TESTS OF OUTLIERS

By

SHALAL H. AL-JOBOURI

THESIS

SUBMITTED TO THE DEPARTMENT OF STATISTICS

AND THE COMMITTEE

OF GRADUATE STUDIES IN THE COLLEGE OF

ADMINISTRATION AND ECONOMICS

BAGHDAD UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

IN

STATISTICS

JUNE 1976

I certify that this thesis was prepared under my supervision at the University of Baghdad as a partial requirement for the degree of Master of Science in Statistics.

Signature :

Advisor :

Department:

Date :

In view of the available recommendation I forward this thesis for debate by the Examining Committee :

Signature :

Name :

Chairman of the Committee of Graduate Studies in the College of Administration and Economics.

Date :

بسم الله الرحمن الرحيم

P R E F A C E

The problem of screening data has occupied the minds of statisticians for many years. All statistical methods depend on some underlying assumptions which has to characterize the data before the analysis. One of such assumption is the assumption that the data belong to one population and it is free from any deviated values. The statistical methods of screening data for such outside values is called tests of outliers which is the title of our thesis.

The work which we have done can be considered unique in it's subject. We were able to look and review all the procedures of detecting outliers up to the present time. In the meantime we have developed new procedures for testing outliers in the multivariate case.

Chapter I involves all the tests which are developed in the case of the univariate variable. Summary of the test procedures and their uses including the corresponding tables of the critical values.

[illegible]

[illegible]

variable and in simple linear regression lines.

Chapter III is a valuable work and contains a new contribution to the methods of testing outliers in the field of multivariate data. We have noticed in our investigation that very little is done in the multivariate case which inspired us to go further in this case. In this chapter we took data from the agriculture cencesus to demonstrate the use of this technique.

In the last chapter, we thought from the point of view of completeness of the work we have started, we reviewed the procedures of detecting outliers in the analysis of experiments, we showed the importance of those procedures and its impacts on the results of the analysis.

This thesis, for the first time, put together all the work done in the field of testing for outliers. So it can be considered a unique reference book of its kind in this subject. In addition, it contains a new contribution work in the field of multivariate data. Some examples were given in all chapters to explain the method of application.

This work required us to look in all literature and journals up-to-date. It involves knowledge and research in quite a number of fields, such as; statistical methods, distribution and probability theory, estimation, analysis of variance, design of experiments and multivariate methods.

We hope, in our work that we have contributed in going a step forward in serving our people and nation on the way of our immortal message.

LIST OF NOTATIONS

X Is a random variable.

Y Is a random variable.

S^2 Is the sum of squares.

s Is the sample standard deviation.

S Is the variance covariance matrix.

e_i Is the error.

$\underline{X}$ Is a random vector.

T^2 Is a Hotelling T^2 Distribution.

D^2 Is the Mahalanobise distance.

s_v^2 Is an independent mean square to the common variance σ^2.

N.S. Non significance.

ANOVA Is an analysis of variance.

$Y = a + b X$ Is linear regression.

LIST OF TABLES

Page

Page

CHAPTER ONE

OUTLIERS IN ONE VARIATE

1.1 - Introduction -

An outlying observation or "outlier" is the observation that appears to deviate markedly from other observations of the sample in which it occurs. It may be merely an extreme manifestation of the random variability inherent in the data, in this case we will keep it. Or it may be the result of gross deviation from prescribed experimental procedure, or an error in calculation or recording the numerical value. Investigation of the reason and rejection of the observation because it is being from a different population than that of the sample values. If a physical reason was discovered for an outlier, then it is either (i) rejected (ii) corrected or (iii) rejected, and an additional observation is taken.

But if a physical reason is not known then in addition to the (ii) above the observation is rejected and a truncated sample theory is used in the analysis.

A statistical test may always be used to lend support to a judgment that a physical reason does actually exist for an outlier or to initiate action to find a physical cause. There is a number of criteria for testing outliers. in some of these, the doubtful observation is included in the calculation of the numerical value of a sample criterion (or statistic) which is then compared with a critical value based on theory. The critical value is that value of the sample criterion which would be exceeded by chance with some specified (small) probability (α on the basis that all the observations are random sample from a single parent population.

1.2 - Historical Comments -

A survey of statistical literature indicates that the problem of testing the significance of outlying observations received considerable attention prior to 1937. Since this date, however, published literature on the subject seems to have been unusually scant, perhaps because of inherent difficulties in the problem as pointed out by Pearson and Chandra Sekar/1936 [39]. These authors made some important contributions to the problem of outlying observations by bringing clearly in the concept of efficiency of tests which may be used in view of admissible alternative hypotheses.

In 1933, P.R. Rider [36] published a rather comprehensive survey of work on the problem of testing the significance of outlying observations up to that date. The test criteria surveyed by Rider, appears to impose as an initial condition that the standard deviation σ, of the population from which the items were drawn, should be known accurately. In connection with such tests requiring accurate knowledge of σ, we refer to Irwins criteria 1925 [23] which utilize the difference between the first two individuals or the difference between the second and third individuals in random samples for testing the significance of outlying observations.

In 1935, McKay [27] published a note on the distribution of the extreme minus the mean in samples of size n from a normal universe and the distribution of this statistic in samples of n-1 from the same population. McKay gave also an approximate expression for the upper percentage points of the distribution but did not tabulate the exact distribution due to the complicity of the multiple integrals involved.

Under certain circumstances, accurate knowledge concerning σ may be available as, for example in using "daily control" tests of student (William Gosset [17]). The population standard deviation may be estimated in some cases with sufficient precision from past data.

In 1935 also W.R. Thompson [45] apparently had this very point in mind when he devised an exact test in his paper, "On Criterion For The Rejection Of Observations And The Distribution Of The Ratio Of The Deviation To The Sample Standard Deviation". Thompson showed that if

$$t_i = (X_i - \bar{X})/s \qquad \ldots\ldots\ldots\ldots \qquad (1.1)$$

where $\bar{X}$ and s is the sample mean and sample standard deviation, and X_i is an observation selected arbitrarily from a random sample of n items drawn from a normal population, then the probability density function of