

and so $\{\sum_{i=1}^{k}V_{n,n-k+i}\}^2$ may be written as

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

$$\{\{(k-1)(n-k)/nk\}\{\sum_{i=1}^{k}(x_{n-k+1}-\bar{x}_{n-k})\}^2+Y+\nu s_\nu^2\}$$

$$= k(n-k)n^{-1}t^2_{n+\nu-k}/\{(k-1)t^2_{n+\nu-k}+n+\nu-k\},$$

where $t_{n+\nu-k}$ is a variate following Student's t-distribution with $n+\nu-k$ degrees of freedom.

Thus for $K \geq \sqrt{\{((2k-1)n-2k^2)/2n\}}$

$$P\left[\sum_{i=1}^{k}V_{(n-k+i)} > K\right] = \binom{n}{k}P\left[t_{n+\nu-k} > \{nK^2(n+\nu-k)/\{k(n-k)-n(k-1)K^2\}\}^{\frac{1}{2}}\right]. \tag{4.20}$$

It is of interest to note that a similar result is not available for the Grubbs type statistic $\sum_{i=1}^{k}(x_{(i)}-\bar{x}_{n-k})^2/R_n^2$; that is there is no constant K_o such that for $K > K_o$

$$P\left[\sum_{i=1}^{k}(x_{(i)}-\bar{x}_{n-k})^2/R_n^2 < K\right] = \binom{n}{k}P\left[\sum_{i=1}^{k}(x_i-\bar{x}_{n-k})^2/R_n^2 < K\right],$$

unless both sides of the equation are zero. To see this first note that the Grubbs type statistic is equivalent to the statistic

$$(n-k)\sum_{i=1}^{k}V^2_{(n-k+i)}+\{\sum_{i=1}^{k}V_{(n-k+i)}\}^2,$$

(this follows from the second of identities (4.13)). Further, each of the surfaces

$$(n-k)u_j+2u_j\sum_1^k u_i+(n-k-1)\sum_1^k u_i^2+\{\sum_1^k u_i\}^2 = n-k-1$$

is tangential to the surface

$$(n-k)\sum_1^k u_i+\{\sum_1^k u_i\}^2 = K^*, \quad \text{if } K^* = (n-k),$$

and wholly encloses that surface if $K^* < (n-k)$.

It follows that if K_o is such that

the inequality $(n-k)\sum_{i=1}^k V^2_{n,n-k+i}+\{\sum_{i=1}^k V_{n,n-k+i}\}^2 > K_o$ implies that the set $(V_{n,n-k+1},\ldots,V_{n,n})$ must be the set consisting of the k largest $V_{n,r}$, $(V_{n,(n-k+i)},\ldots,V_{n,(n)})$, then $K_o \geq (n-k)$.

However $P\,[(n-k)\sum_{i=1}^k V^2_{n,n-k+i}+\{\sum_{i=1}^k V_{n,n-k+i}\}^2 > (n-k)] = 0,$

since $h_{n,k}(u_1,\ldots,u_n)$ is zero outside the region R_h^k, and the result follows.

4.2 An Upper and a Lower Outlier

In many situations it may be desired to test both the largest and the smallest observations of a sample as outlying. The problems associated with applying a sequential procedure to such a situation are rather more complex than those encountered when testing a group of possible outliers at one end of the sample. The phenomenon of masking is highlighted in the studies, by Irwin (1925), Grubbs (1950) and Tietjen and Moore (1972), of the classical set of data relating to the observations made by Lieutenant Herndon in 1846 of the vertical semi-diameter of Venus discussed originally by Chauvenet (1863), and reproduced in Table 4.1.

Table 4.1

-1·40	-0·44	-0·30	-0·24	-0·22
-0·13	-0·05	-0·06	0·10	0·18
0·20	0·39	0·48	0·63	1·01

Both Irwin and Grubbs, testing the observations sequentially, declare the lowest observation, -1·40, but not the highest, 1·01, to be outlying, at the 5% level of significance. However Tietjen and Moore, testing both observations simultaneously, declare the pair to be outlying at the 5% level of significance. Chauvenet, although using sequential procedures, based upon a variety of different criteria, also declared both observations to be outlying, but the significance levels of the tests he used are considerably higher than 5%.

The calculation of the significance probabilities of two (or

even more) outliers detected 'sequentially', not all lying at the same end of the sample, appears to be exceedingly complex and no study has been made of this problem in the literature.

These difficulties are avoided by using a criterion which tests both observations simultaneously. Grubbs (1950) proposed the use of the statistic

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

(where $\bar{x}_{r,s} = \frac{1}{s-r+1}\sum_{i=r}^{s} x_{(i)}, s > r$)

but did not obtain any distributional results for it. Tietjen and Moore (1972) obtain approximate percentage points, using Monte Carlo procedures, for a variety of outlier testing criteria of the Grubbs type, but not specifically for this one. The statistic they use in their analysis of the data of Chauvenet makes some allowance for the possibility that the most extreme pair of observations may occur at the same end or at different ends of the sample, though exactly how much allowance is made is not clear. This point will be discussed further in the next section.

It is easily seen that the statistic proposed by Grubbs is essentially equivalent to the likelihood ratio statistic for testing the null hypothesis, H_o, that all observations arise from the same normal population $N(\mu,\sigma^2)$, where both μ and σ are unknown, against the alternative hypothesis, H_1, that all but two observations arise from the same normal population $N(\mu,\sigma^2)$ and the two remaining observations arise from two distinct normal populations $N(\mu_1,\sigma^2)$

and $N(\mu_2,\sigma^2)$, each with the same variance as the rest of the sample, but with $\mu_1 < \mu < \mu_2$. Since

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

Grubbs' statistic is equivalent to

$$U_{(1)}^2+2U_{(1)}U_{(n)}/(n-1)+U_{(n)}^2.$$

In cases when there is available an external estimate of σ^2, s_ν^2 say, where νs_ν^2 is distributed as $\sigma^2\chi_\nu^2$, then this statistic may be modified to

$$V_{(1)}^2+2V_{(1)}V_{(n)}/(n-1)+V_{(n)}^2.$$

In the case when σ^2 is known exactly then a similar argument shews that an appropriate likelihood-based statistic (for testing H_o against H_1) is

$$u_{(1)}^2+2u_{(1)}u_{(n)}/(n-1)+u_{(n)}^2.$$

The null distribution of each of these statistics in principle may be derived from the joint distributions of $(U_{(1)},U_{(n)})$, $(V_{(1)},V_{(n)})$ or $(u_{(1)},u_{(n)})$ as appropriate.

Consider in particular the joint distribution of $(V_{(1)},V_{(n)})$. Suppose $(V_{(1)},V_{(n)})$ has joint density function $\ell_n(.,.)$ and joint

distribution function $L_n(.,.)$. $\ell_n(u,v)$ is clearly zero outside the region R^2_ℓ defined by

$$u \leq 0 \leq v,$$

$$u^2+2uv/(n-1)+v^2 < (n-2)/(n-1),$$

(if $v=0$, then $\ell_n(u,v)$ will also be zero unless $\pm\{n(n-1)\}^{-\frac{1}{2}} < u < v < \{n(n-1)\}^{-\frac{1}{2}}$).

Now for $(u,v) \in R^2_\ell$

$$\ell_n(u,v)dudv = P\,[V_{(1)} \in (u,u+du), V_{(n)} \in (v,v+dv)]$$

$$= n(n-1)\,P\,[V_{n,n-1} \in (u,u+du), V_{n,n} \in (v,v+dv),$$

$$(x_{n-1}-\bar{x}_{n-2})/R_{n-2} < V_{n-2,(1)} < V_{n-2,(n-2)} < (x_n-\bar{x}_{n-2})/R_{n-2}]$$

$$= n(n-1)\,P\,[V_{n,n-1} \in (u,u+du), V_{n,n} \in (v,v+dv)] \times$$

$$\times\; P\,[\{(n-1)u+v\}/\{(n-2)((n-2)-(n-1)u^2-2uv-(n-1)v^2)\}^{\frac{1}{2}} <$$

$$V_{n-2,(1)} < V_{n-2,(n-2)} < \{u+(n-1)v\}/\{(n-2)((n-2)-(n-1)u^2-2uv-(n-1)v)\}^{\frac{1}{2}}] \tag{4.21}$$

(noting that both $V_{n,n-1}$ and $V_{n,n}$ are independent of $V_{n-2,(n-2)}$).

Now the second of the two probabilities in (4.21) may be written as

$$P\,[V_{n-2,(n-2)} < \{u+(n-1)v\}/\xi_n(u,v)] -$$

$$P\,[V_{n-2,(1)} < \{(n-1)u+v\}/\xi_n(u,v), V_{n-2,(n-2)} < \{u+(n-1)v\}/\xi_n(u,v)]$$

(writing $\xi_n(u,v)$ for $\{(n-2)((n-2)-(n-1)u^2-2uv-(n-1)v^2)\}^{\frac{1}{2}}$)

$$= C_{n-2}\,[\{u+(n-1)v\}/\xi_n(u,v)] - L_{n-2}\,[\{(n-1)u+v\}/\xi_n(u,v), \{u+(n-1)v\}/\xi_n(u,v)].$$

Thus

$$\ell_n(u,v) = n(n-1)\mu_n(u,v)C_{n-2}\,[\{u+(n-1)v\}/\xi_n(u,v)] - L_{n-2}\,[\{(n-1)u+v\}/\xi_n(u,v), \{u+(n-1)v\}/\xi_n(u,v)]. \tag{4.22}$$

Now since $\ell_n(x,y)$ is zero outside the region R_ℓ^2 and since

$$L_n(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \ell_n(\xi,\eta)\,d\xi\,d\eta$$

$L_n(x,y)$ is zero either if $x < -\{(n-1)/n\}^{\frac{1}{2}}$ or if $y < 0$. Thus

$$L_{n-2}\,[\{(n-1)u+v\}/\xi_n(u,v), \{u+(n-1)v\}/\xi_n(u,v)] = 0$$

if $\{(n-1)u+v\}/\xi_n(u,v) < -\{(n-3)/(n-2)\}^{\frac{1}{2}}$,

i.e. if $2(n-1)u^2-4uv+(n-2)v^2 > n-3$ and $(n-1)u+v < 0$. (4.23)

Hence

$$\ell_n(u,v) = n(n-1)\mu_n(u,v)C_{n-2}\,[\{u+(n-1)v\}/\xi_n(u,v)] \qquad (4.24)$$

for $(u,v) \in R_\ell^2$ satisfying either inequalities (4.23), or $u+(n-1)v < 0$.

Further $C_n(x) = 1$ if $x > \{(n-1)/n\}^{\frac{1}{2}}$ and so

$$C_{n-2}[\{u+(n-1)v\}/\xi_n(u,v)] = 1$$

if

$$(n-2)u^2+4uv+2(n-1)v^2 > n-3 \quad \text{and} \quad u+(n-1)v > 0, \qquad (4.25)$$

and (4.22) becomes

$$\ell_n(u,v) = n(n-1)\mu_n(u,v) \qquad (4.26)$$

for $(u,v) \in R_\ell^2$ satisfying inequalities (4.23) and (4.25).

Application of a similar method to the joint density and distribution functions of $(u_{(1)},u_{(n)})$, $m_n(.,.)$ and $M_n(.,.)$ respectively say, gives the recursive formula

$$m_n(u,v) = n(n-1)(2\pi)^{-1}\{n/(n-2)\}^{\frac{1}{2}}\exp\{-\tfrac{1}{2}\{(n-1)(u^2+v^2)+2uv\}/n-2\} \times$$

$$\times\ [A_{n-2}\{u+(n-1)v\}/(n-2)-M_{n-2}\,[\{(n-1)u+v\}/(n-2),\{u+(n-1)v\}/(n-2)]]$$

$$\text{for } u < 0 < v. \qquad (4.27)$$

Since $M_n(x,y) = 0$ if $y < 0$

$$m_n(u,v) = n(n-1)(2\pi)^{-1}\{n((n-2)\}^{\frac{1}{2}}\exp\{-\tfrac{1}{2}(n-1)(u^2+v^2)+2uv\}/(n-2)\} \times$$

$$\times A_{n-2}[\{u+(n-1)v\}/(n-2)]$$

for $u+(n-1)v < 0$. (4.28)

The similarities with equations (4.22) and (4.24) are clear. However since $(u_{(1)},u_{(n)})$ can take values over the entire quarter plane $u < 0 < v$ there is no counterpart of equation (4.26).

The null distribution of any outlier detecting criterion, based upon the two extreme studentized (or standardised) deviations from the mean, may be obtained by the integration of the joint density $\ell_n(u,v)$ (or $m_n(u,v)$) over the appropriate regions.

In the case of statistics of the Grubbs type an argument similar to that given in the preceding section, for the case of Grubbs type statistics for testing upper outliers, shews that it is not possible to obtain even the tail of the distribution in the simple form of equation (4.20).

4.2.1 The Studentized Range

Another statistic which may be used for testing simultaneously the upper and lower observations of a sample as outlying is, in the case when σ is unknown, $V_{(n)}-V_{(1)}$, and in the case when σ is known, $u_{(n)}-u_{(1)}$. This statistic is essentially the studentized, or standardised, range of the sample. Although this statistic is not likelihood-based, for any appropriate alternative hypothesis,

it does possess considerable intuitive appeal; its value will be inflated when both $x_{(1)}$ and $x_{(n)}$ are outliers, it is independent of both the location and scale of measurement used, and it may be considered as a generalisation of the likelihood-based statistic for testing two possible outliers at the same end of the same, $V_{(n)}+V_{(n-1)}$, since both statistics are the sums of the studentized absolute deviations of the queried observations from the mean of the sample. The use of the statistic $u_{(n)}-u_{(1)}$ was proposed by Student (1927) for testing the single most extreme observation as outlying, rather than the upper and lower extreme values simultaneously. It is interesting to note that Student advocates what is basically a sequential procedure for testing for outliers, though the particular situation considered by Student allowed the replacement of rejected outliers by new observations before each repetition of the test, in essence a 'topping-up' procedure.

The distribution of $(n-1)^{\frac{1}{2}}(V_{(n)}-V_{(1)})$, the studentized range, in the particular case $\nu=0$ was considered by David, Hartley and Pearson (1954), and Pearson and Stephens (1964), though not in the context of outlier detection. They obtain the upper tail of the distribution for small samples and provide approximate upper and lower 0·5%, 1%, 2·5%, 5% and 10% points for sample sizes 3(1)20(5)100,150,200,500,1000. These are reproduced in Table 29c of Biometrika Tables for Statisticians vol 1.

The work of David, Hartley and Pearson relating to the upper tail of the distribution may be extended to the cases $\nu\neq0$ using the results given above. Now if all four of the following inequalities hold

$$2(n-1)V_{n,1}^2+4V_{n,1}V_{n,n}+(n-2)V_{n,n}^2 > n-3, \qquad (n-1)V_{n,1}+V_{n,n} < 0,$$

$$(n-2)V_{n,n}^2+4V_{n,1}V_{n,n}^2+2(n-1)V_{n,n} > n-3 \quad \text{and} \quad V_{n,1}+(n-1)V_{n,n} > 0 \tag{4.29}$$

then (4.26) shews that $(V_{n,1},V_{n,n})$ must be the pair $(V_{n,(1)},V_{n,(n)})$, in that order. Further the line $v-u = 3/2$ is tangential to both of the ellipses

$$2(n-1)u^2+4uv+(n-2)v^2 = n-3$$

and

$$(n-2)u^2+4uv+2(n-1)v^2 = n-3,$$

and so if $V_{n,n}-V_{n,1} > 3/2$ then inequalities (4.29) will be satisfied. (It should be noted that necessarily $(V_{n,1},V_{n,n})$ must satisfy the inequality $(n-1)V_{n,1}^2+2V_{n,1}^2V_{n,n}+(n-1)V_{n,n} < (n-2)$ and so $V_{n,n}-V_{n,1} > 3/2$ implies that both the following inequalities hold; $(n-1)V_{n,1}+V_{n,n} < 0$ and $V_{n,1}+(n-1)V_{n,n} > 0$).
It follows that for $K \geq 3/2$

$$P\,[V_{n,(n)}-V_{n,(1)} > K] = n(n-1)\,P\,[V_{n,n}-V_{n,1} > K]\,. \tag{4.30}$$

In the case $\nu=0$ this corresponds to the result of David, Hartley and Pearson (1954).

The probability on the right hand side of (4.30) may be derived by an extension of the method used by David, Hartley and Pearson (1954). R_n^2 may be written as

$$R_n^2 = \tfrac{1}{2}(x_n-x_1)^2+Y+\nu s_\nu^2$$

where Y is distributed as $\sigma^2\chi^2$ with n-2 degrees of freedom independently of (x_n-x_1). It follows that $(V_{n,n}-V_{n,1})^2$ may be written as

$$(x_n-x_1)^2/\{\tfrac{1}{2}(x_n-x_1)^2+Y+\nu s_\nu^2\}$$

$$= 2t^2_{n+\nu-2}/\{t^2_{n+\nu-2}+(n+\nu-2)\}$$

where $t_{n+\nu-2}$ is a variate following Student's t-distribution with $n+\nu-2$ degrees of freedom.

Thus for $K > 3/2$

$$P\,[V_{n,(n)}-V_{n,(1)} > K] = n(n-1)\,P\,[t_{n+\nu-2} > (n+\nu-2)^{\frac{1}{2}}K/\{2-K^2\}^{\frac{1}{2}}]. \quad (4.31)$$

Putting $\nu=0$ retrieves the result of David, Hartley and Pearson.

It may be noted that since the line

$$v-u = 2$$

is tangential to the ellipse

$$(n-1)u^2+2uv+(n-1)v^2 = n-2,$$

it follows that $V_{n,(n)}-V_{n,(1)} \leq 2$ for all samples and so

$$P\,[V_{n,(n)}-V_{n,(1)} > 2] = 0.$$

This is a generalisation to the cases $\nu\neq 0$ of a result of Thomson (1955).

4.2.2 An Example

Consider again the data in table 4.1 discussed earlier. Here, with 15 observations, the range of the sample is 1·01+1·40 = 2·41, and R_{15} = 4·296 (with ν=0), giving a value of $V_{(15)}-V_{(1)}$ of 1·168, or equivalently a value of the studentized range of 4·37. The upper 2½% and 1% points of this statistic are 4·29 and 4·44 respectively (Table 29c Biometrika Tables), so that the pair -1·40 and 1·01 are declared outliers at the 2½% level of significance. The apparent discrepancy with the result of Tietjen and Moore (1972) is explained in part by the partial allowance made in the latters' analysis for not specifying in advance which ends of the sample the outliers may occur.

It is interesting to note that if only the most extreme observation, in this case -1·40, is rejected and the test repeated on the reduced sample of 14 observations a value of 3·61 is obtained for the studentized range, which is less than even the 10% point 3·95, illustrating again the phenomenon of masking. It is curious that were instead the maximum value 1·01 to be omitted and the test repeated on the reduced sample of 14 observations then a value of the studentized range of 4·094 would be obtained, which falls between the 5% and 2½% points 4·09 and 4·21, indicating that both the observations -1·40 and 1·01 may be considered as outlying.

This illustrates the difficulties encountered in using a statistic based upon the range of the sample for testing only one observation as outlying (as suggested by Student) rather than considering such a criterion as testing simultaneously both the largest and the smallest observations.

4.3 Outliers at Unspecified Ends of the Sample

4.3.1 The case μ unknown

The criteria and tests discussed in the preceding sections are designed to test as outlying observations at a specified end (section 4.1) or ends (section 4.2) of the sample. That is the null hypothesis H_o (that all observations arise from a normal population $N(\mu,\sigma^2)$) was tested against an alternative of the form that k of the observations in the sample come from normal distributions $N(\mu_i,\sigma^2)(i=1,\ldots,k)$ where it is specified in advance whether, for each i, $\mu_i > \mu$ or $\mu_i < \mu$. This corresponds to the 'one-sided' case of chapter 3. Corresponding to the 'two-sided' case discussed in section 3.6, it is possible to consider multiple outliers at unspecified ends of the sample and obtain likelihood-based criteria for testing the most 'extreme' set of k observations as outlying. The k observations declared to be the most extreme set of size k will be those whose omission from the sample produces the largest increase in the likelihood of the sample. It is clear that this is a generalisation of the case of a single outlier discussed in (3.6). For example in the case of two outliers the criterion is

$$\{\min\{\sum_{i=1}^{n-2}(x_{(i)}-\bar{x}_{n-2})^2,\sum_{i=3}^{n}(x_{(i)}-\bar{x}_{3,n})^2,\sum_{i=2}^{n-1}(x_{(i)}-\bar{x}_{2,n})^2\}\}/\sum_{i=1}^{n}(x_{(i)}-\bar{x}_n)^2, \quad (4.32)$$

and the two observations actually tested as outlying are those corresponding to that sum of squares out of the three above which is found to be the smallest. The extension to two-sided criteria for

testing k possible outliers involves considering (k+1) sets of possible outliers at specified ends of the sample, deciding which of these k+1 sets is the most 'extreme', in the sense outlined above, and then testing that set, using as test criterion the minimum of the k+1 separate test criteria.

It must be noted that this method of identifying possible outliers is essentially different from that proposed by Tietjen and Moore (1972). They identify as the k possible outliers to be tested that set consisting of the k observations whose absolute distances from the mean of the whole sample are the greatest. Take for example k=2 and consider a situation where all but four observations in a sample of size n have value 0, and the remaining four values are $(-1+\varepsilon),1,(1+\varepsilon),2$ (where $0 < \varepsilon < 1$), and suppose that $(1+3\varepsilon)/\varepsilon < n < (3+2\varepsilon)/\varepsilon$. Then $\bar{x}_n = (3+2\varepsilon)/n$, $(1+\varepsilon)-\bar{x}_n = (n+n\varepsilon-3-2\varepsilon)/n$ and $\bar{x}_n-(-1+\varepsilon) = (3+2\varepsilon+n-n\varepsilon)/n$, and so

$$\bar{x}_n-(-1+\varepsilon) > (1+\varepsilon)-\bar{x}_n, \quad \text{(since } n < (3+2\varepsilon)/\varepsilon\text{)}.$$

Also
$$\sum_{i=1}^{n-2}(x_{(i)}-\bar{x}_{n-2})^2 = 1+(1-\varepsilon)^2-\varepsilon^2/(n-2)$$

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

and so
$$\sum_{i=1}^{n-2}(x_{(i)}-\bar{x}_{n-2})^2 < \sum(x_{(i)}-\bar{x}_{2,n-1})^2$$

(since $n > (1+3\varepsilon)/\varepsilon$).

Thus the two values selected as possible outliers by Tietjen and Moore's procedure are 2 and $(-1+\varepsilon)$, while the pair of values

selected by the likelihood-based method described above are 2 and $(1+\varepsilon)$. Of course on purely intuitive grounds the latter pair might be considered more 'extreme' than the former since the observation $(1+\varepsilon)$ lies further from the main body of the sample than the observation $(-1+\varepsilon)$. It may be noted that a sequential procedure would select initially the value 2 and then the value $(1+\varepsilon)$.

The above example shews that the statistic E_2 used by Tietjen and Moore (1972), in their analysis of the data of 4·1, is not as they assert, the same as the statistic (4.31). Further, the example illustrates that if it is desired to use the tables of Tietjen and Moore (1972) it is important to ensure that the k observations tested as outlying are indeed those which lie furthest in absolute distance from the mean of the whole sample, rather than that set of k observations which on purely subjective grounds appears to be the most 'extreme'.

The null distributions of the likelihood-based two-sided criteria are extremely complicated and it would seem that a Monte Carlo study would be more productive than a direct algebraic approach.

4.3.2 The case μ known

In cases when the mean of the population is known, but the variance σ^2 is unknown, it may be desired to test the null hypothesis H_o against an alternative of the form that all but k observations come from the same normal population $N(\mu,\sigma^2)$, and k observations come from a normal population $N(\mu,\sigma_1^2)$ where $\sigma_1^2 > \sigma^2$. In such cases there will be no *a priori* reasons for specifying the ends of the sample at which outliers may occur.

It is easy to shew that an appropriate likelihood-based statistic for testing the k observations furthest in absolute distance from the mean μ, $(x^*_{(n-k+1)},\ldots,x^*_{(n)})$ say, is

$$Z_{n,(k)} = \sum_{i=1}^{k}(x_{(n-k+i)}-\mu)^2 / \sum_{i=1}^{n}(x_i-\mu)^2,$$

which, under the null hypothesis H_o, is the ratio of the sum of the k largest of a set of gamma variates with degrees of freedom parameter $\frac{1}{2}$ to the sum of those variates. Putting $r=\frac{1}{2}$ in (2.32) gives the following inequality for the upper tail probability of $Z_{n,(k)}$:

$$P\,[Z_{n,(k)} > u] < \binom{n}{k} P\,[F_{k,n-k} > \{(n-k)/k\}\{u/(1-u)\}]\ , \tag{4.33}$$

where $F_{k,n-k}$ is an F-variate with (k,n-k) degrees of freedom.

Chapter 5

Outliers in Linear Models

The study of outliers in univariate samples, the subject of the previous three chapters, may be considered as a special, though important, case of more general situations. This chapter and the next are concerned with two such generalisations; this chapter with outliers in data described by a general linear model and the next with outliers in multivariate samples. Both of these contain the univariate sample as a particular case. In both chapters attention will be restricted to normal data, that is to data following a linear model with a normal error structure in this chapter and to samples from a multivariate normal distribution in the next.

It is only in comparatively recent years that attention has been paid to the detection of outliers in data arising from linear models. Early contributions are those of Daniel (1960) and Bross (1961) in the particular context of outliers in factorial experiments, and Srikantan (1961) in the context of outliers in regression models. More recent is the work of Cox and Snell (1968 and 1971), working from a general viewpoint, Andrews (1971), Elashoff (1972), Tietjen, Moore and Beckman (1973), Prescott (1975) and Lund (1975), who all consider problems of testing for outliers in simple linear regression models, and that of Stefansky (1971 and 1972), John and Prescott (1975) and Gentleman & Wilk (1975) who consider outliers in certain designed experiments which have the property that all residuals have a common variance. The latter consider various graphical procedures, based upon half normal plots, and extend the work of Daniel (1959)

from univariate samples to two-way tables. The results of this chapter apply to the general problem of detecting outliers in data following the linear model

$$Y = X\beta+\varepsilon, \tag{5.1}$$

where Y is an $n \times 1$ vector of observations, X is an $n \times m$ matrix of known constants (where $n > m+2$), β is an $m \times 1$ vector of unknown parameters and ε is an $n \times 1$ vector of errors jointly distributed normally with mean 0 and variance σ^2V, where V is a known $n \times m$ positive definite diagonal matrix, and σ is unknown scalar. It will be assumed that X is of full rank m. Without loss of any generality V may be taken to be the $n \times n$ identity matrix I_n. This follows since if P is the unique $n \times n$ diagonal matrix such that

$$P^2 = V$$

and if $Y^* = P^{-1}Y$, $X^* = P^{-1}X$, $\varepsilon^*=P^{-1}\varepsilon$ then the model (5.1) may be expressed as

$$Y^* = X^*\beta+\varepsilon^*$$

where $\varepsilon^* \sim N(0,\sigma^2I_n)$.

Particular attention is given to the special cases when the model (5.1) represents a polynomial regression of degrees one, two and three, (i.e. linear, quadratic and cubic regression), on a single independent variable measured at equally spaced points. The

empirical studies by Tietjen, Moore and Beckman (1973) suggest that the concentration on equally spaced values of the independent variable is, in the case of linear regression models at least, not unduly restrictive. These models fall outside the restricted class considered by Stefansky (1971, 1972) since the residuals do not, in general, have a common variance.

The detection of outliers in data described by such a linear model must involve consideration of the differing variances of the residuals, a point emphasised by Behnken and Draper (1972), since the 'most extreme' observation (as judged by an objective criterion) may not necessarily be that observation with the largest residual. For example in data having a linear regression upon a single independent variable the 'most extreme' observation is not necessarily that observation lying furthest from the 'fitted' regression line, since allowance must be made for the differences in variances of the residuals or deviations from the 'fitted' line. In particular a large deviation from the 'fitted' line at a value of the independent variable near its mean is of more note than an equally large deviation at one of the extremes of the range of the independent variable. This fact has been ignored by Andrews (1972) and in some standard texts on statistical methods, for example Snedecor and Cochran (1967), (sections 6.13 and 11.11). The definitions of 'most extreme' employed in the following sections is one based upon the likelihood of the sample. It is a direct extension of that used in the previous three chapters and, in the case of a single extreme, is equivalent to that implied by the use of the 'maximum normed residual' discussed in detail by Stefansky (1971, 1972) and, less explicitly, by the other authors

referred to above.

An important distinction between the problems of the detection of outliers in univariate samples and in data following more complex linear models is that in the more complex models it is no longer useful to consider separately 'one-sided' and 'two-sided' criteria. That is the observations tested as outlying will be those which are the most 'extreme', without regard to whether they are 'large' or 'small' (however 'large' and 'small' may be defined for data following a general linear model). Of course it is possible to define 'one-sided' criteria, for single outliers at least, for example by referring to the sign of the residual. In the context of simple linear regression this would entail testing only that observation which lay furthest above the 'fitted' regression line, (taking account of the differing variances of the deviations). Srikantan (1961) considers just such criteria. However 'one-sided' criteria are of limited practical application and furthermore do not in general have the desirable property that their definition can be extended to 'one-sidedness' for multiple outliers, since the removal of one suspect observation from the data may alter the sign of the residuals of other suspect observations after fitting the model to the reduced set of data. For these reasons only criteria which are equivalent to 'two-sided' criteria in the univariate case will be considered in this chapter.

Throughout the following sections it will be assumed that the vector β in the model (5.1) is unknown and to be estimated. If β were known the problem would reduce to the detection of outliers in univariate samples considered in the previous two chapters.

The following section establishes the notation and derives a likelihood-based criterion for testing outliers in linear models. Succeeding sections develop some general results concerning this criterion and some applications to particular models and practical examples.

5.1 Preliminaries

5.1.1 The Likelihood Ratio Test

Let $Y = (y_1,\ldots,y_n)'$ be an $n \times 1$ vector of observations of an n-dimensional random variable. Let X be a known $n \times m$ matrix with elements x_{ij}, with $n > m+2$, and assume X is of full rank m. Let β be an unknown $m \times 1$ vector and let I_n be the $n \times n$ identity matrix.

Denote by Y_i the $(n-1) \times 1$ vector obtained by omitting the i^{th} component of Y, (i.e. the vector $(y_1, y_2,\ldots,y_{i-1},y_{i+1},\ldots,y_n)'$). Similarly denote by X_i the $(n-1) \times m$ matrix obtained by omitting the i^{th} row of X.

Let H_o be the hypothesis that Y is an observation of a normal random variable with mean $X\beta$ and variance $\sigma^2 I_n$, where σ is an unknown scalar. Under H_o the log-likelihood of Y is

$$-\tfrac{1}{2}n\log(2\pi)-n\log\sigma-\tfrac{1}{2}(Y-X\beta)'(Y-X\beta)/\sigma^2.$$

This is maximised when

$$\beta = (X'X)^{-1}X'Y = \hat{\beta} \quad \text{(say)}$$

and $$\sigma^2 = (Y-X\hat{\beta})'(Y-X\hat{\beta})/n,$$

giving a maximised log-likelihood under H_o of

$$-\tfrac{1}{2}n\log(2\pi)-\tfrac{1}{2}n-\tfrac{1}{2}n\log\{(Y-X\hat{\beta})'(Y-X\hat{\beta})/n\}.$$

Let H_1 be the alternative hypothesis that one observation, y_n without loss of generality, arises from a normal distribution with mean μ and variance σ^2, and that the remaining (n-1) observations, represented by the vector Y_n, arise from a normal distribution with mean $X_n\beta$ and variance $\sigma^2 I_{n-1}$. Under H_1 the log-likelihood of Y is

$$-\tfrac{1}{2}n\log(2\pi)-n\log\sigma-\tfrac{1}{2}(Y_n-X_n\beta)'(Y_n-X_n\beta)/\sigma^2-\tfrac{1}{2}(y_n-\mu)^2/\sigma^2.$$

This is maximised when

$$\beta = (X_n'X_n)^{-1}X_n'Y_n = \hat{\beta}_n \quad \text{(say)},$$

$$\sigma^2 = (Y_n-X_n\hat{\beta}_n)'(Y_n-X_n\hat{\beta}_n)/n,$$

and $\mu = y_n$,

giving a maximised log-likelihood under H_1 of

$$-\tfrac{1}{2}n\log(2\pi)-\tfrac{1}{2}n-\tfrac{1}{2}n\log\{(Y_n-X_n\hat{\beta}_n)'(Y_n-X_n\hat{\beta}_n)/n\}.$$

The difference between the maximised log-likelihoods under the null and alternative hypotheses is thus

$$\tfrac{1}{2}n\log\{\{(Y-X\hat{\beta})'(Y-X\hat{\beta})\}/\{(Y_n-X_n\hat{\beta}_n)'(Y_n-X_n\hat{\beta}_n)\}\},$$

which may be written as

$$\tfrac{1}{2}n\log(R/R_n)$$

where $R = (Y-X\hat{\beta})'(Y-X\hat{\beta})$

and $R_i = (Y_i-X_i\hat{\beta}_i)'(Y_i-X_i\hat{\beta}_i)$ $\qquad (i=1,\ldots,n)$.

It follows that an appropriate criterion for testing an observation as outlying is any monotonic function of $\max_{i=1}^{n}\{R/R_i\}$, where the observation actually tested as outlying, is that for which R/R_i is maximum. That is the "extreme" observation is defined to be that observation whose omission from the sample produces the greatest decrease in the residual sum of squares after fitting the model (5.1). Throughout this chapter the 'extreme observation' will be that as here defined.

5.1.2 The Criterion $T_{(n)}$

Let $T_i = (n-m-1)(R-R_i)/R_i$ and suppose $T_{(1)} < T_{(2)} < \ldots < T_{(n)}$ are the ordered values of the T_i. Since $T_{(n)}$ is a monotonic function of $\max_{i=1}^{n}\{R/R_i\}$, $T_{(n)}$ is an appropriate likelihood-based statistic for testing the extreme observation as outlying.

For arbitrary i, T_i follows an F-distribution with 1 and (n-m-1) degrees of freedom. This may be seen by considering the model $Y = X + Z_i\gamma+\varepsilon$, where Z_i is the vector with i^{th} component equal to 1 and zero's elsewhere, and γ is an unknown scalar. R_i will be equal to the residual sum of squares, with n-m-1 degrees of freedom, after fitting this model and R will be the residual sum of squares, with n-m degrees of freedom, after fitting the model under the constraint $\gamma=0$.

Throughout this chapter the statistic $T_{(n)}$ will be used to test the extreme observation as outlying. It is important to note that this statistic is essentially equivalent to that used by

Srikantan (1961), and termed the 'maximum studentized residual' by Ferguson (1961). Here 'studentized residual' is taken to mean the residual divided by an estimate of its standard deviation. In the case of the residuals having a common variance it is also equivalent to the 'maximum normed residual' of Stefansky (1971 and 1972). To see this define the $1 \times m$ vector x_i' to be the i^{th} row of X, i.e.

$$x_i' = (x_{i1}, x_{i2}, \ldots, x_{im}),$$

and let the vector of residuals be the $n \times 1$ vector

$$e = (e_1, e_2, \ldots, e_n)' = Y - X\hat{\beta},$$

$$\text{then } e_i = y_i - x_i'\hat{\beta}.$$

The statistic used by Srikantan for testing the extreme observation as outlying is

$$t_{(n)} = \max_{i=1}^{n}\{t_i\},$$

$$\text{where } t_i = e_i^2/\{R(1-x_i'(X'X)^{-1}x_i)\}. \tag{5.2}$$

$$\text{Now } R = Y'Y - Y'X(X'X)^{-1}X'Y$$

$$\text{and } R_i = Y_i'Y_i - Y_i'X_i(X_i'X_i)^{-1}X_i'Y_i.$$

$$\text{Since } Y_i'Y_i = Y'Y - y_i^2, \tag{5.3a}$$

$$X_i'X_i = Y'X - x_i x_i', \tag{5.3b}$$

and $$X_i'Y_i = X'Y - y_i x_i, \tag{5.3c}$$

R_i may be written as

$$R_i = Y'Y - y_i^2 - (Y'X - y_i x_i')X'X)^{-1}\{I_m - x_i x_i'(X'X)^{-1}\}^{-1}(X'Y - y_i x_i).$$

Expanding $\{I_m - x_i x_i'(X'X)^{-1}\}^{-1}$ as

$$I_m + x_i x_i'(X'X)^{-1} + x_i x_i'(X'X)^{-1} x_i x_i'(X'X)^{-1} + \ldots$$

and collecting terms gives

$$R_i = Y'Y - Y'X(X'X)^{-1}X'Y - y_i^2 -$$

$$\{y_i^2 x_i'(X'X)^{-1}x_i - 2y_i x_i'(X'X)^{-1}X'Y + (x_i'(X'X)^{-1}X'Y)^2\}W$$

where $W = 1 + x_i'(X'X)^{-1}x_i + (x_i'(X'X)^{-1}x_i)^2 + \ldots$

$$= \{1 - x_i'(X'X)^{-1}x_i\}^{-1},$$

whence $$R - R_i = e_i^2/\{1 - x_i'(X'X)^{-1}x_i\}. \tag{5.4}$$

Hence $$T_i = (n-m-1)t_i/(1-t_i), \tag{5.5}$$

and in particular, the statistics $T_{(n)}$ and $t_{(n)}$ are related by the equation

$$T_{(n)} = (n-m-1)t_{(n)}/(1-t_{(n)}). \qquad (5.5a)$$

It follows that the statistics $T_{(n)}$ and $t_{(n)}$ are essentially equivalent, as asserted. A particular consequence of the equivalence is that the 'extreme observation' may be defined alternatively to be that observation with the largest studentized residual, in the sense defined above.

It is readily seen that the null distribution of $T_{(n)}$, $F_n(.|X)$ say, depends upon the particular form of the design matrix X. It is not possible to derive this distribution by a recursive procedure in a manner analogous to that described in earlier chapters. Such a procedure would relate $F_n(.|X)$ to $F_{n-1}(.|X_n)$, the distribution of the likelihood-based criterion based upon the reduced design matrix X_n. The reason for this is that $T_{(n)}$ is essentially a 'two-sided' criterion; an examination of the particular case $X = (1,1,\ldots,1)'$ (i.e. the univariate sample case considered in section 3.6) makes this clear. It is possible to derive recursively the distribution of 'one-sided' criteria, such as $\max_{i=1}^{n} e_i/\{1-x_i'(X'X)^{-1}x_i\}^{\frac{1}{2}}\}$, at least for some forms of the design matrix X, but the matter is not pursued here.

It is possible to derive the upper tail of the distribution of $T_{(n)}$ using upper bounds on the magnitude of $T_{(n-1)}$, the second largest of the T_i. These bounds are established in the following section and their application to the distribution of $T_{(n)}$ is considered in the section 5.3.

5.2 Upper Bounds for $T_{(n-1)}$

The following notation will be needed: let X_{ij} be the $(n-2) \times m$ matrix obtained by omitting both the i^{th} and j^{th} rows of X, (assuming $i \neq j$ throughout), let Y_{ij} be the $(n-2) \times 1$ vector obtained by omitting both the i^{th} and j^{th} components of Y, and let R_{ij} be the residual sum of squares after fitting the model of the form (5.1) to the observations Y_{ij}, that is

$$R_{ij} = Y_{ij}'Y_{ij} - Y_{ij}'X_{ij}(X_{ij}'X_{ij})^{-1}X_{ij}'Y_{ij}.$$

For brevity write $Z = (X'X)^{-1}$, $Z_i = (X_i'X_i)^{-1}$ and $Z_{ij} = (X_{ij}'X_{ij})^{-1}$.

Applying identity (5.4) to the reduced sample Y_j gives

$$R_j - R_{ij} = (y_i - x_i'\hat{\beta}_j)^2/\{1 - x_i'Z_j x_i\}. \tag{5.6}$$

Now $$x_i'\hat{\beta}_j = x_i'Z_jX_j'Y_j$$

$$= x_i'Z\{I_m - x_jx_j'Z\}^{-1}(X'Y - y_jx_j)$$

(on application of identities 5.3).

Expanding $\{I_m - x_jx_j'Z\}^{-1}$ and rearranging terms yields

$$x_i'\hat{\beta}_j = x_i'ZX'Y + \{y_j - x_j'ZX'Y\}\{\sum_{r=0}^{\infty}(x_j'Zx_j)^r\}x_i'Zx_j,$$

giving

$$y_i - x_i'\hat{\beta}_j = e_i + e_jx_i'Zx_j(1 - x_j'Zx_j)^{-1}. \tag{5.7}$$

Substitution of (5.7) in (5.6) and the addition of identity (5.4) (with i replaced by j) gives, after some algebra,

$$R-R_{ij} = e_i^2/\{1-x_i'Z_jx_i\}+e_j^2/\{1-x_j'Z_ix_j\}+2e_ix_i'Z_{ij}x_j. \tag{5.8}$$

Thus, upon substitution of t_i defined in (5.2),

$$(R-R_{ij})/R = A_{ij}(t_i+t_j)+2B_{ij}(t_it_j)^{\frac{1}{2}} \tag{5.9}$$

(writing $A_{ij} = \{1-x_i'Zx_i\}/\{1-x_i'Z_jx_i\}$

$= \{1-x_j'Zx_j\}/\{1-x_j'Z_ix_j\}$ (after some algebra),

and $$B_{ij} = x_j'Z_{ij}x_j\{1-x_i'Zx_i\}^{\frac{1}{2}}\{1-x_j'Zx_j\}^{\frac{1}{2}}). \tag{5.10}$$

Now necessarily $R_{ij} \geq 0$, since R_{ij} is a sum of squares. It follows that the point $(\sqrt{t_i},\sqrt{t_j})$ must lie within the ellipse defined by

$$A_{ij}(x^2+y^2)+2B_{ij}xy = 1 \tag{5.11}$$

Thus if

$$t_i > \max\{\tfrac{1}{2}(A_{ij}+B_{ij})^{-1},\tfrac{1}{2}(A_{ij}-B_{ij})^{-1}\}$$

then $t_i > t_j$.

Let $$k_{ij} = \max\{\tfrac{1}{2}(A_{ij}+B_{ij})^{-1},\tfrac{1}{2}(A_{ij}-B_{ij})^{-1}\}. \tag{5.12}$$

It may be noted that $k_{ij} > 0$. This follows since $R > R_{ij}$, implying that the right hand side of equation (5.9) is positive definite, so that $A_{ij} > B_{ij}$. Clearly $k_{ij} < 1$.

$$\text{Define } k_i = \max_{\substack{j=1 \\ j \neq i}}^{n} \{k_{ij}\}, \tag{5.13}$$

then

$$t_i > k_i \text{ implies that } t_i > t_j \text{ for all } j \neq i. \tag{5.14}$$

Now define $K_i = (n-m-1)k_i/(1-k_i)$.

Then, since $0 < k_i < 1$, (5.14) gives

$$T_i > K_i \text{ implies that } T_i > T_j \text{ for all } j \neq i. \tag{5.15}$$

If $K = \max_{i=1}^{n} K_i$ then (5.15) gives the rather weaker result that

$$T_i > K \text{ implies } T_i > T_j \text{ for all } j \neq i, \tag{5.16}$$

which is equivalent to that obtained by Srikantan (1961), by a different argument. If the design matrix X is such that the residuals have common variance then it will follow that all the k_i, defined in (5.13), will be equal, so that statements (5.15) and (5.16) will be identical and will reduce to the result of Stefansky (1971 and 1972).

It may be noted that (5.16) is equivalent to the inequality

$$T_{(n-1)} < K, \tag{5.17}$$

that is K is an upper bound for $T_{(n-1)}$. Further (5.15) may be stated as

'if there exists an i, $1 \leq i \leq n$, such that $T_i > K_i$
then $T_{(n-1)} < K_i$'.

(5.18)

5.3 The Upper Percentage Points of $T_{(n)}$

In section 5.1 attention was drawn to the important fact that the null distribution of $T_{(n)}$ depends upon the particular form of the design matrix X. In this section it is shewn that it is possible to obtain upper bounds for the upper percentage points of $T_{(n)}$ irrespective of the particular form of X. These are tabulated in Table 5.1 for various significance levels and sample sizes. It is shewn further that for the smaller sample sizes, depending upon the form of the design matrix X, these upper bounds coincide with the actual percentage points of $T_{(n)}$. The special cases when the model (5.1) represents a polynomial regression of degrees one, two and three, with a constant term, are considered in section 5.4

Consider the identity

$$P\,[T_{(n)} > u] = \sum_{j=1}^{n} P\,[T_j > u,\ T_j = T_{(n)}]\ . \qquad (5.19)$$

Since $P\,[T_i > u,\ T_i = T_{(n)}] \leq P\,[T_i > u\,]$

it follows that

$$P\,[T_{(n)} > u] \leq n\,P\,[T_j > u]\ . \qquad (5.20)$$

Now for arbitrary j, T_j follows an F-distribution with (1,n-m-1) degrees of freedom (see section 5.1.2). Hence the upper 100α percentage point of $T_{(n)}$ is bounded above by the upper $100\alpha/n$ percentage point, $T_n(\alpha;m)$ say, of the F-distribution with (1,n-m-1) degrees of freedom.

Table 5.1 gives the values of the upper bounds $T_n(\alpha;m)$ of the percentage points of $T_{(n)}$ for each value of m, m=2(1)6, n=(m+3)(1)30 and α=0·1, 0·05, 0·025, 0·01, 0·001. (Values for m=1, which includes the univariate sample case, may readily be derived from the tables of Quesenberry and David (1961) or Srikantan (1961)). Recently Lund (1975) has calculated upper bounds for the equivalent statistic $t_{(n)}$ for m=1(1)6,8,10,15,25, n=(m+4)(1)20(5)50(10)100 and α=0·1, 0·5, 0·01.

The values in Table 5.1 were obtained by solving the equation

$$\int_0^{T_n(\alpha:m)} dF_{1,n-m-1} = 1-\alpha/n$$

(where here $F_{1,n-m-1}$ is the distribution function of the F-distribution of (1,n-m-1) degrees of freedom), using a modified Newton-Raphson iterative procedure. The values are expected to be correct to within one unit in the last figure.

Let $\{\tau(1),\tau(2),\ldots,\tau(n)\}$ be a permutation of the integers $\{1,2,\ldots,n\}$ such that

$$K_{\tau(1)} \leq K_{\tau(2)} \leq \cdots \leq K_{\tau(n)}.$$

Let u > 0 and suppose that r is the unique positive integer such that

$$K_{\tau(r)} < u \leq K_{\tau(r+1)},$$

(i.e. suppose u lies between the r^{th} and $(r+1)^{th}$ largest of the K_i). If $u < K_{(1)}$ define r=0, if $u > K_{(n)} \equiv K$ define r=n.

Table 5.1

Upper Bounds for the Upper percentage points of $T_{(n)}$

a) m=2

n	10%	5%	2·5%	1%	0·1%
5	48·5051	98·5025	198·5013	498·5005	4998·5000
6	23·5871	38·8321	63·0396	118·1305	557·0338
7	17·2043	25·6796	37·6755	61·4872	201·6098
8	14·5161	20·4821	28·3654	42·7905	113·9082
9	13·0968	17·8162	23·7742	34·1039	79·4624
10	12·2464	16·2356	21·1107	29·2452	62·1667
11	11·6949	15·2090	19·4005	26·1970	52·0883
12	11·3179	14·4999	18·2245	24·1312	45·6076
13	11·0505	13·9881	17·3750	22·6520	41·1408
14	10·8561	13·6063	16·7384	21·5484	37·9007
15	10·7122	13·3144	16·2477	20·6986	35·4567
16	10·6046	13·0869	15·8610	20·0278	33·5558
17	10·5236	12·9070	15·5507	19·4873	32·0403
18	10·4628	12·7630	15·2982	19·0447	30·8075
19	10·4175	12·6469	15·0901	18·6771	29·7875
20	10·3842	12·5525	14·9170	18·3683	28·9315
21	10·3604	12·4755	14·7718	18·1063	28·2044
22	10·3443	12·4126	14·6492	17·8821	27·5802
23	10·3343	12·3611	14·5451	17·6888	27·0396
24	10·3293	12·3191	14·4563	17·5210	26·5674
25	10·3284	12·2850	14·3893	17·3748	26·1522
26	10·3309	12·2574	14·3150	17·2465	25·7848
27	10·3362	12·2355	14·2590	17·1336	25·4577
28	10·3438	12·2182	14·2107	17·0339	25·1652
29	10·3534	12·2050	14·1692	16·9455	24·9023
30	10·3646	12·1953	14·1335	16·8669	24·6652

Table 5.1(cont.)

b) m=3

n	10%	5%	2·5%	1%	0·1%
6	58·5042	118·5021	598·5010	598·5004	5998·5000
7	26·3955	43·2922	70·1207	131·1750	617·5830
8	18·6163	27·6787	40·5041	65·9610	215·7595
9	15·4161	21·6718	29·9368	45·0589	119·6090
10	13·7450	18·6350	24·8073	35·5075	82·4890
11	12·7484	16·8495	21·8605	30·2208	64·0529
12	12·1026	15·6956	19·9806	26·9277	53·3902
13	11·6601	14·9008	18·6933	24·7073	46·5708
14	11·3449	14·3277	17·7662	23·1232	41·8898
15	11·1140	13·9002	17·0728	21·9446	38·5050
16	10·9414	13·5730	16·5390	21·0390	35·9583
17	10·8107	13·3175	16·1185	20·3252	33·9815
18	10·7108	13·1148	15·7811	19·7509	32·4082
19	10·6342	12·9521	15·5064	19·2808	31·1301
20	10·5755	12·8201	15·2798	18·8907	30·0738
21	10·5307	12·7123	15·0912	18·5631	29·1883
22	10·4969	12·6238	14·9327	18·2852	28·4368
23	10·4721	12·5509	14·7986	18·9473	27·7921
24	10·4544	12·4906	14·6845	17·8422	27·2340
25	10·4425	12·4408	14·5869	17·6642	26·7469
26	10·4355	12·3998	14·5031	17·5089	26·3187
27	10·4325	12·3660	14·4309	17·3726	25·9399
28	10·4328	12·3385	14·3687	17·2526	25·6029
29	10·4359	12·3162	14·3148	17·1464	25·3015
30	10·4414	12·2985	14·2682	17·0522	25·0308

Table 5.1(cont.)

c) m=4

n	10%	5%	2·5%	1%	0·1%
7	68·5036	138·5018	278·5009	698·5004	6998·5000
8	29·0731	47·5442	76·8714	143·6106	675·3054
9	19·9428	29·5565	43·1610	70·1629	229·0493
10	16·2582	22·7848	31·4067	47·1808	124·9412
11	14·3516	19·4010	25·7738	36·8204	85·3200
12	13·2191	17·4249	22·5632	31·1351	65·8204
13	12·4858	16·1529	20·5255	27·6142	54·6131
14	11·9826	15·2785	19·1350	25·2498	47·4780
15	11·6230	14·6485	18·1358	23·5682	42·5971
16	11·3583	14·1786	17·3896	22·3198	39·0772
17	11·1591	13·8185	16·8155	21·3622	36·4344
18	11·0070	13·5370	16·3635	20·6082	34·3866
19	10·8895	13·3131	16·0008	20·0022	32·7590
20	10·7980	13·1328	15·7053	19·5065	31·4383
21	10·7267	12·9861	15·4615	19·0952	30·3480
22	10·6712	12·8657	15·2582	18·7499	29·4346
23	10·6281	12·7663	15·0872	18·4570	28·6600
24	10·5950	12·6839	14·9423	18·2063	27·9959
25	10·5701	12·6153	14·8188	17·9901	27·4213
26	10·5518	12·5582	14·7129	17·8023	26·9200
27	10·5390	12·5106	14·6217	17·6384	26·4795
28	10·5308	12·4711	14·5430	17·4946	26·0899
29	10·5264	12·4383	14·4748	17·3677	25·7434
30	10·5252	12·4112	14·4157	17·2555	25·4336

Table 5.1(cont.)

d) m=5

n	10%	5%	2·5%	1%	0·1%
8	78·5031	158·5016	318·5008	798·5003	7998·5000
9	31·6414	51·6223	83·3458	155·5372	730·6652
10	21·1977	31·3328	45·6740	74·1373	241·6191
11	17·0514	23·8330	32·7909	49·1789	129·9623
12	14·9227	20·1221	26·6834	38·0561	87·9843
13	13·6627	17·9672	23·2254	31·9965	67·4856
14	12·8476	16·5846	21·0400	28·2621	55·7674
15	12·2878	15·6358	19·5529	25·7632	48·3362
16	11·8868	14·9528	18·4863	23·9901	43·2677
17	11·5905	14·4431	17·6906	22·6764	39·6208
18	11·3665	14·0524	17·0789	21·6699	36·8877
19	11·1943	13·7464	16·5973	20·8783	34·7730
20	11·0603	13·5026	16·2109	20·2424	33·0943
21	10·9550	13·3059	15·8959	19·7226	31·7334
22	10·8719	13·1452	15·6358	19·2914	30·6109
23	10·8062	13·0130	15·4187	18·9294	29·6713
24	10·7543	12·9034	15·2360	18·6223	28·8748
25	10·7135	12·8121	15·0808	18·3595	28·1923
26	10·6817	12·7357	14·9484	18·1327	27·6020
27	10·6573	12·6717	14·8346	17·9357	27·0872
28	10·6391	12·6179	14·7365	17·7637	26·6349
29	10·6259	12·5728	14·6515	17·6126	26·2351
30	10·6171	12·5349	14·5777	17·4793	25·8795

Table 5.1(cont.)

e) m=6

n	10%	5%	2·5%	1%	0·1%
9	88·5028	178·5014	358·5007	898·5003	8998·5000
10	34·1162	55·5520	89·5843	167·0292	784·0073
11	22·3915	33·0223	48·0643	77·9175	253·5746
12	17·8026	24·8256	34·1016	51·0708	134·7163
13	15·4630	20·8043	27·5439	39·3250	90·5045
14	14·0827	18·4805	23·8521	32·8118	69·0617
15	13·1907	16·9939	21·5177	28·8763	56·8615
16	12·5777	15·9752	19·9498	26·2505	49·1510
17	12·1378	15·2423	18·8197	24·3915	43·9056
18	11·8119	14·6954	17·9775	23·0162	40·1389
19	11·5646	14·2758	17·3304	21·9637	37·3206
20	11·3735	13·9468	16·8210	21·1367	35·1426
21	11·2240	13·6843	16·4121	20·4726	33·4155
22	11·1057	13·4720	16·0788	19·9299	32·0167
23	11·0115	13·2983	15·8033	19·4800	30·8636
24	10·9362	13·1549	15·5733	19·1022	29·8990
25	10·8760	13·0356	15·3793	18·7817	29·0817
26	10·8279	12·9358	15·2145	18·5073	28·3817
27	10·7896	12·8520	15·0736	18·2705	27·7765
28	10·7594	12·7814	14·9523	18·0648	27·2488
29	10·7360	12·7217	14·8476	17·8850	26·7854
30	10·7181	12·6713	14·7567	17·7270	26·3758

Then (5.15) gives

$$P\left[T_{\tau(i)} > u,\ T_{\tau(i)} = T_{(n)}\right] = P\left[T_{\tau(i)} > u\right] \quad \text{for each } i,\ 1 \leqslant i \leqslant r. \tag{5.21}$$

Now (5.19) may be written as

$$P\left[T_{(n)} > u\right] = \sum_{j=1}^{n} P\left[T_{(j)} > u,\ T_{(j)} = T_{(n)}\right].$$

$$\text{Thus } P[T_{(n)} > u] = rP[F_{1,n-m-1} > u] + \sum_{j=r+1}^{n} P[T_{\tau(j)} > u,\ T_{\tau(j)} = T_{(n)}] \tag{5.22}$$

(where $F_{1,n-m-1}$ denotes a variate following an F-distribution with (1,n-m-1) degrees of freedom).

In particular if r=n, i.e. if u > K, then

$$P\left[T_{(n)} > u\right] = P\left[F_{1,n-m-1} > u\right]. \tag{5.23}$$

Thus if the upper $100\alpha/n$ percentage point of $F_{1,n-m-1}$ exceeds K then it coincides with the upper 100α percentage point of $T_{(n)}$. Hence

$$P\left[T_{(n)} > T_n(\alpha;m)\right] \leqslant \alpha,$$

and equality is achieved if $T_n(\alpha;m) \geqslant K$. (5.24)

It may be noted that the calculation of the upper bounds on $T_{(n-1)}$, the K_i and K of section 5.2, for any particular design matrix

X may be performed with only a small amount of computation additional to that needed in any standard analysis of the data. In particular the matrix $(X'X)^{-1}$ will be needed not only for evaluating the K_i but also in any later analysis. Values of the K_i for some standard polynomial regression models are given in the next section.

5.4 Polynomial Regression

Table 5.2 gives values of K_i, the upper bounds on $T_{(n-1)}$ defined in section 5.2, for the particular cases when the model 5.1 represents a polynomial regression of degrees one, two and three (Tables 5.2 (a), (b) and (c) respectively) on equally spaced values of the independent variable, for all sample sizes up to 30. Since values of the K_i are clearly invariant under transformations of location and scale of the independent variable, in these polynomial cases the design matrix X may be taken, without loss of generality, to be the matrix whose $(i,j)^{th}$ element is i^{j-1}, $1 \leq i \leq n$, $1 \leq j \leq m$.

Considering the symmetry of such models it is clear that for each i

$$K_i = K_{n-i+1}.$$

For compactness therefore values of K_i are tabulated only for $1 \leq i \leq \frac{1}{2}n$ if n is even, and $1 \leq i \leq \frac{1}{2}(n+1)$ if n is odd.

In these particular models it is seen that the largest K_i always occurs when i=1 (or i=n), that is the first column of each table gives the values of K, defined in section 5.2, for various sample sizes. A comparison with Table 5.1 shews, upon application of the result (5.24), that in the case of linear regression on equally spaced values of the independent variable the upper bounds $T_n(\alpha;m)$ given in Table 5.1(a) for the upper 100α percentage points of $T_{(n)}$ coincide with the actual upper 100α percentage points of $T_{(n)}$ for sample sizes n up to and including n=7,9,11,14,21

Table 5.2

Values of K_i for Polynomial Regressions

a) Linear Regression

Sample Size	Values of Independent Variable, i						
n	1	2	3	4	5	6	7
4	14·73	14·73	-	-	-	-	-
5	14·39	14·39	14·39	-	-	-	-
6	14·52	14·52	6·70	-	-	-	-
7	14·93	14·93	8·84	6·14	-	-	-
8	15·50	15·50	10·65	8·06	-	-	-
9	16·17	16·17	12·23	9·82	8·11	-	-
10	16·92	16·92	13·64	11·43	9·79	-	-
11	17·72	17·72	14·94	12·93	11·37	10·09	-
12	18·56	18·56	16·17	14·34	12·87	11·64	-
13	19·42	19·42	17·33	15·67	14·29	13·11	12·08
14	20·31	20·31	18·46	16·94	15·65	14·53	13·53
15	21·21	21·21	19·56	18·17	16·96	15·90	14·94
16	22·13	22·13	20·64	19·36	18·23	17·22	16·30
17	23·05	23·05	21·70	20·52	19·46	18·50	17·63
18	23·99	23·99	22·75	21·65	20·66	19·75	18·92
19	24·93	24·93	23·80	22·77	21·84	20·98	20·18
20	25·88	25·88	24·83	23·87	22·99	22·17	21·41
21	26·84	26·84	25·86	24·96	24·13	23·35	26·62
22	27·79	27·79	26·88	26·04	25·25	24·51	23·81
23	28·76	28·76	27·90	27·11	26·36	25·65	24·98
24	29·72	29·72	28·92	28·17	27·46	26·78	26·15
25	30·69	30·69	29·93	29·22	28·54	27·90	27·28
26	31·66	31·66	30·95	30·27	29·62	29·01	28·41
27	32·64	32·64	31·96	31·31	30·69	30·10	29·53
28	33·61	33·61	32·97	32·35	31·76	31·19	30·64
29	34·59	34·59	33·98	33·39	32·82	32·27	31·75
30	35·57	35·57	34·98	34·42	33·87	33·35	32·84

Table 5.2(a) (cont.)

Sample Size	Values of Independent Variable, i							
n	8	9	10	11	12	13	14	15
15	14·07	-	-	-	-	-	-	-
16	15·46	-	-	-	-	-	-	-
17	16·82	16·06	-	-	-	-	-	-
18	18·14	17·40	-	-	-	-	-	-
19	19·43	18·72	18·06	-	-	-	-	-
20	20·69	20·01	19·36	-	-	-	-	-
21	21·93	21·27	20·65	20·05	-	-	-	-
22	23·15	22·52	21·91	21·33	-	-	-	-
23	24·35	23·74	23·15	25·59	22·04	-	-	-
24	25·53	24·94	24·37	23·83	23·30	-	-	-
25	26·69	26·13	25·58	25·05	24·54	24·04	-	-
26	27·85	27·30	26·77	26·26	25·76	25·28	-	-
27	28·99	28·46	27·95	27·45	26·97	26·50	26·04	-
28	30·12	29·61	29·11	28·63	28·16	27·70	27·26	-
29	31·24	30·74	30·26	29·80	29·34	28·90	28·46	28·03
30	32·35	31·87	31·41	30·95	30·51	30·08	29·65	29·24

Table 5.2 (cont.)

(b) Quadratic Regression

Sample Size	Values of Independent Variable, i						
n	1	2	3	4	5	6	7
5	48·27	48·27	4·04	-	-	-	-
6	44·41	44·41	6·80	-	-	-	-
7	41·78	41·78	7·24	7·24	-	-	-
8	40·00	40·00	7·63	8·74	-	-	-
9	38·81	38·81	9·58	9·47	9·47	-	-
10	38·05	38·05	12·23	10·07	10·70	-	-
11	37·62	37·62	14·75	10·73	11·54	11·54	-
12	37·43	37·43	17·10	11·50	12·27	12·67	-
13	37·44	37·44	19·28	13·15	12·99	13·57	13·57
14	37·59	37·59	21·31	15·12	13·76	14·37	14·65
15	37·86	37·86	23·19	17·02	14·59	15·15	15·58
16	38·23	38·23	24·94	18·87	15·48	15·95	16·43
17	38·68	38·68	26·59	20·65	17·01	16·78	17·25
18	39·20	39·20	28·15	22·36	18·70	17·64	18·08
19	39·77	39·77	29·63	24·01	20·36	18·55	18·92
20	40·39	40·39	31·04	25·61	21·98	19·50	19·78
21	41·05	41·05	32·40	27·16	23·57	20·94	20·67
22	41·74	41·74	33·71	28·66	25·12	22·48	21·59
23	42·46	42·46	34·98	30·12	26·64	24·01	22·53
24	43·20	43·20	36·22	31·53	28·12	25·50	23·54
25	43·97	43·97	37·42	32·91	29·57	26·98	24·90
26	44·76	44·76	38·61	34·26	30·99	28·42	26·35
27	45·56	45·56	39·77	35·59	32·39	29·85	27·78
28	46·38	46·38	40·91	36·88	33·76	31·26	29·19
29	47·21	47·21	42·03	38·15	35·11	32·64	30·59
30	48·06	48·06	43·15	39·40	36·43	34·00	31·97

Table 5.2(b) (cont.)

Sample Size	Values of Independent Variable, i							
n	8	9	10	11	12	13	14	15
15	15·58	-	-	-	-	-	-	-
16	16·63	-	-	-	-	-	-	-
17	17·58	17·58	-	-	-	-	-	-
18	18·46	18·62	-	-	-	-	-	-
19	19·32	19·58	19·58	-	-	-	-	-
20	20·17	20·49	20·60	-	-	-	-	-
21	21·03	21·37	21·57	-	-	-	-	-
22	21·91	22·24	22·50	22·60	-	-	-	-
23	22·80	23·13	23·40	23·57	23·57	-	-	-
24	23·70	24·02	24·29	24·57	24·59	-	-	-
25	24·62	24·91	25·19	25·43	25·57	25·57	-	-
26	25·57	25·81	26·10	26·33	26·52	26·58	-	-
27	26·53	26·72	27·00	27·25	27·45	27·56	27·56	-
28	27·56	27·65	27·90	28·16	28·36	28·52	28·58	-
29	28·87	28·59	28·07	29·07	29·29	29·46	29·56	29·56
30	30·25	29·53	29·74	29·98	30·21	30·39	30·52	30·57

Table 5.2 (cont.)

(c) Cubic Regression

Sample Size	Values of Independent Variable, i						
n	1	2	3	4	5	6	7
6	115·65	115·79	13·95	-	-	-	-
7	106·38	106·36	11·14	5·59	-	-	-
8	98·99	98·99	10·19	8·84	-	-	-
9	93·16	93·16	11·23	11·23	7·89	-	-
10	88·49	88·49	12·58	12·58	10·05	-	-
11	84·72	84·72	13·37	13·37	12·10	10·05	-
12	81·67	81·67	13·99	13·99	13·79	11·78	-
13	79·19	79·19	14·60	15·10	15·10	13·54	12·15
14	77·19	77·19	16·17	16·14	16·14	15·17	13·65
15	75·57	75·57	18·92	17·03	17·03	16·61	15·22
16	74·27	74·27	21·63	17·83	17·86	17·86	16·75
17	73·23	73·23	24·26	18·62	18·96	18·96	18·19
18	72·43	72·43	26·80	19·42	19·44	19·44	19·51
19	71·82	71·82	29·24	20·25	20·86	20·86	20·71
20	71·37	71·37	31·58	21·11	21·74	21·82	21·82
21	71·06	71·06	33·82	22·73	22·60	22·85	22·85
22	70·88	70·88	35·96	24·73	23·46	23·83	23·83
23	70·81	70·81	38·02	26·71	24·33	24·76	24·76
24	70·83	70·83	39·99	28·65	25·21	25·67	25·77
25	70·93	70·93	41·87	30·56	26·12	26·57	26·78
26	71·11	71·11	43·69	32·43	27·04	27·47	27·75
27	71·35	71·35	45·43	34·27	28·13	28·37	28·70
28	71·65	71·65	47·11	36·06	29·64	29·27	29·63
29	72·00	72·00	48·74	37·82	31·35	30·19	30·55
30	72·40	72·40	50·31	39·55	33·05	31·12	31·47

Table 5.2(c) (cont.)

Sample Size	Values of Independent Variable, i							
n	8	9	10	11	12	13	14	15
15	14·21	-	-	-	-	-	-	-
16	15·58	-	-	-	-	-	-	-
17	17·02	16·26	-	-	-	-	-	-
18	18·46	17·54	-	-	-	-	-	-
19	19·86	18·89	18·30	-	-	-	-	-
20	21·19	20·26	19·52	-	-	-	-	-
21	22·44	21·62	20·80	20·32	-	-	-	-
22	23·62	22·93	22·11	21·50	-	-	-	-
23	24·72	24·19	23·43	22·74	22·34	-	-	-
24	25·77	25·40	24·72	24·02	23·49	-	-	-
25	26·78	26·55	25·97	25·31	24·69	24·36	-	-
26	27·75	27·65	27·19	26·58	25·95	25·48	-	-
27	28·71	28·71	28·36	27·21	27·21	26·65	26 37	-
28	29·73	29·73	29·49	29·00	28·46	27·89	27·48	-
29	30·72	30·72	30·59	30·18	29·68	29·12	28·63	28·39
30	31·69	31·69	31·65	31·33	30·87	30·35	29·84	29·48

for α=0·1,0·05,0·025,0·01,0·001 respectively. This agrees with an equivalent result for α=0·05 and α=0·01 given by Srikantan (1961). In the case of quadratic regression on equally spaced values of the independent variable the upper bounds given in Table 5.1(b) coincide with the actual 100α percentage points for sample sizes n up to and including n=6,6,8,9,15 for α=0·1,0·05,0·025,0·01,0·001 respectively. In the case of cubic regression the upper bounds given in Table 5.1(c) coincide with the actual values for sample sizes n up to and including n=7,7,8,11 for α=0·05,0·025,0·01,0·001 respectively, (note that for α=0·1 there is no value of n for which the upper bound coincides with the actual value). It is clear that in the case of polynomial regression of order higher than three the upper bounds $T_n(\alpha;m)$ will coincide with the actual upper 100α percentage points of $T_{(n)}$ only for very small values of α and even then only for small sample sizes.

An examination of the values of K_i second largest in magnitude, which in practice means, for the particular models under discussion, the fifth K_i in ordered sequence, since for these models K_1 and K_2 agree to several significant figures, so $K_2=K_{n-1} \simeq K_1=K_n=K$), and application of result (5.22) indicates that although equality of the upper bounds $T_n(\alpha;m)$ with the actual percentage points of $T_{(n)}$ is not achieved, these bounds will nevertheless be close to the actual values for sample sizes rather larger than those quoted above as ensuring equality of bound and actual percentage point. In the case of linear regression this will be so for samples of sizes one or two greater than those quoted above, and for cubic regression it will be so for samples of sizes seven or eight greater than those quoted above.

Although the results given above apply specifically to polynomial regressions on equally spaced values of the independent variable they have wider applicability. Tietjen, Moore and Beckman (1973), in a small scale Monte Carlo experiment, studied the effect of various different spacings of the independent variable, in the case of simple linear regression, upon the distribution of a statistic functionally related to $T_{(n)}$. They generated two samples of 5,000 values of their test criterion for each of four different configurations of values of the independent variable and for each of four sample sizes. The four configurations of the independent variable they considered were firstly equally spaced values, secondly $[\frac{1}{2}n]$ values of 1 and the others at 0, thirdly all but two values at 0 and the other two at 1 and fourthly all but two values at $\frac{1}{2}$ and the other two at 0 and 1. The four sample sizes were 5, 10, 16 and 20. They found that the empirical five per cent point of the distribution of their statistic was little affected by the configuration of the values of the independent variable. Of course this is not surprising; the result (5.24) shews that for small sample sizes the upper tail of the distribution of $T_{(n)}$ is independent of the configuration of the values of the independent variable, and result (5.22) indicates that for rather larger sample sizes the distribution of $T_{(n)}$ is well approximated by a scaled F-distribution with (1,n-3) degrees of freedom, again independently of the relative values of the independent variable. In the cases of polynomial regression of higher orders it is also true that for small samples the extreme tail of the distribution of $T_{(n)}$ is independent (or at least "approximately independent", in the sense suggested above) of the relative spacings of the independent

variable.

Tietjen, Moore and Beckman (1973) extend the empirical study of the distribution of a test criterion for a single outlier in linear regression, referred to above, to sample sizes 4(1)12(2)20,24,30. They again obtain two samples of 5,000 values of their test criterion for each of four distinct arrangements of the independent variable. They derive empirical 100α percentage points (for α=0·1,0·05,0·01) of the distributions of their criterion from each of the eight samples and then tabulate the 'averaged' values of these empirical percentiles, for each α, despite the fact that the eight empirical percentiles are derived from the four different distributions of their criterion corresponding to the four distinct arrangements of the independent variable. They also tabulate empirical 'percentiles' of their criterion for sample sizes 36,48,60 and 100 based on a total of only 10,000 observations for each sample size, but do not state which configuration (or configurations) of the values of the independent variable was used to obtain these values. It may be noticed, however, that for the smaller samples at least their tabulated 'critical values' for their criterion differ by only one or two units in the second decimal place from the exact values for the criterion obtained on equally spaced values of the independent variable. The latter may easily be derived from Table 2 of Srikantan (1961) (for α=0·05 and α=0·01 and $n \leq 20$) or from Table 5.1(a), (for α=0·1,0·05,0·01 and $n \leq 30$).

Tietjen, Moore and Beckman (1973) claim that the critical values at each level α of their criterion for a single outlier in linear regression on n points, given in their Table I, differ by less than

0·01 from the critical values at level $\alpha/2$ of the statistic used by Grubbs (1950) for testing for a single outlier at a specified end of a univariate sample of size (n+1), given in his Table IA. They proceed to emphasise that "critical values [for their criterion] may be obtained from Grubbs' (1950) Table IA by using his tabulated values for (n+1) in place of (n), [and with α replaced by $\alpha/2$]." A comparison of the two tables referred to (for α=0·1 and 0·05) reveals that the empirical critical values at level α in samples of size n tabulated by Tietjen, Moore and Beckman (1973) are close to those of Grubbs (1950) at level $\alpha/2$ for samples of size (n-1), and not (n+1). Further, the differences between the values tabulated are in some cases rather greater than 0·01. A comparison with the extended tables of Grubbs and Beck (1972) for $n \geq 30$ and for α=0·01 shews that while there is the same broad correspondence between the sets of critical values the actual differences are considerably larger for α=0·01 and for the larger sample sizes, (the difference is 0·083 for α=0·01 and n=60 for example).

A point of interest in this comparison between the tables of Tietjen, Moore and Beckman (1973) and Grubbs and Beck (1972) (which may be taken to include that of Grubbs (1950)) is that although the empirical values of Tietjen, Moore and Beckman are consistently close to the values of Grubbs & Beck, the former are almost consistently greater than the latter, (in fact only ten of the fifty-seven values of the former are less than the comparable values of the latter; seven of these are empirical values for α=0·01, which may be expected to be subject to large sampling errors). It is not difficult to appreciate why this should be. It may easily be seen, using (5.5a),

that the criterion of Tietjen, Moore and Beckman, $U_{(n)}$ say, is related to $T_{(n)}$ (defined in 5.2) by the equation

$$T_{(n)} = \frac{(n-3)U^2_{(n)}}{(n-2)-U^2_{(n)}} \quad .$$

It follows from (5.24) that the upper 100α percentage of $U_{(n)}$, for small values of α and n at least, is approximately

$$[(n-2)F_{1,n-3}(\alpha/n)/\{n-3+F_{1,n-3}(\alpha/n)\}]^{\frac{1}{2}}, \qquad (5.25)$$

where $F_{1,n-3}(\alpha/n)$ is the upper $100\alpha/n$ percentage point of $F_{1,n-3}$. Now if $U_{(n)}$ is the statistic used by Grubbs (1950) then $U_{(n)}$ is the maximum of a set $(U_1,U_2,\ldots,U_n)$ where for arbitrary i $U_i\sqrt{(n-2)}/\sqrt{(n-1-U_i^2)}$ follows a Student's t-distribution with (n-2) degrees of freedom (see Thompson (1935)).

It follows that the upper $100\alpha/2$ percentage point of $U_{(n-1)}$ is approximately

$$[(n-2)F_{1,n-3}(\alpha/(n-1))/\{n-3+F_{1,n-3}(\alpha/(n-1))\}]^{\frac{1}{2}}. \qquad (5.26)$$

The similarity of (5.25) and (5.26) explains the similarity of the tables of Tietjen, Moore and Beckman (1973) and Grubbs (1950), while the one being consistently greater than the other is explained by the difference in the divisors, n and (n-1) respectively, of the arguments of $F_{1,n-3}(.)$ in the two expressions.

5.5 Tests for Multiple Outliers

Consider first the case of two outliers. Suppose H_o is the hypothesis that Y is an observation of a normal random variable with mean $X\beta$ and variance $\sigma^2 I_n$, and that this is tested against the alternative hypothesis H_1 that two observations, y_i and y_j say, arise from two distinct normal distibutions with means μ_1 and μ_2 and with a common variance σ^2, the remaining (n-2) observations Y_{ij} arising from a normal distribution with mean $X_{ij}\beta$ and variance $\sigma^2 I_{n-2}$. It is easily seen that the difference in maximised log-likelihoods under the null and alternative hypotheses is

$$\tfrac{1}{2}n\log(R/R_{ij}).$$

It follows that an appropriate criterion for testing an observation as outlying is any monotonic function of

$$\max_{i<j}\{R/R_{ij}\}.$$

The pair of observations declared to be the "extreme pair", and tested as outlying, is defined to be that pair whose omission from the sample produces the greatest decrease in the residual sum of squares after fitting a model of the form (5.1).

Let

$$T_{ij} = \tfrac{1}{2}(n-m-2)(R-R_{ij})/R_{ij}, \qquad (i\neq j),$$

and define

$$T_{(n),2} = \max_{i<j}\{T_{ij}\},$$

Then $T_{(n),2}$ is a monotonic function of $\max_{i<j}\{R/R_{ij}\}$, and so is an appropriate statistic for testing the extreme pair of observations as outlying.

It is easily seen, using an extension of the argument of 5.1.2, that for arbitrary i and j, $(i \neq j)$, T_{ij} follows an F-distribution with (2,n-m-2) degrees of freedom. Further, by an argument similar to that of section 5.3, the following inequality holds;

$$P\,[T_{(n),2} > u] \leq \binom{n}{2} P[F_{2,n-m-2} > u]\ , \qquad (5.27)$$

Where $F_{2,n-m-2}$ is a variate following an F-distribution with (2,n-m-2) degrees of freedom. In particular it may be noted that the upper 100α percentage points of $T_{(n),2}$ are bounded above by the $100\alpha/\binom{n}{2}$ percentage points of $F_{2,n-m-2}$.

The generalisation to k outliers is now clear, (where it is assumed that k is 'sensibly' small in relation to n, certainly less than $\frac{1}{2}n$). For the alternative hypothesis that k observations arise from k distinct normal distribution with different means but a common variance σ^2, the remaining (n-k) observations arising from a linear model of the form (5.1), it is seen that the set of k observations declared to be the 'extreme set of size k', and tested as outlying, is defined to be that set of size k whose omission from the sample produces the greatest decrease in the residual sum of squares after fitting a model of the form (5.1).

With obvious extensions in notation, the criterion used for testing the extreme set of size k is

$$T_{(n),k} = \max\{T_{i_1 i_2 .. i_k}\} \quad \text{(where the maximum is taken over the set } (i_1 < i_2 < \ldots < i_k)),$$

where $T_{i_1 i_2 .. i_n} = (n-m-k)(R-R_{i_1 i_2 .. i_k})/(kR_{i_1 i_2 .. i_k})$.

Since for arbitrary $i_1, i_2, \ldots, i_k$, $T_{i_1 i_2 .. i_k}$ follows an F-distribution with (k,n-m-k) degrees of freedom, the following inequality for the upper tail probability of $T_{(n),k}$ holds;

$$P[T_{(n),k} > i] \leq \binom{n}{k} P[F_{k,n-m-k} > u] . \qquad (5.28)$$

Further, the upper 100α percentage points of $T_{(n),k}$ are bounded above by the upper $100\alpha/\binom{n}{k}$ percentage points of $F_{k,n-m-k}$.

It should be stated that there are practical difficulties involved in employing the criteria discussed above for the detection of multiple outliers. For even moderately large values of n and values of k larger than two or three, these difficulties may prohibit, in practice, the use of these criteria.

The greatest difficulty, of course, is identifying the set of k observations which are the most extreme. Except in the restricted class of models where the variances of the residuals are equal, it may not be obvious from a cursory examination of the data which k observations are the most extreme, and it may be necessary to calculate $\binom{n}{k}$ residual sums of squares to establish which is the extreme set. This contrasts with the case of the univariate sample, discussed in

Chapter 4, (and with the case of models with residuals having common variance), where at most only k+1 residual sums of squares need be calculated.

A second difficulty is that for large values of n and k the right hand side of inequality (5.28) may be so large as to render it valueless as an upper bound on the significance probability of an observed value of $T_{(n),k}$.

In view of these difficulties, particularly for large values of n and k, a "sequential procedure" for detecting outliers successively, analogous to the methods discussed in chapter 4, may be preferred, despite the inherent disadvantages associated with such procedures (for example the loss of power owing to 'masking').

5.6 Criteria Incorporating Independent Estimates of the Variance

In certain experiments there may be available an independent unbiased estimate, s_ν^2, of the variance σ^2, based upon ν degrees of freedom, where νs_ν^2 is distributed as χ_ν^2. For example in orthogonally designed experiments one factor might represent the differing levels of a continuously varying quantity, and it may be desired to investigate whether the observations have a polynomial regression upon this quantity. Such a situation is commonly met in problems of biological assay where, typically, it is of interest to fit a pair of straight lines, under the constraints of either parallelism or intersection at the origin, to the mean responses of numbers of subjections to various concentrations of test and standard preparations of some drug. In such situations an independent estimate of the variance σ^2 is available from the inter-subject variation at each concentration.

When an independent estimate s_ν^2 of σ^2 is available it may be desired to modify the test criterion $T_{(n)}$ of section 5.1.2 (and equivalently the criterion $t_{(n)}$ of Srikantan (1961)) to utilize the additional information. This is achieved by taking as test criterion

$$T^*_{(n)} = \max_{i=1}^{n}\{T^*_i\}$$

where $T^*_i = (n+\nu-m-1)(R-R_i)/(\nu s_\nu^2+R_i)$ (5.29)

(or equivalently in the case of Srikantan's criterion,

$$t^*_{(n)} = \max_{i=1}^{n}\{t^*_i\}$$

where $t^*_i = e_i^2/\{(R+\nu s_\nu^2)(1-x_i'(X'X)^{-1}x_i)\}$.

It is easily seen that the upper tail probability of $T^*_{(n)}$ satisfies

$$P\,[T^*_{(n)} > u] \leq nP\,[F_{1,n+\nu-m-1} > u]\,, \qquad (5.30)$$

and that the upper 100α percentage point of $T^*_{(n)}$ is bounded above by the upper $100\alpha/n$ percentage point of $F_{1,n+\nu-m-1}$.

Further it is evident that if

$$K^*_i = (n+\nu-m-1)k_i/(1-k_i) = (n+\nu-m-1)K_i/(n-m-1) \qquad (5.31)$$

(where k_i is as defined in (5.13))

then

$$T^*_i > K^*_i \text{ implies that } T^*_i > T^*_j \text{ for all } j\neq i, \qquad (5.32)$$

and that $K^* = \max_{i=1}^{n} K^*_i$ is an overall upper bound for the second largest T^*_i. It follows that a result analogous to that of (5.22) holds, and in particular the upper 100α percentage point of $T^*_{(n)}$ is equal to the upper $100\alpha/n$ percentage point of $F_{1,n+\nu-m-1}$ if the latter is greater than K^*.

In the case of polynomial regression of degrees one, two and three upon equally spaced values of the independent variable the bounds K^*_i and K^* may easily be derived from Table 5.2 using relation

(5.31).

The extension of the above results to criteria for testing simultaneously the k extreme observations as outlying are clear. In particular the test criterion is

$$T^*_{(n),k} = \max\{(n+\nu-m-k)(R-R_{i_1 i_2 .. i_k})/(\nu s_\nu^2 + R_{i_1 i_2 .. i_k})\},$$

and

$$P\,[T^*_{(n),k} > k] \leq \binom{n}{k} P\,[F_{1,n+\nu-m-k} > u]\,. \qquad (5.32)$$

Further the upper 100α percentage point of $T^*_{(n),k}$ is bounded above by the upper $100\alpha/\binom{n}{k}$ percentage point of $F_{1,n+\nu-m-k}$.

5.7 Some Examples

The following three examples concern chemical experiments to determine equilibrium constants of various bivalent metal ions (in combination with a number of complexing ligands) in electrolytic reactions. The examples are selected from a large number of similar analyses on data obtained in the course of a series of experiments performed in the Department of Chemistry at the University of Hull. These experiments involved bivalent ions of seven different metals and six different complexing ligands; the values of the equilibrium constants obtained in certain of these experiments, particularly those involving the heavier metals, suggested to the experimenter that the data might contain outliers. In any particular analysis the value of the equilibrium constant was obtained as the slope of a regression line of a dependent variable on a single independent variable. Neither a cursory examination nor a simple plot of the data would necessarily reveal possible outliers since the variances of the observations were unequal.

The independent variable in the first two examples is the free concentration of the complexing ligand in the electrolyte with a particular concentration of the metal ion; in both cases the metal involved was lead at a concentration of 0·0778 (measured in suitable units). In the third example the independent variable is the reciprocal of the concentration of the metal ion in the electrolyte in the absence of any complexing ligand; the metal involved was again lead.

In the first two examples the dependent variable was essentially

the reciprocal of the observed rate of the electrolytic reaction, in the third it was essentially the rate of the electrolytic reaction. The observed rates had previously been obtained as the slopes of regression lines fitted to subsidiary sets of data; the observed optical density of the electrolyte at specified times in the reaction was regressed upon time as the independent variable. The sets of times at which the optical density was observed were different for different values of the independent variable (the concentrations of the metal ions and complexing ligands), consequently the variances of the observations of the dependent variable (i.e. the estimated slopes of the subsidiary regression lines and their reciprocals) were unequal.

An identification of a likely cause of outliers in these particular sets of data involves an examination of the method of measuring the optical density of the electrolyte. The relevant feature of the method is that at each specified time in the reaction a small sample of the electrolyte was removed and filtered five times: should the experimenter inadvertantly filter the liquid four or six times instead (apparently a recognised laboratory error) then a spuriously high or low value of the optical density at that time is recorded. This in turn produces a spurious value of the rate of reaction (either high or low depending on whether the mistake is made towards the beginning or end of the reaction). Ostensibly it would have been preferable to attempt to identify these 'primary' outliers, and so obtain a 'corrected' value of the rate of the reaction. Typically, however, the optical density was measured only three or four times in each reaction, (the complete reaction for the lighter metals took only a few minutes). Further it was possible that there

were causes of outliers in the data other than that suggested above (for example, gross errors in the calculation of the rate of reaction) and which could only be detected by examining the data as a whole.

The function of the optical density that was plotted against time was the logarithm of the difference between the optical density of the electrolyte at the specified time and the optical density of the electrolyte upon completion of the reaction. This is of importance in the third example.

The data for the three examples are given in tables 5.3(i)-(iii). In each example the model to be fitted to the data was the standard linear model given in (5.1), with m=2 and with the first column of X taken to consist entirely of ones. The values of Y and the second column of X are given in Tables 5.3. The square roots of the diagonal elements of the matrix V, (where the variance of the errors ε is $V\sigma^2$) are given in the fourth columns of Tables 5.3.

Table 5.3(i)

Point No. i	Conc.Nitrate x_i	[Reaction Rate]$^{-1}$ y_i	Relative Standard deviations of y_i
1	0·098	1·07	0·10
2	0·198	1·85	0·09
3	0·297	2·09	0·09
4	0·398	1·49	0·10
5	0·498	4·42	0·12
6	0·599	4·72	0·10

Table 5.3(ii)

Point No. i	Conc. Chloride x_i	[Reaction Rate]$^{-1}$ y_i	Relative Standard deviations of y_i
1	0·007	0·011	0·06
2	0·011	0·017	0·07
3	0·014	0·232	0·19
4	0·022	0·323	0·08
5	0·029	0·180	0·10
6	0·036	0·348	0·09
7	0·044	0·307	0·12
8	0·058	1·168	0·13
9	0·074	0·862	0·10
10	0·092	0·989	0·24
11	0·111	1·150	0·18
12	0·129	1·480	0·28
13	0·149	1·497	0·23
14	0·168	1·450	0·17
15	0·168	2·010	0·16
16	0·187	1·818	0·22
17	0·224	2·398	0·50

Table 5.3(iii)

Point No. i	[Conc. Lead]$^{-1}$ x_i	(Reaction Rate) y_i	Relative Standard deviations of y_i
1	102·88	552·0	0·30
2	64·10	356·0	0·36
3	51·55	316·0	0·12
4	34·25	204·0	0·12
5	25·71	161·0	0·12
6	20·58	136·5	0·11
7	17·15	118·0	0·10
8	14·71	104·5	0·05
9	12·85	97·2	0·05
10	12·85	97·2	0·06
11	12·85	97·2	0·05
12	11·43	89·5	0·04
13	10·29	76·8	0·05
14	6·85	58·8	0·03
15	4·12	43·2	0·07
16	3·42	37·2	0·05
17	2·94	33·0	0·03
18	2·57	35·4	0·02
19	2·29	30·0	0·03
20	2·06	33·0	0·02

5.7.1 Example (i); Effects of Nitrate

With a sample size of 6, calculation of the statistic $T_{(6)}$ (defined in section 5.1.2) gives

$$T_{(6)} = 22{\cdot}40,$$

with this maximum occurring at point 4. That is the fourth observation is the extreme of the set (in the sense defined in section 5.1.1). To assess the evidence that this observation is an outlier it is necessary to calculate the bounds K_i (defined in section 5.2). These are given in Table 5.4(i) below. The largest K_i (i.e. the bound K) is marked with an asterisk.

Table 5.4(i)

i:	1	2	3	4	5	6
K_i:	4·4	3·1	5·0	8·2	11·9	11·9*

Entering Table 5.1(a) with n=6, α=0·1 (and m=2) gives a value of the upper bound for the 10% point of $T_{(6)}$, $T_6(0{\cdot}1;2)$ (defined in section 5.3), as

$$T_6(0{\cdot}1;2) = 23{\cdot}59.$$

Since this value is larger than the value of K, 11·9, the upper bound $T_6(0{\cdot}1;2)$ is in fact the actual 10% point of $T_{(6)}$, (see (5.2.4)).

The observed value of $T_{(6)}$, 22·40 is less than this percentage point, which indicates that there is very little evidence to shew that the extreme observation is an outlier.

The intercept and slope of the line fitted to all the data are 0·148 and 6·971 respectively.

5.7.2 Example (ii); Effects of Chloride

With a sample size of 17, calculation of $T_{(17)}$ gives

$$T_{(17)} = 20{\cdot}17,$$

with the maximum occurring at point 8. This is greater than 19·49, the upper bound, $T_{17}(0{\cdot}01;2)$, for the upper 1% point of $T_{(17)}$, Table 5.1(a)). There is thus strong evidence to indicate that the observation at point 8 is an outlier.

Omission of point 8 from the data and calculation of the statistic $T_{(16)}$ on the reduced sample of size 16 gives

$$T_{(16)} = 5{\cdot}95,$$

with the maximum occurring at point 14. Clearly therefore there is very little evidence for there being more than one outlier in the original data.

The intercepts and slopes of the lines fitted to the data omitting the outlier at point 8 are -0·046 and 10·763 respectively, (the line fitted to all the data, including the outlier, has intercept and slope -0·026 and 10·937).

It may be noted that in this example, unlike the first, it is not essential to calculate the bounds K_i to assess the significance of the observed value of the criterion $T_{(17)}$. The observed value is certainly significant at the one per cent level; for a more refined assessment the bounds K_i would be required. These are given in Table 5.4(ii) below.

Table 5.4(ii)

i:	1	2	3	4	5	6	7	8	9
K_i:	31·5*	31·5	17·9	25·1	21·0	21·2	18·3	17·0	19·2

i:	10	11	12	13	14	15	16	17
K_i:	16·6	18·9	17·6	19·7	25·0	25·0	22·8	17·9

5.7.3 Example (iii); Effects of Lead

With sample size of 20, calculation of $T_{(20)}$ gives

$$T_{(20)} = 6{\cdot}36,$$

with the maximum occurring at point 1. Clearly this provides little evidence of the presence of a single outlier in the data. The intercept and slope of the line fitted to the complete set of data are 20·511 and 5·627 respectively. The residual sum of squares after fitting the line is 11·137.

After this analysis had been completed it was discovered that

there was considerable doubt as to the validity of the observations at concentrations of the lead ion greater than 0·0778, (i.e. values of the independent variable less than 12·85). The difficulty was that at the higher concentrations of the lead ion, in the absence of the complexing ligands, the reaction rate was extremely slow, (at a concentration of ·0778 the reaction has a half-life of about 5 minutes), and was in fact very much slower than the experimenter had thought. It was therefore very probable that with these higher concentrations of the ion the reactions were incomplete when the final measurements of the optical densities of the electrolytes were made, thus invalidating the observations of the rates of the reactions for these higher concentrations of the lead ions. At the concentration of the lead ion of precisely 0·0778 particular care was taken with the experiments (in fact the experiments were performed in triplicate, points 9, 10 and 11); this was because this same concentration of the lead ion was to be used later in conjunction with the various complexing ligands, and values from these particular experiments would be needed for a different analysis.

It was therefore decided to discard the data relating to concentrations greater than 0·0778 and analyse only the data for concentrations less than or equal to this value, that is just the data for the first eleven points given in Table 5.3(iii).

Before examining this reduced set of data for outliers it is of interest to fit a regression line to this set of 11 points. The intercept and slope of this line are 28·391 and 5·294 respectively, and the residual sum of squares is 2·617. Calculation of the statistic $T_{12...20}$ of section 5.5 gives

$$T_{12\ldots20} = ((20-2-9)/9)(11{\cdot}137-2{\cdot}617)/2{\cdot}617$$

$$= 3{\cdot}26.$$

If all 20 observations were from the same linear model then this would be an observation of $F_{9,9}$. The upper 5% point of $F_{9,9}$ is 3·18. There is thus some statistical evidence to lend support to the decision (arrived at originally by purely chemical considerations) to discard the last 9 observations. Of course this does not imply that these 9 observations are outliers (in the usual sense of the term at least); with only 20 observations it would hardly be possible ever to say that 9 of them were outliers. Attention was directed specifically to this set of 9 observations because of the knowledge of gross errors in the experiments that produced them; it was not because they were the most extreme set of size 9 that they were investigated.

Examination of the set of 11 observations for the presence of outliers gives a value of $T_{(11)}$ of 146·406, with the maximum occurring at point 3. It is clear that this is highly significantly large (the upper bound for the 0·1% point is 52·0883, Table 5.1(a)), even without calculation of the bounds K_i given in Table 5.4(iii) below. Omission of point 3 from the data and calculation of $T_{(10)}$ on the reduced sample of size 10 gives

$$T_{(10)} = 37{\cdot}06,$$

with the maximum occurring at point 8. The bounds K_i for the

Table 5.4(iii)

i:	1	2	3	4	5	6
K_i:	31·5*	12·9	31·5	14·6	11·2	9·9

i:	7	8	9	10	11
K_i:	10·3	14·9	15·7	13·6	15·7,

reduced set of 10 values of the independent variable are given in Table 5.4(iv) below.

Table 5.4(iv)

i:	1	2	3	4	5	6
K_i:	24·7*	18·4	-	24·7	12·8	9·9

i:	7	8	9	10	11
K_i:	8·6	11·1	11·6	10·5	11·6

The upper bounds for the upper 1% and 0·1% points of $T_{(10)}$ are (Table 5.1(a)) 29·2452 and 62·1667 respectively; comparison with the starred value in Table 5.4(iv) shews that both of these are in fact the actual percentage points. There is thus strong evidence that the observation at point 8 is also an outlier.

Omission of point 8 and calculation of the statistic $T_{(9)}$ on the reduced sample gives

$$T_{(9)} = 2{\cdot}73,$$

with the maximum occurring at point 1, which clearly provides little or no evidence for the existence of any further outliers in the data. The intercept and slope of the line fitted to this reduced set of data on 9 points are 32·366 and 5·036 respectively, with a residual sum of squares of only 0·022.

It is interesting to see that the examination of the data on the original set of 20 points failed to reveal the two outliers at points 3 and 8 which were eventually identified by an analysis of the smaller sample. This is an example of the well-known phenomenon of 'masking'.

Chapter 6

Outliers in Multivariate Samples

This chapter considers some of the many problems encountered in the detection of outliers in multivariate samples. The criteria and methods discussed in sections 6.2 to 6.5 are immediate extensions of those described in Chapter 3 and 4 and are ones based upon the likelihood of the sample rather than solely upon intuitive considerations. In particular it is shewn that one commonly used statistic for the detection and testing of outliers in multivariate samples, the criterion proposed on intuitive grounds by Wilks (1963), is, in fact, likelihood-based under certain conditions. Further, in the case of a single outlier, it is shewn that this statistic has an intuitively appealing interpretation in terms of a one-dimensional projection of the multidimensional sample of observations. As in the previous chapter, attention is restricted to normal data, that is, in this case, to samples from multivariate normal distributions.

The following section, 6.1, first discusses some of the more general aspects of the problems of the detection of outliers in multivariate samples, in particular the problems of ordering multivariate data, and then considers some of the intuitively based methods and criteria which have been proposed previously.

6.1 Introduction

A distinguishing feature of the general problem of detecting outliers in samples from distributions of dimensions higher than one is that there is no distinction between 'one-sided' and 'two-sided' criteria, that is the extreme observation, or extreme set of observations, (with a suitable definition of 'extreme'), is tested as outlying without regard to the 'direction' in which it is extreme. A qualification must be made at this point since there are rather special and restricted situations in which it is appropriate to detect outliers by examining the marginal samples, where both 'one-sided' and 'two-sided' criteria may be employed. Such situations will be considered below.

The lack of distinction between 'one-sided' and 'two-sided' criteria in detecting outliers in multivariate samples arises because, unlike the univariate case, there is no basis for completely ordering the multivariate sample in such a way that both the 'largest' and the 'smallest' observations are inherently 'surprising' or 'questionable' in some sense. (Of course even in univariate samples the maximum and minimum are 'surprising' strictly only for certain particular forms of parent distribution, although these are the ones commonly met with in practice).

There are a wide variety of methods of ordering multivariate data, methods which result in either a complete or a partial ordering of the data. For example there are those based upon some scalar function of the observations such as the distance, or generalized distance, of the observations from some arbitrary reference point or

from the mean of the sample, (which produce a complete ordering) and methods based upon the successive concentric convex hulls of the data (which result in only a partial ordering). Barnett (1976) considers many such methods and classifies four ordering principles; in addition to the 'reduced ordering' (i.e. based upon a scalar function of the observations) and the 'partial ordering' methods described above, he distinguishes 'marginal ordering' (i.e. based upon one of the marginal samples), which of course may be considered to be a special case of reduced ordering, and 'conditional ordering' (i.e. based upon one of the marginal samples conditional upon the ordering of other marginal sets of observations). The third principle (marginal ordering) will be returned to below. The fourth principle has applications to problems of outlier detection should extra information be available on one, or more, components of the multivariate distribution, or indeed on some concomitant variable not included in the original multivariate sample. However these will not be discussed here.

While such methods of reduced and partial ordering multivariate data do exist, they nevertheless order the observations essentially on the basis of extremeness; that is, it is only the observations at one end of the ordered sample which will be tested as outliers. For example it is only the observation furthermost from the mean or those observations on the outermost convex hull that intuitively are considered 'surprising', and that are the ones to be tested as outliers, (again strictly speaking these observations are 'surprising' only for particular forms of the parent distribution). It is because the ordering of multivariate data, in general, is based essentially upon increasing degrees of extremeness that the distinction between 'one-sided' and 'two-sided' outlier detecting criteria does not arise in

multivariate samples, (with the exception of the particular situation considered below).

The situation in which it may possibly be useful to employ 'one-sided' criteria is when the marginal samples are examined separately for the presence of outliers. Such a procedure might be appropriate, for example, when the various components of each observation have been determined and recorded separately (e.g. measurements of heights and weights of people), and where it is thought outliers in the data would reflect gross errors of measurement or recording, rather than the presence of individuals from a population other than that under study (e.g. dwarves or giants). Typically such gross errors would affect only one component of an observation and so an examination of the marginal samples separately might reveal the outliers. However the outlier will be detected by such an examination only if the aberrant component observation occurs as the extreme of its particular marginal sample, whereas if there is some correlation between the various components then an examination of the complete multivariate sample as a whole might reveal the outlier even if none of its components were the extreme of its particular marginal sample.

There are a variety of methods available for the detection of outliers in multivariate samples. Firstly there are those which depend upon graphical techniques such as those described by Healy (1968), Gnandesikan and Kettenring (1972) and Rohlf (1975). These may be considered as extensions to multivariate data of the methods of Daniel (1959) and Gentleman and Wilk (1975). Secondly there are methods depending upon first reducing the dimensionality of the data,

for example by performing a principal component analysis as discussed by Hawkins (1974). In a third category are those methods which depend upon the calculation of a statistic whose value is sensitive to the presence of outliers. These methods are exemplified by the work of Wilks (1963) and Rohlf (1975). These three categories are not exclusive; Gnanadesikan and Kettenring (1972) consider combinations of methods in the first two and indeed methods in these two categories, while useful for pin-pointing possible outliers (i.e. the 'extreme' observations), are most commonly used in conjunction with the calculation of some test statistic relating to the identified possible outlier. Falling outside these three categories are the various Bayesian procedures, such as that of Guttman (1973), which are direct extensions of those employed in the univariate case.

The following sections are concerned with methods in the third category defined above. The particular statistics considered are those which are likelihood-based in the various cases of known and unknown mean and variance, and their natural extensions. These statistics are not the only ones available; as in the univariate case, there are many others which have considerable intuitive appeal. For example there is the (internally) studentized bivariate range considered by Gentle, Kodel and Smith (1975), and its obvious extension to higher dimensions. This would be particularly appropriate for the detection of a pair of outliers in multivariate data with uncorrelated components. If the components are correlated then the (externally) studentized generalised range considered by Siotani (1959) and its internally studentized equivalent would perhaps be better. The use of all of these statistics implies a definition of the extreme observation (or

extreme set of observations) in the sample as that for which the value of the statistic is most extreme (i.e. either large or small as appropriate), and it is this observation (or set of observations) which is tested as outlying. In all cases the number of outliers suspected in the sample is specified in advance and the particular test criterion used is specific to this number.

An essentially different type of intuitively based test has been discussed by Rohlf (1975), and which he refers to as the 'generalised gap test'. He proposed examination of the shortest simply connected graph (or minimum spanning tree) of the data and takes as test criterion the ratio of the square of the largest arc to the sum of the squares of all the arcs in the graph. He shews that this criterion has approximately the same distribution as the statistic discussed in Chapter 2, $T_{(n)}$, for the detection of a single upper outlier in gamma samples, (but note that Rohlf ignores the fact that the degrees of freedom parameter of the gamma distribution has to be estimated from the data). A feature of this method is that it is not specified in advance how many outliers are suspected; the observations declared as outlying are those contained in the smaller of the two sub-graphs obtained by removal of the largest arc. It is curious that Rohlf does not consider the application of this method to univariate data, where the test criterion has the simple form

$$\max_{i=1}^{n-1} (x_{(i+1)}-x_{(i)})^2 / \sum_j (x_{(j+1)}-x_{(j)})^2.$$

That is the observations tested as outlying are those separated from the body of the data by the largest interval between the ordered

observations.

While Rohlf's 'generalised gap test' has the possible advantage of avoiding any requirement to specify in advance the number of outliers suspected in the sample, it has the counterbalancing disadvantage of only detecting groups of outliers which are adjacent on the shortest simply connected graph, (i.e. groups of outliers that are 'close' to one another). In the one dimensional analogue of the test, for example, only sets of observations which occur at one end of the sample would be declared outliers, and the test would reveal only one of a pair of outliers if they occur at different ends of the sample. It is apparent also that Rohlf's 'generalised gap test' makes no allowance for any correlation between components of the parent multivariate distribution. The various likelihood-based criteria do not suffer from either of these disadvantages; the test criterion for a set of outliers is not dependent upon the proximity of the outliers to each other, and further any correlation between the components ia automatically allowed for.

It may be noted that Rohlf's 'generalised gap test' is not, as he implies, a multivariate generalisation of the various Dixon tests (Dixon (1950, 1951)).

It may be seen that the likelihood-based criteria discussed in the following sections induce partial 'orderings' of the multivariate sample outside the four categories defined by Barnett (1976). For each value m, (where m is less than the size of the sample, n) the most extreme set of size m is determined as that set for which the value of the test criterion for m observations has maximum value. For each m this defines a partial ordering on the sample as $x > y$ if $x \in M$

and $y \in \bar{M}$, where M is the extreme set of size m. It is easily seen that the most extreme set of size m is not necessarily contained in the extreme set of size m+1, (see, for an example in the univariate case, the artificial data considered in section 4.3.1). It follows that the n-1 partial orderings of the sample cannot be combined to produce an ordering of the complete sample on the basis of 'extremeness' as defined by the test criteria. That is the orderings of the sample induced by the outlier test criteria are not complete 'reduced orderings', although they are defined in terms of scalar functions of the observations, but are instead a sequence of partial orderings. The ordering induced by Rohlf's 'generalized gap test' is a partial ordering which essentially classifies each observation as either 'extreme' or 'not extreme'; whether or not there is a single extreme observation depends upon the particular sample.

The following section establishes the notation and derives the likelihood-based criteria for single outliers in various cases. The succeeding sections consider the extensions to criteria for multiple outliers and some further properties of these statistics.

6.2 Likelihood-Based Test Criteria for Single Outliers

Let $x_j' = (x_{1j},\ldots,x_{pj})$, $j=1,2,\ldots,n$, be a random sample of n observations of a p-dimensional random variable.

Let $\bar{x}_i = n^{-1}\sum_{j=1}^{n} x_{ij}$, $(1 \leq i \leq p)$, the mean of the i^{th} components of the n observations, let $\bar{x}$ be the $p \times 1$ vector $(\bar{x}_1, \bar{x}_2,\ldots,\bar{x}_p)'$, and let X be the $p \times n$ matrix with $(i,j)^{th}$ component $(x_{ij}-\bar{x}_i)$. For each r, $1 \leq r \leq n$, let

$$\bar{x}_{i,(r)} = (n-1)^{-1}\sum_{\substack{j=1 \\ j\neq r}}^{n} x_{ij}, \quad (1 \leq i \leq p),$$

the mean of the i^{th} components of all but the r^{th} of the n observations, let $\bar{x}_{(r)} = (\bar{x}_{1,(r)},\bar{x}_{2,(r)},\ldots,\bar{x}_{p,(r)})'$, and let $X_{(r)}$ be the $p \times (n-1)$ matrix with components $(x_{ij}-\bar{x}_{i,(r)})$, $(1 \leq i \leq p,\ 1 \leq j \leq n,\ j\neq r)$. Similarly define the $p \times (n-k)$ matrix $X_{(r_1\ldots r_k)}$ upon omission of the k observations $x_{r_1,\ldots,r_k}$. Further if A is a square matrix let $|A|$ be the determinant of A. Throughout the following sections the null hypothesis, referred to as H_o, will be that the n observations are a random sample from a p-dimensional multivariate normal distribution with mean μ and variance Λ, where μ is a $p \times 1$ vector and Λ is a $p \times p$ positive-definite symmetric matrix. This will be tested against various alternative hypotheses.

6.2.1 The case of both μ and Λ known

This case, though of minor practical importance is considered here briefly for completeness. It generalises the results of 3.2.

Under H_o the log-likelihood of the sample is

$$-\tfrac{1}{2}np\log(2\pi)-\tfrac{1}{2}n\log(|\Lambda|)-\tfrac{1}{2}\sum_{i=j}^{n}(x_i-\mu)'\Lambda^{-1}(x_i-\mu).$$

Let H_1 be the alternative hypothesis that one observation, x_n without loss of generality, arises from a p-dimensional normal distribution $N(\mu^*,\Lambda)$, and the remaining observations arise from the distribution $N(\mu,\Lambda)$. Under H_1 the maximised log-likelihood is, upon substitution of $\hat{\mu}^* = x_n$,

$$-\tfrac{1}{2}np\log(2\pi)-\tfrac{1}{2}n\log\{|\Lambda|\}-\tfrac{1}{2}\sum_{i=1}^{n-1}(x_i-\mu)'\Lambda^{-1}(x_i-\mu).$$

The difference in maximised log-likelihoods is thus $(x_n-\mu)'\Lambda^{-1}(x_n-\mu)$, and it follows that the likelihood- based criterion for testing the 'extreme' observation as outlying is $Z_{(n)}$ where

$$Z_{(n)} = \max_{i=1}^{n}\{(x_i-\mu)'\Lambda^{-1}(x_i-\mu)\},$$

and where the 'extreme' observation is that for which $(x_i-\mu)'\Lambda^{-1}(x_i-\mu)$ is maximum, i.e. that observation whose generalized distance from the population mean μ is the greatest.

For arbitrary i, $(x_i-\mu)'\Lambda^{-1}(x_i-\mu)$ follows a χ^2 distribution with p degrees of freedom. Thus the following inequality for the upper tail probability of $Z_{(n)}$ holds

$$P\,[Z_{(n)} >> u] \;\leq\; n\,P\,[\chi_p^2 > u]\,, \qquad (6.1)$$

In particular the upper 100α percentage points of $Z_{(n)}$ are bounded above by the upper $(100\alpha/n)$ percentage points of χ_p^2.

6.2.2 The case μ unknown and Λ known

Under H_o (with Λ known) the maximised log-likelihood is

$$-\tfrac{1}{2}np\log(2\pi)-\tfrac{1}{2}n\log\{|\Lambda|\}-\tfrac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})'\Lambda^{-1}(x_i-\bar{x}).$$

Under the alternative hypothesis, H_1, that one observation x_n say, arises from $N(\mu^*,\Lambda)$ and the other $(n-1)$ observations are from $N(\mu,\Lambda)$ the maximised log-likelihood is

$$-\tfrac{1}{2}np\log(2\pi)-\tfrac{1}{2}n\log(|\Lambda|)-\tfrac{1}{2}\sum_{i=1}^{n-1}(x_i-\bar{x}_{(n)})'\Lambda^{-1}(x_i-\bar{x}_{(n)}).$$

The difference between these maximised log-likelihoods is

$$\tfrac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})'\Lambda^{-1}(x_i-\bar{x})-\tfrac{1}{2}\sum_{i=1}^{n-1}(x_i-\bar{x}_{(n)})'\Lambda^{-1}(x_i-\bar{x}_{(n)})$$

$$= \tfrac{1}{2}n(x_n-\bar{x})'\Lambda^{-1}(x_n-\bar{x})/(n-1).$$

It follows that a likelihood-based statistic for testing the 'extreme' observation as outlying is $u_{(n)}$ where

$$u_{(n)} = \max_{i=1}^{n}\{(x_i-\bar{x})'\Lambda^{-1}(x_i-\bar{x})\},$$

and where in this case the 'extreme' observation is that observation whose generalised distance from the sample mean is the greatest.

Since, for arbitrary i, $n(x_i-\bar{x})'\Lambda^{-1}(x_i-\bar{x})/(n-1)$ follows a χ^2 distribution with p degrees of freedom the following inequality holds for the upper tail probability of $u_{(n)}$;

$$P\,[u_{(n)} > u] \leq n\,P\,[\chi^2_p > nu/(n-1)]\,. \qquad (6.2)$$

In particular the upper 100α percentage points of $u_{(n)}$ are bounded above by $\{(n-1)/n\}\chi^2_p(\alpha/n)$, where $\chi^2_p(\alpha/n)$ is the upper $100\alpha/n$ percentage point of χ^2_p.

6.2.3 The case μ known Λ unknown

Under the null hypothesis H_o (with μ known) the log-likelihood is maximised when $\Lambda=1/n\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)'$, giving a maximised log-likelihood under H_o of

$$-\tfrac{1}{2}np\log(2\pi)-\tfrac{1}{2}n\log\{\left|\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)'\right|\}+\tfrac{1}{2}n\log(n)-\tfrac{1}{2}n.$$

Under the alternative hypothesis, H_1, that all observations other than x_n arise from the normal distribution $N(\mu,\Lambda)$ and x_n arises from a normal distribution $N(\mu^*,\Lambda)$, where μ^* is to be estimated, the log-likelihood is maximised when $\Lambda=1/(n-1)\sum_{i=1}^{n-1}(x_i-\mu)(x_i-\mu)'$ and $\mu^*=x_n$, giving a maximised log-likelihood of

$$-\tfrac{1}{2}np\log(2\pi)-\tfrac{1}{2}n\log\{\left|\sum_{i=1}^{n-1}(x_i-\mu)(x_i-\mu)'\right|\}+\tfrac{1}{2}n\log(n)-\tfrac{1}{2}n.$$

The difference between the maximised log-likelihoods is therefore

$$\tfrac{1}{2}n\log\{\left|\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)'\right|/\left|\sum_{i=1}^{n-1}(x_i-\mu)(x_i-\mu)'\right|\}$$

$$= -\tfrac{1}{2}n\log\{1-T_n\}$$

where $T_j = (x_j-\mu)'\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)'\right)^{-1}(x_j-\mu)$, $(j=1,2,\ldots,n)$.

Thus a likelihood-based statistic for testing the 'extreme' observation as outlying is $T_{(n)}$ where $T_{(n)} = \max_{j=1}^{n}(T_j)$, and where in this case the 'extreme' observation is that observation whose (internally) studentized generalised distance from the population mean is the greatest.

Now

$$(x_j-\mu)'\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)'\right)^{-1}(x_j-\mu) = t_j/(1+t_j),$$

where

$$t_j = (x_j-\mu)'\left(\sum_{\substack{i=1\\ i\neq j}}^{n}(x_i-\mu)(x_i-\mu)'\right)^{-1}(x_j-\mu),$$

and where, for arbitrary j, $(n-p)t_j/p$ follows an F-distribution with p and (n-p) degrees of freedom. Hence the following inequality holds for the upper tail probability of $T_{(n)}$;

$$P\,[T_{(n)} > u] \leq n\,P\,[F_{p,n-p} > \{(n-p)/p\}\{u/(1-u)\}]. \qquad (6.3)$$

In particular the upper 100α percentage point of $T_{(n)}$ is bounded above by

$$\{pF_{p,n-p}(\alpha/n)\}/\{n-p+pF_{p,n-p}(\alpha/n)\},$$

where $F_{p,n-p}(\alpha/n)$ is the upper $(100\alpha/n)$ percentage point of $F_{p,n-p}$.

6.2.4 The case when both μ and Λ are unknown

Under the null hypothesis H_o the log-likelihood is maximised when $\mu=\bar{x}$ and $\Lambda=1/n\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})' = XX'/n$, giving a maximised log-likelihood of

$$-\tfrac{1}{2}np\log(2\pi)-\tfrac{1}{2}n\log\{|XX'|\}+\tfrac{1}{2}n\log(n)-\tfrac{1}{2}n.$$

Under the alternative hypothesis, H_1, that all observations other than x_n arise from a normal distribution $N(\mu,\Lambda)$ and x_n arises from a normal distribution $N(\mu^*,\Lambda)$ the maximised log-likelihood is

$$-\tfrac{1}{2}np\log(2\pi)-\tfrac{1}{2}n\log\{|X_{(n)}X'_{(n)}|\}+\tfrac{1}{2}n\log(n)-\tfrac{1}{2}n.$$

The difference between the maximised log-likelihoods is thus

$$\tfrac{1}{2}n\log\{|XX'|/|X_{(n)}X'_{(n)}|\}.$$

Hence a likelihood-based criterion for testing the 'extreme' observation as outlying is

$$\max_{i=1}^{n}\{|XX'|/|X_{(i)}X'_{(i)}|\},$$

where the extreme observation is that observation which produces the maximum. Clearly this criterion is equivalent to

$$r_1 = \min_{i=1}^{n}\{|X_{(i)}X'_{(i)}|/|XX'|\},$$

which is the criterion proposed by Wilks (1963) and which he terms the minimum 'one-outlier scatter ratio'.

Since

$$|X_{(i)}X'_{(i)}|/|XX'| = 1-n(x_i-\bar{x})'(XX')^{-1}(x_i-\bar{x})/(n-1), \qquad (6.4)$$

the extreme observation may alternatively be defined as that observation whose (internally) studentized generalised distance from the sample mean is the greatest. If $U_i = (x_i-\bar{x})'(XX')^{-1}(x_i-\bar{x})$ and

$$U_{(n)} = \max_{i=1}^{n}\{U_i\},$$

then the statistics r_1 and $U_{(n)}$ are related by the equation $r_1 = 1-nU_{(n)}/(n-1)$, and either may be used to test the extreme observation as outlying.

An extension of the argument used in 3.5.1 shews that, for arbitrary i,

$$U_i = \{(n-1)/n\}/\{1+(n-p-1)/pF_{p,n-p-1}\},$$

where $F_{p,n-p-1}$ is a variate following an F-distribution with p and (n-p-1) degrees of freedom. Application of identity (6.4) shews that, for arbitrary i, $|X_{(i)}X'_{(i)}|/|XX'|$ follows a beta distribution $B(\frac{1}{2}(n-p-1),\frac{1}{2}p)$, which is the result of Wilks (1962). Using this latter result Wilks (1963) calculates lower bounds for the lower 100α percentage points, of r_1 for α=0·01, 0·025, 0·05, 0·1, p=1(1)5, and n=5(1)30(5)100(100)500, and shews generally that

$$P\,[r_1 < u] \leq n\,p\,[\beta_{\frac{1}{2}(n-p-1),\frac{1}{2}p} < u]\ . \qquad (6.5a)$$

These lower bounds for r_1 may be used to derive equivalent upper bounds for $U_{(n)}$, and in particular it follows from (6.5a) that

$$P\,[U_{(n)} < u] \leq n\,P\,[F_{p,n-p-1} > \{(n-p-1)/p\}\{nu/(n-1-nu)\}]. \qquad (6.5b)$$

6.3 Likelihood-based test criteria for Multiple Outliers

In this section the criteria and the definitions of extreme observations, discussed in the previous section, are extended to the cases of multiple outliers. The null hypothesis H_o is tested against the alternative hypothesis (referred to throughout this section as H_1) that k observations arise from k distinct normal distributions, each with a common variance Λ and with unknown means $\mu_1^*, \mu_2^*, \ldots, \mu_k^*$, and that the remaining n-k observations are from a normal distribution $N(\mu,\Lambda)$. The various cases of μ and Λ known and unknown are considered separately. It is assumed that k is 'small' in relation to n, certainly less than both $\frac{1}{2}n$ and $\frac{1}{2}(2n-p(p+1))$.

In the case of known mean and variance it is easily seen that for the alternative hypothesis, H_1, a likelihood-based criterion for testing the k most extreme observations as outlying is

$$Z_{(n),k} = \sum_{i=1}^{k} Z_{(n-k+i)},$$

where $Z_{(1)} < Z_{(2)} < \ldots < Z_{(n)}$ are the ordered values of $Z_i = (x_i-\mu)'\Lambda^{-1}(x_i-\mu)$, $i=1,2,\ldots,n$. The k most extreme observations are defined as those k observations whose generalised distances from the population mean are the greatest. Clearly

$$P\,[Z_{(n),k} > u] \leq \binom{n}{k} P\,[\chi^2_{kp} > u] \qquad (6.6)$$

and further the upper 100α percentage points of $Z_{(n),k}$ are

bounded above by the upper $100\alpha/\binom{n}{k}$ percentage points of χ^2_{kp}.

In the case of unknown mean and known variance it may readily be seen that for the alternative hypothesis H_1 stated above (where now μ is unknown and to be estimated, the likelihood-based criterion for testing the k 'most extreme' observations as outlying is

$$u_{(n),k} = \max_{\tau}\{\sum_{i=1}^{k}(x_{\tau(i)}-\bar{x})'\Lambda^{-1}(x_{\tau(i)}-\bar{x})+(n-k)^{-1}(\sum_{i=1}^{k}(x_{\tau(i)}-\bar{x}))'\Lambda^{-1}(\sum_{i=1}^{k}(x_{\tau(i)}-\bar{x})\} \tag{6.7}$$

where $\{\tau(1),\tau(2),\ldots,\tau(n)\}$ is a permutation on the first n positive integers, and where the k 'most extreme' observations are those for which the expression in (6.7) is maximum, over all permutations τ (notice these are not necessarily the k observations whose generalised distances from the population mean are the greatest).

In the case of known mean and unknown variance the likelihood-based criterion for the alternative hypothesis H_1 stated above (where now μ is known and Λ is unknown and to be estimated) for testing the k 'most extreme' observations as outlying is $T_{(n),k} = \min_{\tau}(T_\tau)$, where T_τ is the ratio of determinants given by

$$T_\tau = |\sum_{i=1}^{n-k}(x_{\tau(i)}-\mu)(x_{\tau(i)}-\mu)'|/|\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)'| , \tag{6.8}$$

and where τ is a permutation on the first n integers.

In this case the k 'most extreme' observations are those for which T_τ is minimum. Now

$$T_\tau = |A_\tau|/|\{A + \sum_{i=n-k+1}^{n}(x_{\tau(i)}-\mu)(x_{\tau(i)}-\mu)'\}|,$$

where A_τ is the $p \times p$ matrix $\sum_{i=1}^{n-k}(x_{\tau(i)}-\mu)(x_{\tau(i)}-\mu)'$. Since A follows a Wishart distribution $W(p,n-k,\Lambda)$ and since, for each $i=n-k+1,\ldots,n$, the $(x_{\tau(i)}-\mu)$ are independently and identically normally distributed, $N(0,\Lambda)$, each independently of A_τ, it follows (see for example Wilks (1962) p. 562) that, for arbitrary τ, T_τ is distributed as the product of k independent random variables having beta distributions $B(\frac{1}{2}(n-p+1-i),\frac{1}{2}p)$, $i=1,2,\ldots,k$ (or equivalently as the product of p independent random variables having beta distributions $B(\frac{1}{2}(n-k+1-i),\frac{1}{2}k)$, $i=1,2,\ldots,p$). It is thus possible to obtain lower bounds for the lower percentage points of $T_{(n),k}$, using the inequality

$$P\,[T_{(n),k} < u] \leq \binom{n}{k} P\,[T < u], \tag{6.9}$$

where T has the distribution above.

In the case when both the mean and the variance are unknown the likelihood-based criterion, for the alternative hypothesis H_1 stated above (where now both μ and Λ are to be estimated), for testing the k most extreme observations as outlying is $r_k = \min_\tau\{R_\tau\}$, where R is the ratio of determinants

$$R_\tau = |X_{(\tau(1)\ldots\tau(k))}X'_{(\tau(1)\ldots\tau(k))}|/|XX'|,$$

and where τ is a permutation on the first n integers. R_τ is the 'k-outlier scatter ratio' of Wilks (1963). Wilks (1963) shews that for arbitrary τ, R_τ is distributed as the product of k independent random variables having beta distributions $B(\frac{1}{2}(n-p-i),\frac{1}{2}p)$, $i=1,2,\ldots k$,

and in the case k=2 he derives lower bounds for the lower 100α percentage points of r_2 for α=0·01, 0·025, 0·05, 0·1, p=1(1)5 and n=5(1)30(5)100(100)500.

It should be noted that the practical difficulties involved in applying the criteria discussed above may prohibit their use for large values of n and k. In particular the identification of the most extreme set of size k may, except in the simple case of known mean and variance, involve a considerable amount of computation. Further, the upper bounds for the tail probabilities of the various criteria given in (6.6) and (6.9) may be so large as to render them valueless. It may therefore be preferable to adopt a 'sequential' procedure and attempt to identify outliers successively (or possibly in small groups of, say, two or three at a time). The practical advantages of such a 'sequential' procedure may outweigh the inherent disadvantages such as loss of power owing to masking, but note that the subset of observations eventually declared as outlying might not be the most extreme subset of that size as defined by an objective criterion.

6.4 Criteria Incorporating an Independent Estimate of the Variance

In this section consideration is given to the modifications that may be made to the criteria discussed in the previous two sections when there is available an unbiased estimate, S say, of the population variance Λ. It is assumed that S is distributed independently of the sample, so that a 'pooled' estimate of Λ may be obtained by combining S with the estimate of Λ derived from the sample. It is assumed further that S is a symmetric $p \times p$ matrix such that νS follows a Wishart distribution $W(p,\nu,\Lambda)$.

Consider first the case of a single outlier. When the population mean μ is known, the statistic $T_{(n)}$ of section 6.2.3 may be modified to

$$T^*_{(n)} = \max_{j=1}^{n}\{T^*_j\}, \text{ where } T^*_j = (x_j-\mu)'\Big(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)'+\nu S\Big)^{-1}(x_j-\mu).$$

It may be noticed that this implies a corresponding modification of the definition of the 'extreme' observation as that whose (pooled) studentized generalised distance from the population mean is the greatest. It is evident that for arbitrary j, $\{(n-p+\nu)/p\}\{T^*_j/(1-T^*_j\}$ follows an F-distribution with p and $(n-p+\nu)$ degrees of freedom. Thus the following inequality for the upper tail probability of $T^*_{(n)}$ holds;

$$P\,[T^*_{(n)} > u] \leq n\,P\,[F_{p,n-p+\nu} > \{(n-p+\nu)/p\}\{u/(1-u)\}]. \qquad (6.10)$$

Further the upper 100α percentage points of $T^*_{(n)}$ are bounded above by

$$\{pF_{p,n-p+\nu}(\alpha/n)\}/\{n-p+\nu+pF_{p,n-p+\nu}(\alpha/n)\},$$

where $F_{p,n-p+\nu}(\alpha/n)$ is the upper $100(\alpha/n)$ percentage point of $F_{p,n-p+\nu}$.

In the case when both the population mean and variance are unknown the statistic $U_{(n)}$ of section 6.2.4 may be modified to $U^*_{(n)} = \max_{i=1}^{n} U^*_i$, where $U^*_i = (x_i-\bar{x})'(XX'+S)^{-1}(x_i-\bar{x})$, $i=1,2,\ldots,n$, (with the corresponding modification of the definition of the extreme observation). Clearly this is equivalent to modifying the statistic r_1 of 6.2.4 to r^*_1 where

$$r^*_1 = \min_{i=1}^{n}\{|X_{(i)}X'_{(i)}+\nu S|/|XX'+\nu S|.$$

Further the following equivalent inequalities hold for the tail probabilities of $U^*_{(n)}$ and r^*_1;

$$P[r^*_1 < u] \leq n\, P[\beta_{\frac{1}{2}(n+\nu-p-1),\frac{1}{2}p} < u] \qquad (6.11a)$$

$$P[U^*_{(n)} > u] \leq n\, P[F_{p,n+\nu-p-1} > \{(n+\nu-p-1)/p\}\{nu/(n-1-nu)\}]. \qquad (6.11b)$$

In the case of multiple outliers the modifications $T^*_{(n),k}$ and r^*_k of the statistics $T_{(n),k}$ and r_k, defined in section 6.3, are clear. It is evident that $T^*_{(n),k} = \min_{\tau}\{T^*_\tau\}$, where, for arbitrary τ, T^*_τ is distributed as the product of k independent random variables having beta distributions $B(\frac{1}{2}(n+\nu-p+1-i),\frac{1}{2}p)$ $i=1,2,\ldots,k$. Further $r^*_k = \min_{\tau}\{R^*_\tau\}$ where, for arbitrary τ, R^*_τ is distributed as the product of k independent random variables having beta distributions

$B(\frac{1}{2}(n+\nu-p-i),\ i=1,2,\ldots,k.$

Different modifications of the likelihood-based criteria are of use when a 'sequential' procedure is adopted for the detection of multiple outliers. These are the 'externally studentized' criteria, whose use avoids the problems of masking. For example when the population mean is known $T_{(n)}$ may be modified to

$$\max_{i=1}^{n}\{(x_i-\mu)'S^{-1}(x_i-\mu)/\nu\},$$

which is the maximum of a set of n variates each distributed as $p/(\nu-p+1)F_{p,\nu-p+1}$. When the mean is unknown $U_{(n)}$ may be modified to

$$\max_{i=1}^{n}\{(x_i-\bar{x})'S^{-1}(x_i-\bar{x})/\nu\},$$

which is the maximum of a set of n variates each distributed as $\{(n-1)/n\}\{p/(\nu-p+1)\}F_{p,\nu-p+1}$. This statistic is essentially identical to that considered by Siotani (1959). The 'externally studentized' versions of the criteria for multiple outliers may readily be derived, but they are of little practical importance.

6.5 q-Dimensional Projections of the Sample

Wilks (1963) proposed the criteria r_k, for testing for k outliers in p-dimensional samples with unknown mean and variance, on purely intuitive grounds. He shews that r_k has a geometric interpretation as the ratio of two sums of squares of volumes of (p+1)-simplexes (each simplex being formed by p observations together with the sample mean,) the first sum excluding the k suspect observations and second including them. In sections 6.2 and 6.3 it was shewn that the criteria could also be regarded as likelihood-based under certain conditions. In this section it is shewn that in the case of a single outlier the criterion r_1 has another intuitively appealing property in terms of a q-dimensional projection of the p-dimensional sample.

Let A be a $q \times p$ matrix (where $q \leq p$ and A is assumed to be of full rank q) and let $y_j = Ax_j$ $(j=1,2,\ldots,n)$. Then y_j, $j=1,2,\ldots,n$, is a q-dimensional projection of the sample x_j, $j=1,2,\ldots,n$. Let $Y = AX$ and $Y_{(i)} = AX_{(i)}$.

Let $R_i = |X_{(i)}X'_{(i)}|/|XX'|$, the 'one-outlier scatter ratio' of the i^{th} observation of the sample, and let $R_i(A) = |Y_{(i)}Y'_{(i)}|/|YY'|$, the 'one-outlier scatter ratio' of the i^{th} observation of the projected sample. Let $r_1 = \min_{i=1}^{n}\{R_i\}$. It will be shewn that if A_i is that projection which minimises $R_i(A)$ for all projections A then $R_i(A_i) = R_i$, that is the value of the one-outlier scatter ratio of the i^{th} observation of the projected sample (where the projection is chosen to minimise that ratio) is equal to the one-outlier scatter ratio of the i^{th} observation of the original p-dimensional sample.

Now $R_i(A) = 1-ny'_i(YY')^{-1}y_i/(n-1)$.

$$= 1-nx_i'A'(AXX'A')^{-1}Ax_i/(n-1).$$

Suppose Δ is an elemental increment in A, then

$$(n-1)\{1-R_i(A+\Delta)\}/n = x_i'(A'+\Delta')((A+\Delta)XX'(A'+\Delta'))^{-1}(A+\Delta)x_i$$

$$= x_i'(A'+\Delta')(I_q+G^{-1}(\Delta XX'A'+AXX'\Delta'))^{-1}G^{-1}(A+\Delta)x_i,$$

where $G = AXX'A'$, (to terms in the first order of Δ),

$$= x_i'(A'G^{-1}A+2(I_p-A'G^{-1}AXX')\Delta'G^{-1}A)x_i.$$

Thus

$$(n-1)\{R_i(A+\Delta)-R_i(A)\}/n = 2x_i'(A'G^{-1}AXX'-I_p)\Delta'G^{-1}Ax_i,$$

(to terms in the first order of Δ).

If A_i is such that $R_i(A_i+\Delta) = R_i(A_i)$ for all elemental increments Δ then

$$A_i'(A_iXX'A_i')^{-1}A_iXX' = I_p$$

so

$$A_i'(A_iXX'A_i')^{-1}A_i = (XX')^{-1}$$

and $R_i(A_i) = R_i$ as asserted.

(It may be noted that A_i is not uniquely determined since of C is any $q \times q$ non-singular matrix it is easily verified that

$R_i(CA_i) = R_i(A_i)$).

In particular $\min_{i=1}^{n} R_i(A_i) = r_1$, and it follows that a single outlier in a p-dimensional sample can be detected by examining all q-dimensional projections of that sample, where q is fixed and $1 \leq q \leq p$; for each projection A the minimum one-outlier scatter ratio, $r_1(A)$, say, is computed on the projected sample, if A* is that projection for which

$$r_1(A^*) = \min_{A}\{r_1(A)\}$$

then $r_1(A^*) = r_1$ and further if A^*x_m is the extreme observation in the q-dimensional projected sample A^*x_j, j=1,2,...,n, then x_m is the extreme observation in the p-dimensional sample x_j, j=1,2,...,n.

Consideration of the particular case j=1 illustrates the intuitive appeal of this result. The most extreme observation in a p-dimensional sample can be detected by examining the sample from all possible 'viewpoints' (i.e. by examining all one-dimensional projections of the sample) and noting which particular 'viewpoint' (or one-dimensional projection) produces the smallest (most extreme) value of the test criterion for the extreme observation in the sample, as viewed one-dimensionally, (i.e. in the one-dimensional projection of the sample). The extreme observation in the p-dimensional sample is then pin-pointed as that observation corresponding to the extreme observation in the sample as viewed from that particular viewpoint defined above. Furthermore the criterion for testing that extreme observation as outlying may be calculated from that one-dimensional projection of the sample.

It is easily seen that a corresponding result is not available for the cases of two or more outliers. Consider for example the case p=2 and q=1, and suppose the sample of n=4m observations is such that there are m values each of (1,0), (0,1), (-1,0) and (0,-1). The two-outlier scatter ratio for the two observations (1,0) and (0,1) is (n-4)/n. It is readily seen that the minimum value of the two-outlier scatter ratio for these two points, when the sample is projected onto a straight line, occurs when the sample is projected onto the line x=y and that in this case the two-outlier scatter ratio has the value (n-4)/(n-2). It follows that it is not, in general, possible to detect multiple outliers in multidimensional samples by examining one-dimensional projections of the sample.

BIBLIOGRAPHY

ANDREWS, D.F.(1971) "Significance Tests Based on Residuals". Biometrika 58, 139-148.

ANSCOMBE, F.J. (1960)"Rejection of Outliers". Technometrics 2, 123-147.

ANSCOMBE, F.J. (1960a). "Discussion of the Papers of Messrs Anscombe and Daniel". Technometrics 2, 157-166.

ANSCOMBE, F.J. (1961)"Examination of Residuals". Proc. 4th Berk.Symp., 1-36.

ANSCOMBE, F.J. and TUKEY, J.W. (1963). "The Examination and Analysis of Residuals". Technometric 5, 141-160.

ANSCOMBE, F.J. (1967). "Topics in the Investigation of Linear Relations fitted by the Method of Least Squares". Jour. Roy. Statist. Soc. B, 29, 1-29; discussion 29-52.

BARNETT, V. (1976) "The Ordering of Multivariate Data", Jour. Roy. Statist.Soc. A, 139, Part 3.

BASU, A.P. (1968) "On Some Tests of Hypotheses Relating to the Exponential Distribution When Some Outliers are Present". Jour.Amer.Statist.Assoc. 63, 548-559.

BASU, D. (1955). "On Statistics Independent of a Complete Sufficient Statistic". Sankhyā. 15, 377-380.

BEHNKEN, D.W. and DRAPER, N.R. (1972) "Residuals and Their Variance Patterns". Technometrics 14, 101-111.

BLISS, C.I., COCHRAN, W.G., and TUKEY, J.W. (1956) "A Rejection Criterion Based upon the Range". Biometrika 43, 418-422.

BOX, G.E.P., and TIAO G. C. (1968). "A Bayesian Approach to Some Outlier Problems". Biometrika 55, 119-129.

BROSS, I.D.J. (1961) "Outliers in Patterned Experiments: A Strategic Appraisal". Technometrics 3, 91-102.

BROWN, B.M. (1975) "A Short-Cut Test for Outliers Using Residuals". Biometrika 62, 623-629.

CHAUVENET, W. (1863)."A Manual of Spherical and Practical Astronomy". Philadelphia.

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.

COLLETT, D. and LEWIS, T. (1976). "The Subjective Nature of Outlier Rejection Procedures". Appl. Statist. 25, Part 3.

COX, D.R., and SNELL, E.J. (1968). "A General Definition of Residuals". Jour. Roy.Statist.Soc. B, 30, 248-75.

COX, D.R., and SNELL, E.J. (1971). "On Test Statistics Calculated from Residuals". Biometrika, 58, 589-594.

DANIEL, C. (1959). "Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments". Technometrics 1, 311-341.

DANIEL, C. (1960). "Locating Outliers in Factorial Experiments". Technometrics 2, 149-156.

DARLING, D.A. (1952)."On a Test for Homogeneity and Extreme Values". Ann.Math.Statist. 23, 450-456.

DAVID, H.A. (1956a). "On the Application to Statistics of an Elementary Theorem in Probability". Biometrika 43, 85-91.

DAVID, H.A. (1956b). "Revised Upper Percentage Points of the Extreme Studentized Deviate from the Sample Mean". Biometrika 43, 449-451.

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.

DAVID, H.A., and PAULSON, A.S. (1965). "The Performance of Several Tests for Outliers". Biometrika 52, 429-436.

DEMPSTER, A.P., and ROSNER, B. (1971). "Detection of Outliers and Related Topics". Statistical Decision Theory and Related Topics. Edited by Gupta, S.S. and Yackel, J. 161-180.

DIXON, W.J. (1950) "Analysis of Extreme Values". Ann.Math.Statist. 21, 488-506.

DIXON, W.J. (1951). Ratios Involving Extreme Values". Ann.Math.Statist. 22, 68-78.

DIXON, W.J. (1953). "Processing Data for Outliers". Biometrics 9, 74-89.

DIXON, W.J. (1960). "Simplified Estimation from Censored Normal Samples". Ann.Math.Statist. 31, 385-391.

EISENHART, C., HASTAY, M.W., and WALLIS, W.A. (1947)
Selected Techniques of Statistical Analysis.
New York: McGraw-Hill.

ELASHOFF, Janet Dixon (1972). "A Model for Quadratic Outliers in Linear Regression".
Jour.Amer.Statist.Assoc. 67, 478-485.

ELLENBERG, J.H. (1973)"The Joint Distribution of the Standardised Least Squares Residuals from a General Linear Regression".
Jour.Amer.Statist.Assoc. 68, 941-943.

EPSTEIN, B. (1960a). "Tests for the Validity of the Assumption that the Underlying Distribution of Life is Exponential: Part 1".
Technometrics 2, 83-101.

EPSTEIN, B. (1960b). "Tests for the Validity of the Assumption that the Underlying Distribution of Life is Exponential: Part 2".
Technometrics 2, 167-183.

EPSTEIN, B., and SOBEL, M. (1953). "Life Testing".
Jour.Amer.Statist.Assoc. 48, 486-502.

EPSTEIN, B and SOBEL, M. (1954). "Some Theorems Relevant to Life Testing from an Exponential Distribution".
Ann.Math.Statist. 25, 373-381.

FERGUSON, T.S. (1960)."Discussion of the Papers of Messrs. Anscombe and Daniel".
Technometrics 2, 157-166.

FERGUSON, T.S. (1961). "On the Rejection of Outliers".
Proc. 4th Berkeley Symp.I, 253-287.

FERGUSON, T.S. (1961a). "Rules for Rejection of Outliers". Rev.Inst.Internat.Statist. 29, 29-43.

deFINETTI, B. (1961) "The Bayesian Approach to the Rejection of Outliers". Proc. 4th Berk.Symp. 199-210.

FINNEY, D. J. (1974)."Problems, Data, and Inference". Journ.Roy.Statist.soc.A, 137, 1-23.

FISHER, R.A. (1929). "Tests of Significance in Harmonic Analysis". Proc.Roy.Soc.A, 125, 54-59.

FISHER, R.A. (1940) "On the Similarity of the Distribution Found for the Test of Significance in Harmonic Analysis, and in Stevens's Problem in Geometric Probability". Ann.Eugen. X, 14-17.

FOX, A.J. (1972). "Outliers in Time Series", Jour.Roy.Statist.Soc. B, 34, 350-363.

GEBHARDT, F. (1964). "On the Risk of Some Strategies for Outlying Observations". Ann.Math.Statist. 35, 1524-1536.

GEBHARDT, F. (1966). "On the Effect of Stragglers on the Risk of some Mean Estimators in Small Samples". Ann.Math.Statist. 37, 441-450.

GENTLE, J.E., KODELL, R.L., and SMITH, P.L. (1975). "On the Distribution of the Studentized Bivariate Range". Technometrics, 17, 501-505.

GENTLEMAN, J.F., and WILK, M.B. (1975). "Detecting Outliers in a Two-Way Table: 1 Statistical Behaviour of Residuals". Technometrics, 17, 1-14.

GNANADESIKAN, R., and KETTENRING, J.R. (1972)
"Robust Estimates, Residuals, and Outlier Detection with Multi-Response Data".
Biometrics 28, 81-124.

GODWIN, H.J., (1945). "On the Distribution of the Estimate of Mean Deviation Obtained from Samples from a Normal Population".
Biometrika 33, 254-256.

GREEN, R.F. (1974). "A Note on Outlier Prone Families of Distributions".
Ann.Statist.2, 1293-1295.

GRUBBS, F.E. (1950) "Sample Criteria for Testing Outlying Observations".
Ann.Math.Statist. 21, 27-58.

GRUBBS, F.E. (1969). "Procedures for Detecting Outlying Observations in Samples".
Technometrics 11, 1-21.

GRUBBS, F.E. and BECK, F. (1972). "Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations".
Technometrics 14, 847-854.

GUMBEL, E.J. (1960). "Discussion of the Papers of Messrs Anscombe and Daniel".
Technometrics 2, 157-166.

GUTTMAN, I. (1973). "Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriousity - A Bayesian Approach".
Technometrics 15, 723-738.

GUTTMAN, I. and KHATRI, C.G. (1975).
"A Bayesian Approach to the Detection of Spuriosity".
Applied Statistics. R.P. Guptor (Ed.) 111-145.

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.

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.

HALPERIN, M., GREENHOUSE, S.W., CORNFIELD, J., and ZALOKER, J., (1955)
"Tables of Percentage Points for the Studentised Maximum Absolute Deviate in Normal Samples".
Jour.Amer.Statist.Assoc. 50, 185-195.

HAWKINS, D.M. (1971). "On the Bounds of the Range of Order Statistics".
Jour.Amer.Statist.Assoc. 66, 644-645.

HAWKINS, D.M. (1973). "Repeated Testing for Outliers".
Statistica Neerlandica 27, 1-9.

HAWKINS, D.M. (1974)."The Detection of Errors in Multivariate Data Using Principal Components".
Jour.Amer.Statist.Ass. 69, 340-344.

HEALY, M.J.R. (1968). "Multivariate Normal Plotting".
Appl.Statist. 17, 157-161.

IRWIN, J.O. (1925). "On a Criterion for the Rejection of Oultying Observations".
Biometrika 17, 238-250.

JOHN, J.A. and PRESCOTT, P. (1975). "Critical Values of a Test to Detect Outliers in Factorial Experiments".
Appl.Statist. 24, 56-59.

JOSHI, P.C. (1972). "Efficient Estimation of the Mean of an Exponential Distribution When an Outlier is Present".
Technometrics 14, 137-143.

KABE, D.G.(1970). "Testing Outliers from an Exponential Population".
Technometrics 15, 15-18.

KALE, B.K. (1975). "Trimmed Means and the Method of Maximum Likelihood when Spurious Observations are Present".
Applied Statistics, Edited by Gupta, R.P.
North-Midland Publishing Company.

KALE, B.K. (1975). "Detection of Outliers".
Technical Report No. 63, Dept. Statist.
Univ.of Manitoba. Winnipeg, Canada.

KALE, B.K. and SINHA, S.K. (1971).
"Estimation of Expected Life in the Presence of an Outlier Observation".
Technometrics 13, 755-759.

KARLIN, S., and TRUAX, D. (1960). "Slippage Problems".
Ann.Math.Statist. 31, 296-324.

KENDALL, M.G. and BUCKLAND, W.R. (1957). A Dictionary of Statistical Terms.
Oliver and Boyd. Edinburgh.

KING, E.P. (1953). "On Some Procedures for the Rejection of Suspected Data.
Jour.Amer.Statist.Assoc.48, 531-533.

KRUSKAL, W.H. (1960)."Some Remarks on Wild Observations".
Technometrics 2, 1-3.

KRUSKAL, W.H. (1960a)."Discussion of the Papers of Messrs Anscombe and Daniel".
Technometrics 2, 157-166.

KUDÔ, A. (1956) "On the Testing of Outlying Observations".
Sankhyā 17, 67-76.

KUDÔ, A. (1957). "The Extreme Value in a Multivariate Normal Sample".
Mem.Fac.Sci.Kyushu Univ. (A) 11, 143-156.

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.

LIKEŠ, J. (1966) "Distribution of Dixon's Statistics in the Case of an Exponential Population".
Metrika 11, 46-54.

LUND, R.E. (1975). "Tables for an Approximate Test for Outliers in Linear Regression".
Technometrics, 17, 473-476.

MAY, J.M. (1952). "Extended and Corrected Tables of the Upper Percentage Points of the Studentized Range".
Biometrika 39, 192-193.

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.

MOUNT, K.S. and KALE, B.K. (1973). "On Selecting a Spurious Observation".
Canad.Math.Bulletin, 16, 75-78.

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.

MCMILLAN, R.G. (1971). "Tests for One or Two Outliers in Normal Samples with Unknown Variance".
Technometrics 13, 87-100.

MCMILLAN, R.G. and DAVID, H.A. (1971). "Tests for One or Two Outliers in Normal Samples with Known Variance".
Technometrics 13, 75-85.

NAIR, K.R. (1948a). "The Studentised Form of the Extreme Mean Square Test in the Analysis of Variance".
Biometrika 35, 16-31.

NAIR, K.R. (1948b). "The Distribution of the Extreme Deviate from the Sample Mean and Its Studentized Form".
Biometrika 35, 118-144.

NAIR, K.R. (1952). "Tables of Percentage Points of the 'Studentized' Extreme Deviate from the Sample Mean".
Biometrika 39, 189-195.

NEWMAN, D.(1940). "The Distribution of Ranges in Samples from a Normal Population, Expressed in Terms of an Independent Estimate of the Standard Deviation". Biometrika 31, 20-30.

NEYMAN, J and SCOTT, E.L. (1971). "Outlier Proneness of Phenomena and of Related Distributions". Optimizing Methods in Statistics. Academic Press, New York.

PAULSON, E. (1952). "An Optimum Solution to the k-Sample Slippage Problem for the Normal Distribution". Ann.Math.Statist. 23, 610-616.

PEARSON, E.S., and CHANDRA SEKAR, C. (1936). "The Efficiency of Statistical Tools and a Criterion for the Rejection of Outlying Observations". Biometrika 28, 308-20.

PEARSON, E.S.and HARTLEY,H.O. (1945). "The Probability Integral of the Range in Samples of n Observations from the Normal Population". Biometrika 32, 301-310.

PEARSON, E.S., and HARTLEY, H.O. (1942). "Tables of the Probability Integral of the Studentized Range". Biometrika 33, 89-99.

PEARSON, E.S., and HARTLEY, H.O., (Eds) (1976). Biometrika Tables for Statisticians. Volume 1. Third Edition, 2nd imp.

PEARSON, E.S. and STEPHENS, M.A. ((64). "The Ratio of Range to Standard Deviation in the same Normal Sample". Biometrika 51, 484-487.

PEARSON, K. (1902). "Note on Francis Galton's Problem". Biometrika 1, 390-399.

PEIRCE, B. (1852). "Criterion for the Rejection of Doubtful Observations". Astronomical Journal 2, 161-163.

PILLAI, K.C.S. (1952). "On the Distribution of 'Studentized' Range". Biometrika 39, 194-195.

PILLAI, K.C.S. (1969)."Upper Percentage Points of the Extreme Studentized Deviate from the Sample Mean". Biometrika 46, 473-474.

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.

PRESCOTT, P. (1975). "An Approximate Test for Outliers in Linear Models". Technometrics 17, 129-132.

QUESENBERRY, C.P. and DAVID, H.A. (1961). "Some Tests for Outliers". Biometrika 48, 379-390.

RIDER, P.R. (1933). "Criteria for Rejection of Observations". Washington University Studies, New Series, Science and Technology, No. 8. St.Louis.

ROHLF, F.J. (1975). "Generalisation of the Gap Test for the Detection of Multivariate Outliers". Biometrics 31, 93-101.

ROSNER, B. (1975). "On the Detection of Many Outliers". Technometrics, 17, 221-227.

SHAPIRO S.S., and WILK, M.B. (1965) "An Analysis of Variance Test for Normality (Complete Samples)."
Biometrika 52, 591-611.

SHAPIRO S.S., WILK, M.B. and CHEN, H.J. (1965).
"A Comparative Study of Various Tests for Normality".
Jour.Amer.Statist.Assoc. 60, 1343-1372.

SINHA, S.K. (1972). "Reliability Estimation in Life Testing in the Presence of an Outlier Observation".
Oper.Res. 20, 888-894.

SINHA, S.K. (1973a)."Estimation of the Parameters of a Two-Parameter Exponential Distribution when an Outlier may be Present".
Utilitas Math. 3, 75-82. (with correction, Utilitas Math. 3, 333-354.)

SINHA, S.K. (1973b)."Life Testing and Reliability Estimation for Non-homogeneous Data - A Bayesian Approach".
Comm. in Statist. 2, 235-243.

SINHA, S.K. (1973c)."Distribution of Order Statistics and Estimation of Mean Life When an Outlier may be Present".
Canad.Jour.Statist. 1, 119-171.

SIOTANI, M. (1959). "The Extreme Value of the Generalized Distances of the Individual Points in the Multivariate Normal Sample".
Ann.Inst.Statist.Math., Tokyo 10, 183-208.

SNEDECOR, G.W., and COCHRAN, W.G., (1967). Statistical Methods.
Iowa State University Press. Ames, Iowa.

SRIKANTAN, K.S. (1961)."Testing for the Single Outlier in a Regression Model".
Sankhyā 23, 251-260.

STEFANSKY, W. (1971). "Rejecting Outliers by Maximum Normal Residual".
Ann.Math.Statist. 42, 35-45.

STEFANSKY, W. (1972). "Rejecting Outliers in Factorial Designs".
Technometrics 14, 469-479.

STEVENS, W.L. (1939). "Solution to a Geometrical Problem in Probability".
Ann.Eugen. IX, 315-320.

'STUDENT', (1927). "Errors of Routine Analysis".
Biometrika 19, 151-164.

THOMPSON, B.K. (1973). Personal Communication.

THOMPSON, W.A.Jr., and WILKE, T.A. (1963).
"On an Extreme Rank Sum Test for Outliers".
Biometrika 50, 375-383.

THOMPSON, W.R. (1935)."On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation".
Ann.Math.Statist. 6, 214-219.

THOMSON, G.W. (1955)."Bounds for the Ratio of Range to Standard Deviation".
Biometrika 42, 268-269.

TIAO, G.C., and GUTTMAN, I. (1967). "Analysis of Outliers with Adjusted Residuals".
Technometrics 9, 541-559.

TIETJEN, G.L., and MOORE, R.H. (1972). "Some Grubbs-Type Statistics for the Detection of Several Outliers". Technometrics 16, 583-597.

TIETJEN, G.L., MOORE, R.H. and BECKMAN, R.J. (1973). "Testing for a Single Outlier in Simple Linear Regression". Technometrics 15, 717-721.

TIPPETT, L.H.C. (1925). "On the Extreme Individuals and the Range of Samples Taken from a Normal population". Biometrika 17, 364-387.

TUKEY, J.W. (1949). "One Degree of Freedom for non-additivity". Biometrics 5, 232-42.

TUKEY, J.W. (1960). "Discussion of the Papers of Messrs Anscombe and Daniel". Technometrics 2, 157-166.

VEALE, J.R. (1975). "Improved Estimation of Expected Life when one Identified Spurious Observation may be Present". Jour.Amer.Statist.Assoc. 70, 398-401.

VEALE, J.R., and HUTSBERGER, D.V. (1969). "Estimation of a Mean when one Observation may be Spurious". Technometrics 11, 331-339.

VEALE, J.R. and KALE, B.K. (1972). "Tests of Hypothesis for Expected Life in the Presence of a Spurious Observation". Utilitas Math. 2, 9-23.

WALSH, J.E. (1950). "Some Non-Parametric Tests of Whether the Largest Observations of a Set are too Large or too Small". Ann.Math.Statist. 21, 583-592.

WALSH, J.E. (1958). "Large Sample Non-Parametric Rejection of Outlying Observations". Ann.Inst.Statist.Math. 10, 223-232.

WILKS, S.S. (1962). Mathematical Statistics. John Wiley and Sons, Inc. New York.

WILKS, S.S. (1963). "Multivariate Statistical Outliers". Sankhyā 25, 407-426.