MODERN DEVELOPMENTS IN CHI-SQUARE GOODNESS-OF-FIT TESTING

by

Dora Hutchinson

B.Sc. Simon Fraser University 1977

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department

of

Mathematics

SIMON FRASER UNIVERSITY

May 1983

APPROVAL

Name: Dora Hutchinson

Degree: Master of Science

Title of thesis: Modern Developments in Chi-Square Goodness-of-Fit Testing

Examining Committee:

Chairman: B.S. Thomson

R. Lockhart
Senior Supervisor

M.A. Stephens

D.M. Eaves

R. Routledge
External Examiner
Department of Mathematics
Simon Fraser University

Date Approved: April 13, 1983

Title of Thesis/Dissertation:

Modern Developments in Chi-Square Goodness-of-Fit Testing

Author:

(Signature)

Dora Hutchinson

(name)

April 18, 1983

(date)

ABSTRACT

MODERN DEVELOPMENTS IN CHI-SQUARE GOODNESS-OF-FIT TESTING

An examination of the historical development of the Pearson chi-square statistic is presented followed by a review of the recently developed techniques in the field of chi-square goodness-of-fit testing. In particular, a study of the new statistics (the Rao-Robson statistic and the Dzhaparidze-Nikulin statistic) claiming to offer improvement over Pearson's X^2 is provided; Monte Carlo points for these statistics for finite n are determined, power studies are performed and comparisons drawn between competitors, and test procedures are applied to real data to illustrate their usage. Finally, conclusions are drawn as to the success of the modern chi-square methods and a summary of their current applicability given.

ACKNOWLEDGEMENTS

I wish to extend my sincere appreciation and gratitude to my Senior Supervisor, Dr. R. Lockhart, for making the completion of this paper a much easier task. Thanks also go to Dr. M.A. Stephens for his many helpful suggestions.

My acknowledgements would not be complete without recognizing my sister Jan for her invaluable contributions to the typing and organization of this work.

TABLE OF CONTENTS

LIST OF TABLES

I. INTRODUCTION

Since it was first proposed in 1900, Karl Pearson's chi-square statistic has become one of the most popular techniques in goodness-of-fit testing today. However, though there are situations where the chi-square statistic is ideal and irreplaceable, its flexibility and ease of computation are often not sufficient to warrant its use, and the trend is towards the use of more powerful goodness-of-fit statistics. Pearson's statistic, denoted by X^2, has been unable to compete with these statistics in many situations due to the fact that it does not use the total information given by the individual data points; it considers only the number of observations falling within specified cells and consequently lacks power.

Recently there have been attempts made to overcome some of the difficulties historically associated with Pearson's statistic which have increased chi-square's stature in goodness-of-fit testing. Some questions that have been of interest in the past such as "How can degrees of freedom lost due to parameter estimation be recovered?" and "How should the unknown parameters be estimated?" are being re-evaluated in the light of new statistical approaches.

This paper sets out to examine these new techniques with the intention of determining how they have improved on Pearson's chi-square statistic in goodness-of-fit testing and to ascertain how good these improvements are in terms of increased power and/or applicability.

Part II deals with the historical development of the chi-square statistic since 1900, examining the theory behind it, the problems encountered, and the existing solutions to these problems. A review of the old method sufficient to prepare for an examination of the new methods is provided. Despite its long existence, it is still useful.

Part III takes a look at the latest developments and provides an in-depth study of a selection of the more promising techniques.

Part IV presents some practical applications of the new statistics, and Part V provides tabled Monte Carlo points for the new statistics determined for finite n. Part VI examines the comparative power of the new statistics against selected alternatives.

Finally, in Part VII, conclusions are drawn as to the success of the modern chi-square techniques and a summary of their current applicability given.

II. THE HISTORICAL DEVELOPMENT OF THE CHI-SQUARE GOODNESS-OF-FIT STATISTIC

A problem frequently arising in statistics is the goodness-of-fit problem where we wish to test whether or not a random variable X is distributed according to a particular distribution function $F_0(x)$. $F_0(x)$ may be a completely specified function (e.g. $F_0(x)$ is the Normal distribution function with mean μ_0 and variance σ_0^2) or it may be only partially specified (e.g. $F_0(x)$ is in the normal family of distributions.)

The famous statistic, X^2, that Karl Pearson proposed in 1900 is frequently used to tackle this problem. It is a favourite due to its ease of application and the abundant availability of tabled values for the Chi-Square distribution. Pearson's X^2 was developed on the basis of measuring whether or not the observed frequencies of observations on the variable X falling into "cells" were consistent with the "expected" frequencies; that is, the number of observations we would expect to observe given that $F_0(x)$ is indeed the true distribution function. The theory underlying Pearson's X^2 statistic is dependent on whether or not the hypothesized distribution function $F_0(x)$ is completely specified (the Simple Hypothesis Case) or only partially specified (the Composite Hypothesis Case) and requires parameter estimation.

A. The Simple Hypothesis Case

Suppose we are in a goodness-of-fit situation where we wish to test the simple hypothesis H_0: $F(x) = F_0(x)$ where $F_0(x)$ is completely specified. We assume a random sample of n independent observations on the random variable X has been gathered, and that we have arbitrarily divided the range of X into k mutually exclusive cells (the actual selection of these cells will be considered later). Let $\underline{X}'$ denote the row vector of observations, so that

$\underline{X}' = (x_1, x_2, \ldots\ldots, x_n)$. We can now calculate $\underline{N}' = (N_1, N_2, \ldots\ldots, N_k)$ where N_i denotes the number of observations falling into the i^{th} cell, for i=1,...,k. These N_i's constitute a sample of k observations from the Multinomial distribution. If we assume that the null hypothesis is true, we can immediately find the probability p_i that an observation will fall into the i^{th} cell, and subsequently the joint probability that n_1 observations fall into the first cell, n_2 observations fall into the second cell,....., and n_k observations fall into the k^{th} cell.

The multivariate form of the Central Limit Theorem states that $\underline{N}'$ will tend, as n goes to infinity, to have a Normal distribution with mean $\underline{\mu}$, where $\underline{\mu} = (np_1, np_2, \ldots\ldots, np_k)$, and dispersion matrix V, where

$$V = n \times \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & \cdots & -p_1p_{k-1} & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 & \cdots & -p_2p_{k-1} & -p_2p_k \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -p_kp_1 & -p_kp_2 & -p_kp_3 & \cdots & -p_kp_{k-1} & p_k(1-p_k) \end{bmatrix}$$

This matrix V is of dimension k, but since we have the restriction that the sum of the N_i's is equal to n, V is of rank k-1 and hence is not invertible. Since the theory requires that V be invertible, we overcome this problem by deleting a row and column, say the last row and the last column, and will call this new matrix V*.

The inverse of V* is then:

$$V^{*-1} = \frac{1}{n}\begin{bmatrix} \frac{1}{p_1}+\frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \frac{1}{p_k} & \frac{1}{p_2}+\frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{p_k} & \cdots & \frac{1}{p_k} & \frac{1}{p_{k-1}}+\frac{1}{p_k} \end{bmatrix}$$

The quadratic form Q of the exponent of the Multivariate Normal distribution, general form, is then given by:

$$Q = (\underline{N} - \underline{\mu})' V^{*-1} (\underline{N} - \underline{\mu})$$

This can be reduced algebraically as follows:

$$Q = \sum_{i=1}^{k-1} \frac{(N_i - np_i)^2}{np_i} + \frac{1}{np_k} \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} (N_i - np_i)(N_j - np_j)$$

$$= \sum_{i=1}^{k-1} \frac{(N_i - np_i)^2}{np_i} + \frac{1}{np_k} \sum_{i=1}^{k-1} (N_i - np_i) \sum_{j=1}^{k-1} (N_j - np_j)$$

$$= \sum_{i=1}^{k-1} \frac{(N_i - np_i)^2}{np_i} + \frac{1}{np_k} \left(\sum_{i=1}^{k-1} N_i - n \sum_{i=1}^{k-1} p_i\right)^2$$

$$= \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} + \frac{1}{np_k} (n-N_k - n(1-p_k))^2$$

$$= \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} + \frac{1}{np_k} (-N_k + np_k)^2$$

$$= \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

This last expression is, of course, Pearson's chi-square statistic X^2. Hence, X^2 is asymptotically chi-square distributed with k-1 degrees of freedom. This result follows from the theory of the Multivariate Normal distribution which states that the quadratic form in the exponent of this distribution is chi-square distributed with degrees of freedom equal to the rank of V*, in this case k-1.

An alternative proof of the same conclusion that avoids any mathematical complexities (and is a popular item in reviews of chi-square theory) is Fisher's famous proof of 1922. For all its simplicity, it proves to be extremely enlightening for reasons that will later become obvious when the Composite Hypothesis Case is considered.

Suppose we have that $x_1, x_2, \ldots, x_k$ are k independent Poisson distributed variates, the i^{th} variate having parameter np_i associated with it. Then, the probability that $x_1=N_1, x_2=N_2, \ldots, x_k=N_k$ is just

$$P(x_1 = N_1, \ldots, x_k = N_k) = \prod_{i=1}^{k} \frac{e^{-np_i}(np_i)^{N_i}}{N_i!}$$

$$= e^{-n} n^{\Sigma N_i} \prod_{i=1}^{k} \frac{p_i^{N_i}}{N_i!}$$

The sum of the k independent Poisson variates, denoted here by S, is Poisson distributed with parameter $\sum_{i=1}^{k} np_i = n$. Hence, the probability of S being equal to n is:

$$P(S = n) = \frac{e^{-n} n^n}{n!}$$

We can now find the probability that $x_1 = n_1$, $x_2 = n_2, \ldots, x_k = n_k$ conditional on S=n.

$$P\left(x_1 = N_1, \ldots, x_k = N_k \mid S = \sum_{i=1}^{k} N_i = n\right) = \frac{P(x_1 = N_1, \ldots, x_k = N_k)}{P(S=n)}$$

$$= \frac{e^{-n} n^n \prod_{i=1}^{k} \frac{p_i^{N_i}}{N_i!}}{\frac{e^{-n} n^n}{n!}}$$

$$= n! \prod_{i=1}^{k} \frac{p_i^{N_i}}{N_i!}$$

$$= \frac{n!}{n_1!n_2!.....n_k!} \; p_1^{N_1} \; p_2^{N_2} .....p_k^{N_k}$$

This we recognize as the Multinomial distribution. If we define

$$y_i = \frac{N_i - np_i}{(np_i)^{\frac{1}{2}}} \quad \text{for } i = 1,2,.....,k$$

then as n goes to infinity, the distribution of y_i approaches the Normal distribution with mean 0 and variance 1. Since X^2 can be expressed as the sum of these k independent standard normal variates subject to the single linear constraint S=n, we can now establish, by quoting well-known theorems, that in the limit X^2 follows the Chi-Square distribution with k-1 degrees of freedom. This concludes Fisher's proof.

The revealing aspect will be appreciated more fully in the section on parameter estimation, following, where each parameter to be estimated imposes an additional linear constraint.

B. The Composite Hypothesis Case

Suppose now that we are in the situation where we wish to test the composite hypothesis H_0: $F(x) = F_0(x)$ where $F_0(x)$ is a continuous distribution not totally specified but having s of its parameters unknown. The immediate consequence here versus the Simple Hypothesis Case is that the required probabilities p_i, i=1,.....,k, are no longer calculable, at least prior to observing the data. If we denote the s unknown parameters of the distribution function $F_0(x)$ as $\theta_1,\theta_2,.....,\theta_s$, then the p_i's are themselves functions of $\underline{\theta}' = (\theta_1,\theta_2,.....,\theta_s)$. To emphasize this relationship, the

unknown probabilities will be denoted by $p_i(\underline{\theta})$. To calculate X^2 in this situation, now of the form

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - np_i(\underline{\theta}))^2}{np_i(\underline{\theta})}$$

we will require estimates of the unknown parameters, These estimates will here be denoted collectively by $\underline{\hat{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots\ldots, \hat{\theta}_s)$.

This presents a new distribution problem, for it is not obvious that the asymptotic distribution of this statistic will be of the same form as in the Simple Hypothesis Case. Pearson himself did not believe that the estimation of unknown parameters using the sample data would significantly alter the distribution of his statistic. He believed that regarding X^2 as chi-square distributed with k-1 degrees of freedom when unknown parameters were estimated would cause only negligible error in the approximation and would not, therefore, affect practical decisions. His conclusion was perhaps justified for some applications, but his statistic performed so poorly in some of the most common tests employing X^2 that a "degrees-of-freedom battle" ensued. It was not until 1924 with the appearance of Fisher's famous paper on the subject that the battle was resolved.

The Estimation of Parameters

If we now reconsider Fisher's proof in the Simple Hypothesis Case, we can regard the estimation of unknown parameters as simply the imposition of a further s linear constraints. This clearly has the effect of reducing the degrees of freedom from k-1 to k-s-1. However, this conclusion turns out to be further dependent on the method of estimation used.

The common estimators will be considered here, namely the Maximum Multinomial Likelihood estimators and the Maximum Density Likelihood estimators. We will discover that if the Maximum Multinomial Likelihood (MML) estimates are employed (calculated by maximizing the joint density of the N_i's) then a further s linear constraints are imposed as suspected, and the degrees of freedom are reduced to k-s-1 accordingly. These estimates, however, are rarely ever used because of the difficulties associated with their computation.

By far the most popular method of estimation is that based on maximizing the joint density of the x_i's. This method yields the Maximum Density Likelihood (MDL) estimators, and in this situation, the reduction in the degrees of freedom is considerably more complicated. We will discover through the following analysis that the degrees of freedom are bounded by k-1 and k-s-1, but beyond that we can draw no further conclusions. If k is large, then this difference is negligible, but for small k, errors due to the difference between the critical points for these two distributions will be significant.

A further investigation into the conditions which lead to a reduction in the degrees of freedom is provided by Watson in his paper of 1959 where he gives a general approach to the problem. A good review of the theory as applied to Pearson's X^2 statistic is presented by Kendall and Stuart (1963), pages 425-430.

Kendall and Stuart show that if we have estimators with variances and covariances of order n^{-2}, then X^2 is asymptotically chi-square distributed with k-1 degrees of freedom. However, this is not the usual case. Generally, we have that $\hat{\theta}-\theta = O(n^{-\frac{1}{2}})$, and it is in this situation that they present their theory.

In the Simple Hypothesis Case, with $\underline{y}$ defined as:

$$\underline{y} = (y_1, y_2, \ldots\ldots, y_k)$$

$$\text{where } y_i = \frac{N_i - np_i(\underline{\theta})}{(np_i(\underline{\theta}))^{\frac{1}{2}}}$$

they show that the dispersion matrix $V(\underline{y})$ has a trace equal to k-1 so that X^2 is chi-square distributed with k-1 degrees of freedom in the limit.

In the Composite Hypothesis Case, the theory is more complicated, and the results are given below according to the particular estimates being considered.

a. The Maximum Multinomial Likelihood Estimators

The Multinomial Likelihood, L, is:

$$L = \frac{n!}{N_1!N_2! \cdots\cdots N_k!} \left(p_1^{N_1} \cdot p_2^{N_2} \cdots\cdots p_k^{N_k}\right)$$

To maximize L with respect to the unknown parameters $\underline{\theta}$, the log likelihood is computed and then minimized by setting the first partial derivatives to zero. This gives:

$$\frac{\partial \log L}{\partial \theta_j} = \sum_{i=1}^{k} N_i \frac{\partial p_i}{\partial \theta_j} \frac{1}{p_i} = 0 \quad \text{for } j = 1, 2, \ldots\ldots, s$$

The roots of these s equations will provide the MML estimators. Note that each equation is a homogeneous linear relationship with respect to the N_i's so that together they impose s additional linear constraints within Fisher's proof presented earlier. Hence, X^2 is asymptotically chi-square distributed with k-s-1 degrees of freedom.

Kendall and Stuart show that if these estimates are used, then the dispersion matrix of $\underline{y}$ (defined on page 11), $V(\underline{y})$, has a trace equal to k-s-1 implying that X^2 is chi-square distributed in the limit with k-s-1 degrees of freedom.

The Maximum Multinomial Likelihood estimates determined from the grouped data are difficult to obtain. For this reason, when the individual data points are available, they are not used in practice. Rather, the Maximum Density Likelihood estimates obtained from the ungrouped data are utilized. The use of these MDL estimates also implies more efficient use of the information provided by the data given that the ungrouped observations are available.

b. The Maximum Density Likelihood Estimators

The Density Likelihood, L_D, is $L_D = f(x_1)\cdot\cdot\cdot\cdot\cdot f(x_n)$ where x_i, $i=1,.....,n$ are independent observations on the random variable X with probability density function f. The Maximum Density Likelihood estimators are obtained by maximizing L_D and are given by the roots of the equations:

$$\frac{\partial \log L_D}{\partial \theta_i} = 0 \quad \text{for } i = 1, 2,.....,s \ .$$

Kendall and Stuart show that if these estimates are used, then the dispersion matrix of $\underline{y}$, $V(\underline{y})$, has a trace bounded by k-s-1 and k-1 so that in the limit, the distribution of X^2 is not chi-square, but is bounded between a Chi-Square distribution with k-s-1 degrees of freedom and a Chi-Square distribution with k-1 degrees of freedom. Hence, for large k, the effect of using the Chi-Square distribution with k-s-1 degrees of freedom in goodness-of-fit testing will not lead to serious error. However, an examination of a table of Chi-Square distribution values will show that this is not the case

for small k, and this fact should be kept in mind for practical applications.

Application of X^2 to Goodness-of-Fit Testing

In order to use X^2 in hypothesis testing, cells into which the observations can be grouped must be selected. There are two immediate problems posed by this task. The first is in terms of the subjectivity involved; the foregoing theory applies independently of the choice of cells actually made. This subjectivity has been an area of criticism for the chi-square statistic because a variety of results can be obtained from the same data.

Secondly, the theory applies however the cells are chosen as long as they are selected without reference to the data. In practice, of course, the observations are often reviewed to actually determine the class boundaries. No consideration is given in the preceding theory to the case where class boundaries are themselves random variables. It is therefore pertinent to ask how the theory is affected when the classes are determined in this manner. Fortunately, in practice, the limiting distribution of X^2 with random cells is exactly the same as if the fixed cells had been used (see, for example, Moore (1975)).

We look first at the selection of the cell boundaries, assuming that the number of cells, k, has been fixed. In situations where natural groupings exist or in the case of a discrete distribution, this problem may take care of itself. However, if fewer cells than are provided "naturally" are desirable, or in particular, if the distribution of interest is continuous, then it is beneficial to choose cell boundaries that are in some sense optimal; specifically, optimal in terms of maximizing the power of the test. Unfortunately, this problem has not been studied systematically.

The current and generally accepted method appears to be to choose cells

that are equally probable. The benefits of such a choice as given by Moore (N.D.).

(i) The distance sup $| F_1(x) - F(x) |$ to the furthest alternative F_1 indistinguishable from F by X^2 is minimized (this property appears to be misstated in Moore (N.D.)).

(ii) The chi-square test is unbiased. (Mann and Wald (1942) proved only local unbiasedness, but the test is in fact unbiased against arbitrary alternatives F_1.)

(iii) Empirical studies have shown that the Chi-Square distribution is a more accurate approximation to the exact distribution of X^2 when equiprobable cells are employed.

Selection in this way does, however, require that tables be available to provide the necessary values. It also implies that the data must be available ungrouped. Given that either of these requirements is not met, it has been suggested that cells be chosen as equal intervals on the range of the random variable with hypothesized distribution $F_0(x)$ except in the tails which are allowed to go to infinity (Kendall and Stuart, 1963, page 431).

With the problem of selecting class boundaries removed (if not in the most satisfactory manner) we come to the task of selecting k. Studies have been conducted to determine an "optimal" k based on one of two criteria:

(i) Maximization of the power of the test, or

(ii) Attainment of a better approximation of the Chi-Square distribution to the distribution of X^2.

Criterion (i) was the motivation for studies performed by Mann and Wald who, through a sophisticated and rigorous argument not detailed here, arrived at the conclusion that k should be chosen according to

$$k = 4 \left[\frac{4n^2}{z(\alpha)^2} \right]^{1/5}$$

where n = sample size, α = significance level of the test, and $z(\alpha)$ = the point of the standard Normal distribution that places α-probability in the tail.

This formula is generally criticized for providing values of k that are much too large in the sense of criterion (i). It is also criticized on the grounds that the results are accurate only for n greater than or equal to 300.

Dahiya and Gurland (1973) performed a thorough study based on the first criterion utilizing the chi-square test with data-dependent cells (i.e. cells chosen according to the data). They concluded that optimal k was heavily dependent on the alternate hypothesized distribution. For instance, in testing for normality against the alternative Logistic distribution, k = 3 was the optimal choice. For alternatives other than those of or related to the normal family, k of moderate size (7, 12, etc.) proved best.

In terms of the criterion (ii), at least two studies have been performed -- one by Roscoe and Byars (1971) and one by Good, Gover, and Mitchell (1970). Moore (1975) states the results of the Roscoe and Byars study, noting the fact that current recommendations are in terms of the <u>average</u> expected cell frequencies as opposed to Cochran (1954) who gave the commonly accepted rule of thumb in terms of minimum expected frequencies.

Here are the findings:

(i) With equiprobable cells, the average expected cell frequency should be at least 1 (that is, k less than or equal to n) when testing fit at the $\alpha = 0.05$ level; for $\alpha = 0.01$, the average expected frequency should be at least 2 (that is, 2k less than or equal to n).

(ii) When cells are not approximately equiprobable, these average expected frequencies should be doubled.

(iii) These recommendations apply when k is greater than or equal to 3. For k=2 (1 degree of freedom), the chi-square test should be replaced by the test based on the Binomial distribution.

The formula credited to Mann and Wald falls within these guidelines, but it appears to give tests of lower than optimal power. Moore, however, justified his use of the Mann-Wald estimation in his work due to his experiencing greater sensitivity with it than with the Dahiya-Gurland calculations.

III. MODERN METHODS

This paper was largely motivated by an article written by David Moore (N.D.), and this section takes its cue from his publications. In this section that examines the latest developments in the field of chi-square goodness-of-fit testing, you will find a review of those areas which Moore has indicated are worth investigating and omission or brief mention only of those items which he intimates are a dead end. In a sense, and particularly in the remaining two chapters, this paper is an extension of his studies. Otherwise, it may simply reiterate conclusions he has reached without further investigation due either to the apparent futility in the face of better alternatives or to the irrelevancy of the material to this paper.

This chapter is divided into three categories. The first encompasses actual new statistics that have been developed to improve on Pearson's X^2 while attempting to retain its advantages. Following Moore's example (Moore, 1976), these are referred to as "standard" statistics or those statistics whose large-sample theory resembles that of X^2.

The second category includes other "nonstandard" statistics and techniques that are related to the subject matter but are not considered in detail here; these items introduce ideas for potential further study.

The third area looks at specially designed chi-square tests; that is chi-square goodness-of-fit tests based on the most promising new statistic adapted to specific cases. A test of fit for the Multivariate Normal distribution and a test of fit for data that are censored in a particular manner are considered.

Following is an outline of the material taken from the broad range of new chi-square techniques that is included in this chapter.

1. Standard Statistics

   i. The Rao-Robson Statistic
   ii. The Dzhaparidze-Nikulin Statistic

2. Other Techniques

   i. The Kempthorne Statistic
   ii. The Dahiya-Gurland Statistic
   iii. The Effect of Dependence on Chi-Square Tests of Fit

3. Special Applications

   i. A Chi-Square Test for Type II Censored Data
   ii. A Chi-Square Test for Multivariate Normality

1. Standard Statistics

As previously indicated, the chi-square statistics considered here are categorized as "standard" due to their similarities to Pearson's statistic X^2 in terms of their large sample theory. These statistics involve quadratic forms in the standardized cell frequencies $\frac{N_i - np_i}{(np_i)^{\frac{1}{2}}}$ other than the sum of squares used by Pearson. There is a general approach to the construction of such statistics, called "Wald's Method", a good review of which is provided by Moore (1977). To summarize briefly, let $\underline{\theta}$ denote the vector of parameters $\theta_1, \theta_2, \ldots, \theta_s$ and $V(\underline{\theta})$ the vector of standardized frequencies with i^{th} entry:

$$\frac{N_i - np_i(\underline{\theta})}{(np_i(\underline{\theta}))^{\frac{1}{2}}} \qquad \text{for } i=1, 2, \ldots, k$$

Let Q denote a kxk symmetric, nonnegative definite matrix, possibly data-dependent. The generalized form of the Wald's Method statistic W is then:

$$W = V(\underline{\theta})'QV(\underline{\theta})$$

For the particular choice of $Q = I$, where I denotes that kxk identity matrix, W is Pearson's X^2 statistic. For some alternate choices, we arrive at the statistics detailed below.

Statistics of the form discussed in Wald's Method are, in the limit, a linear combination of independent chi-square random variables. For the calculation of their distributions, refer, for example, to Davis (1977).

Given a generalized method for the construction of chi-square statistics, the advantages of X^2 that should be retained by a new quadratic form are of interest. The criteria that a goodness-of-fit chi-square test statistic should ideally satisfy in order to achieve a worthwhile degree of competitiveness are summarized below:

a) The observed value of the statistic should be easily calculable. The main determinant of chi-square statistics' popularity is ease of use. The widespread availability of computers has aided considerably in this respect as the iterative solutions to nonlinear equations and the evaluation of quadratic forms are simplified by computer library routines.

b) The limiting null distribution should be chi-square. This factor enables immediate access to critical points eliminating the need for the construction of special tables for each newly hypothesized distribution.

i) The Rao-Robson Statistic

The item that is deemed most worthy of further investigation and that which receives the most in-depth study here, particularly in terms of application, is the statistic credited to Rao and Robson.

Returning now to Wald's Method, define Q as:

$$Q = (I - BJ^{-1}B')^{-1}$$

where I is again the kxk identity matrix, J the sxs information matrix of the distribution function F(x), and B the kxs matrix with ij^{th} entry

$$p_i^{-\frac{1}{2}} \frac{\partial p_i}{\partial \theta_j}$$

For this choice of Q, W = V'QV is the Rao-Robson statistic. K.C. Rao and D.S. Robson (1975), however, obtained their statistic by an alternate approach which will not be detailed here.

Rao and Robson set out to overcome the problem encountered by Pearson's statistic in the case where the more efficient maximum density likelihood (MDL) estimators are used in its calculation. To reiterate earlier findings, under this condition and for the case where class boundaries have been predetermined, the statistic X^2 is asymptotically distributed as a linear combination of chi-square variables; that is, in the limit, as $n \to \infty$:

$$X^2 = y_1^2 + y_2^2 + \ldots\ldots + y_{k-s-1}^2 + \lambda_1 y_{k-s}^2 + \ldots\ldots + \lambda_s y_{k-1}^2$$

where the y_i's are independent standard normal variables, and the λ_i's are restricted to the unit interval such that $0 \leq \lambda_i < 1$ for i=1, 2,.....,s and may depend on the unknown parameters $\theta_1, \ldots\ldots, \theta_s$. In the more realistic case where the class boundaries are themselves functions of $\underline{\theta}$, Watson (1958) proves that if the parameters involved are those of location and scale, the asymptotic distribution as given above is independent of parameters.

Rao and Robson argued that the asymptotic dependence on both the para-

meters and the functional form of $f(x;\underline{\theta})$ can be eliminated by adding a correction term, denoted by Y^2, which converges in law to:

$$(1 - \lambda_1)y_{k-s}^2 + \ldots\ldots + (1 - \lambda_s)y_{k-1}^2$$

This would enable total recovery of the s degrees of freedom lost by X^2 when the parameter estimates are based on the grouped data, as opposed to only partial recovery resulting from the use of the MDL estimates.

The Rao-Robson statistic, denoted by RR, is then:

$$RR = X^2 + Y^2$$

where, as usual, X^2 denotes Pearson's statistic, and the form of Y^2 remains to be determined. Rao and Robson (1975) present their derivation of Y and the development of their statistic under the assumption that the null distribution is a member of the Exponential family. To arrive at the same statistic via Wald's method, this particular assumption is not required. In either case, the statistic is defined as follows:

$$\text{Let } u_{ij} = \int_{c_{i-1}}^{c_i} \frac{\partial f_1}{\partial \theta_j}\, dx_1 \quad \text{for } i = 1,2,\ldots\ldots,k-1;\ j = 1,2,\ldots\ldots,k-1$$

where f_1 denotes the probability density function of X_1;

let U represent the kxs matrix with ij^{th} entry u_{ij} and U' its transpose;

let T represent the (k-1)x(k-1) matrix with ij^{th} entry:

$$\begin{cases} p_i(1-p_i) & \text{for } i=j \\ -p_i p_j & \text{for } i \neq j \end{cases} \qquad i,j=1,2,\ldots\ldots,k-1$$

let $\underline{N}' = (N_1,\ldots\ldots, N_{k-1})$ denote the vector of cell counts;

let $\underline{p}' = (p_1,\ldots\ldots, p_{k-1})$ denote the vector of cell probabilities.

For large n, $\frac{N}{n^{\frac{1}{2}}}$ has mean $n\underline{p}$ and covariance matrix $T - UVU'$.

Consider the statistic defined by:

$$RR = \frac{1}{n}(\underline{N} - n\underline{p})'(T - UVU')^{-1}(\underline{N} - n\underline{p})$$

If we denote by a_{jk} the jk^{th} entry of the matrix A defined by:

$$A = (V^{-1} - J)^{-1}$$

where J has jk^{th} entry $\sum_{i=1}^{k} \frac{1}{p_i} u_{ij}u_{ik}$, then RR reduces to:

$$RR = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} + \frac{1}{n} \sum_{j,k}^{s} \{\sum_{i}^{k}(\frac{N_i - np_i}{p_i})u_{ij}\}\{\sum_{i}^{k}(\frac{N_i - np_i}{p_i})u_{ik}\} a_{jk}$$

with $\underline{\theta}$ replaced by the MDL estimates. The first term we recognize as Pearson's X^2; the second term is Y^2. Provided that the asymptotic conditional distribution of $\frac{\underline{N} - n\underline{p}}{n^{\frac{1}{2}}}$ given $\hat{\underline{\theta}}$ is Multivariate Normal with mean zero and covariance matrix $T - U\hat{V}U'$, then the statistic RR is asymptotically chi-square distributed with k-1 degrees of freedom (Rao & Robson, 1975). Rao and Robson do not provide the complete supporting theory within their paper; for theoretical details refer instead to Davis (1977).

The specific form of RR will be derived for the Normal and other distributions.

a) The Normal Distribution

Suppose we wish to test that a random sample of n observations, $x_1,\ldots,x_n$, was taken from a normal population with unknown mean and variance. The probability density function is:

$$f(x,\underline{\theta}) = \frac{1}{(2\pi\sigma)^{\frac{1}{2}}} \exp\{\frac{-(x-u)^2}{2\sigma^2}\} \qquad \begin{matrix} -\infty < \mu < \infty \\ \sigma^2 > 0 \end{matrix}$$

The covariance matrix is:

$$\frac{V}{n} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

and $\hat{V} = V$ with σ^2 replaced by the estimate $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$.

Let $\bar{x}$ denote the MDL estimate of μ. Then the natural choice of class intervals is $(\bar{x} + z_{i-1}s, \bar{x} + z_i s)$, for $i=1,2,\ldots,k$, where the z_i's are chosen so that $p_i = \frac{1}{k}$ for every i. In particular, $z_o = -\infty$ and $z_k = \infty$. The values of $z_1,\ldots,z_{k-1}$ may be determined from a standard normal table.

Accordingly, define the boundaries $c_o,\ldots,c_k$ as follows:

$$c_o = -\infty$$

$$c_i = \bar{x} + z_i s \qquad i=1,\ldots,k-1$$

$$c_k = \infty$$

Let f_1 denote the probability density function of the random variable X_1. Then,

$$u_{i1} = \int_{c_{i-1}}^{c_i} \frac{\partial f_1}{\partial \bar{x}} \, dx_1 \qquad i=1,\ldots,k$$

$$= \frac{1}{(2\pi s^2)} \left(\exp \frac{-z_{i-1}^2}{2} - \exp \frac{-z_i^2}{2}\right)$$

and, $$u_{i2} = \int_{c_{i-1}}^{c_i} \frac{\partial f_1}{\partial s^2}\, dx_1 \qquad i=1,\ldots,k$$

$$= \frac{1}{2s^2(2\pi)^{\frac{1}{2}}}\left(z_{i-1}\exp\frac{-z_{i-1}^2}{2} - z_i \exp\frac{-z_i^2}{2}\right)$$

To obtain the a_{jk}'s, j,k=1,2:

$$A = (\hat{V}^{-1} - J)^{-1}$$

$$= \begin{bmatrix} \frac{1}{s^2} - \sum_{i=1}^{k} \frac{1}{p_i} u_{i1}^2 & -\sum_{i=1}^{k} \frac{1}{p_i} u_{i1}u_{i2} \\ -\sum_{i=1}^{k} \frac{1}{p_i} u_{i2}u_{i1} & \frac{1}{2s^4} - \sum_{i=1}^{k} \frac{1}{p_i} u_{i2}^2 \end{bmatrix}^{-1}$$

Using the fact that $\frac{1}{p_i} = k$ for every i, and defining $v_{i1} = su_{i1}$ and $v_{i2} = s^2u_{i2}$ for simplicity, and consistency with the literature,

$$A = \frac{2s^6}{D}\begin{bmatrix} \frac{1 - 2k\,\Sigma\, v_{i2}^2}{2s^4} & \frac{k\,\Sigma\, v_{i1}v_{i2}}{s^3} \\ \frac{k\,\Sigma\, v_{i2}v_{i1}}{s^3} & \frac{1 - k\,\Sigma\, v_{i1}^2}{s^2} \end{bmatrix}$$

where $D = (1 - 2k\,\Sigma\, v_{i2}^2)(1 - k\,\Sigma\, v_{i1}^2) - 2k^2(\Sigma\, v_{i1}v_{i2})^2$

Hence, $$a_{11} = s^2(1 - 2k\,\Sigma\, v_{i2}^2)/D$$

$$a_{12} = a_{21} = 2s^3 k\,\Sigma\, v_{i1}v_{i2}/D$$

$$a_{22} = 2s^4(1 - k\,\Sigma\, v_{i1}^2)/D$$

If we redefine $a_{11} = \frac{a_{11}}{s^2}$, $a_{12} = a_{21} = \frac{a_{12}}{s^3}$, and $a_{22} = \frac{a_{22}}{s^4}$, the 's' terms

will cancel. Substitution into the general form of RR and subsequent simplification yields the Rao-Robson statistic for testing of the Normal distribution:

$$RR = \frac{k}{n}\sum (N_i - \frac{n}{k})^2 + \frac{k^2}{n}\{\sum(N_i - \frac{n}{k})v_{i1}\}^2 a_{11}$$

$$+ \frac{2k^2}{n}\{(\sum (N_i - \frac{n}{k})v_{i1})(\sum (N_i - \frac{n}{k})v_{i2})\}a_{12}$$

$$+ \frac{k^2}{n}\{\sum(N_i - \frac{n}{k})v_{i2}\}^2 a_{22}$$

This formulation agrees in every respect to the derivation given by Rao and Robson (1975) except with respect to the definition of a_{22} which in their case is:

$$a_{22} = s^4(1 - k \sum v_{i1}^2)/D$$

$$\text{versus } a_{22} = 2s^4(1 - k \sum v_{i1}^2)/D$$

A copy of the Fortran program for the calculation of this statistic is included in the Appendix.

Rao and Robson performed simulation studies of the power functions of three statistics, namely Pearson's X^2 with MML estimates, X^2 with the MDL estimates denoted by $\tilde{R}$, and the Rao-Robson statistic based on a test for the Normal distribution. Their results indicate that RR is the most powerful against alternatives including Double Exponential, mixtures of Double Exponential and Normal variates, and mixtures of Normal variates with equal means and unequal variances. Equiprobable cells were employed in all cases. Since it appears that the form of the Rao-Robson statistic derived by Rao and Robson for testing of the Normal distribution employed the definition of $\hat{V}$ following:

$$\frac{\hat{V}}{n} = \begin{bmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{\hat{\sigma}^4}{n} \end{bmatrix}$$

when in fact

$$\frac{\hat{V}}{n} = \begin{bmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix}$$

these simulations are somewhat questionable; this error results in a discrepancy in a_{22} such that a_{22} is one-half of its true value. It is not obvious what impact this would have on the power simulations. However, given a critical point in the upper tail of the Chi-Square distribution as a basis for rejecting the null hypothesis, RR may appear less powerful than it really is if the magnitude of a_{22} is of any significance. Work by the writer indicates the discrepancy does not result in serious error.

b) The Exponential Distribution

If we wish to test that a random sample is exponentially distributed, with θ unknown, then:

$$f(x;\theta) = \frac{1}{\theta} \exp \frac{-x}{\theta} \qquad x \geq 0,\ \theta > 0$$

denotes the probability density function. The variance of θ is θ^2.

Let $\bar{x}$ denote the MDL estimate of θ so that $\hat{V} = \bar{x}^2$. The natural choice of class intervals is $(\bar{x}z_{i-1}, \bar{x}z_i)$, for $i=1,\ldots,k$, with $z_o=0$ and $z_k=\infty$. Values of

$z_1,\ldots,z_{k-1}$ are chosen so that $p_i = \frac{1}{k}$ for every i. This is achieved by taking:

$$z_i = -\log(1 - \frac{i}{k}) \qquad i=1,\ldots,k-1$$

Let $c_i = \bar{x}z_i$, then

$$v_i = \int_{c_{i-1}}^{c_i} \frac{df_1}{d\bar{x}}\, dx_1$$

$$= \frac{1}{\bar{x}} \{z_{i-1}\exp(-z_{i-1}) - z_i\exp(-z_i)\}$$

To obtain a_{11}, and defining $v_i = \bar{x}u_i$,

$$A = a_{11} = (v^{-1} - J)^{-1}$$

$$= (\frac{1}{\bar{x}^2} - \frac{k}{\bar{x}^2} \Sigma\, v_i^2)^{-1}$$

$$= \frac{\bar{x}^2}{1 - k\, \Sigma\, v_i^2}$$

Substitution into the general formula and minor simplification yields the Rao-Robson test statistic for the Exponential distribution:

$$RR = \frac{k}{n} \Sigma\, (N_i - \frac{n}{k})^2 + \frac{k^2}{n} \frac{\{ \Sigma\, (N_i - \frac{n}{k})v_i \}^2}{(1 - k\, \Sigma\, v_i^2)}$$

A copy of the Fortran program for the calculation of this statistic is contained in the Appendix.

c) The Logistic Distribution

Suppose we wish to test that a random sample came from the Logistic

distribution. Then,

$$f(x;\theta) = \frac{\exp y}{\beta(1 + \exp y)^2} \qquad \beta > 0$$

where $\theta=(\alpha,\beta)$ and $y = \frac{-(x - \alpha)}{\beta}$. The covariance matrix V is defined as:

$$V = \begin{bmatrix} 3\beta^2 & 0 \\ 0 & \frac{9\beta^2}{3 + \pi^2} \end{bmatrix}$$

so that

$$\hat{V}^{-1} = \begin{bmatrix} \frac{1}{3\hat{\beta}^2} & 0 \\ 0 & \frac{3 + \pi^2}{9\hat{\beta}^2} \end{bmatrix}$$

where $\hat{\beta}$ denotes the MDL estimate for β; $\hat{\alpha}$ will denote the MDL estimate for α.

Transformation of the data into standard form implies that, for equiprobable cells, the cell boundaries $c_0,\ldots,c_k$ are defined as follows:

$$c_o = -\infty$$

$$c_i = -\log\left(\frac{i}{k} - 1\right) \qquad i=1,\ldots,k-1$$

$$c_k = \infty$$

Then,

$$u_{i1} = \int_{c_{i-1}}^{c_i} \frac{\partial f_1}{\partial \alpha} \, dx_1 \qquad i=1,\ldots,k$$

$$= \frac{1}{\beta}\left\{\frac{\exp c_{i-1}}{(1+\exp c_{i-1})^2} - \frac{\exp c_i}{(1+\exp c_i)^2}\right\}$$

and

$$u_{i2} = \int_{c_{i-1}}^{c_i} \frac{\partial f_1}{\partial \beta}\, dx_1 \qquad i=1,\ldots,k$$

$$= \frac{1}{\beta}\left\{\frac{c_i \exp c_i}{(1+\exp c_i)^2} + \frac{c_{i-1}\exp c_{i-1}}{(1+\exp c_{i-1})^2}\right\}$$

The required a_{jk}'s, j,k=1,2, are the entries of the matrix:

$$A = (\hat{V}^{-1} - J)^{-1}$$

Using the substitution $3 + \pi^2 = c_1$,

$$A = \frac{1}{D}\begin{bmatrix} c_1\hat{\beta}^2 - k\,\Sigma\, u_{i2}^2 & k\,\Sigma\, u_{i1}u_{i2} \\ k\,\Sigma\, u_{i1}u_{i2} & 3\hat{\beta}^2 - k\,\Sigma\, u_{i1}^2 \end{bmatrix}$$

where $D = (3\hat{\beta}^2 - k\,\Sigma\, u_{i1}^2)(c_1\hat{\beta}^2 - k\,\Sigma\, u_{i2}^2) - k^2(\Sigma\, u_{i1}u_{i2})^2$.

The Rao-Robson statistic for testing goodness-of-fit to the Logistic distribution is then:

$$RR = \frac{k}{n}\Sigma\,(N_i - \frac{n}{k})^2 + \frac{k^2}{n}\sum_{j,k}\{\Sigma_i\,(N_i - \frac{n}{k})u_{ij}\}\{\Sigma_i\,(N_i - \frac{n}{k})u_{ik}\}a_{jk}$$

A copy of the Fortran program for calculating this statistic may be found in the Appendix. Note that the parameter estimates for this distribution must be found through numerical iteration and subroutines (provided by Dr. M.A. Stephens) are included for this purpose.

d) The Extreme Value Distribution

If we wish to test for fit of data to the Extreme Value Distribution, then:

$$f(x;\theta) = \frac{1}{\beta} \exp(y - \exp y) \qquad \beta > 0$$

where $\theta=(\alpha,\beta)$ and $y = \frac{-(x - \alpha)}{\beta}$.

The covariance matrix V is:

$$V = \frac{1}{\beta^2} \begin{bmatrix} 1 + \frac{6}{\pi^2}(1-\gamma)^2 & \frac{6(1-\gamma)}{\pi^2} \\ \frac{6(1-\gamma)}{\pi^2} & \frac{6}{\pi^2} \end{bmatrix}$$

Taking the inverse gives:

$$V^{-1} = \beta^2 \begin{bmatrix} 1 & -(1-\gamma) \\ -(1-\gamma) & \frac{\pi^2}{6} + (1-\gamma)^2 \end{bmatrix}$$

Transformation of the data into standard form implies that, for equiprobable cells, the cell boundaries $c_0,\ldots,c_k$ may be defined as:

$$c_0 = -\infty$$
$$c_i = -\log\{-\log(\tfrac{i}{k})\} \quad i=1,2,\ldots,k-1$$
$$c_k = \infty$$

Then,

$$u_{i1} = \int_{c_{i-1}}^{c_i} \frac{\partial f_1}{\partial \alpha} \, dx_1 \qquad \text{for } i=1,\ldots,k$$

$$= \frac{1}{\beta} \{\exp(c_{i-1} - \exp c_{i-1}) - \exp(c_i - \exp c_i)\}$$

and

$$u_{i2} = \int_{c_{i-1}}^{c_i} \frac{\partial f_1}{\partial \beta} \, dx_1 \qquad \text{for } i=1,\ldots,k$$

$$= \frac{1}{\beta} \{c_i \exp(c_i - \exp c_i) - c_{i-1}\exp(c_{i-1} - \exp c_{i-1})\}$$

The required a_{jk}'s are the entries of the matrix A defined by:

$$A = (\hat{V}^{-1} - J)^{-1}$$

Substitution of $c_1=(1-\gamma)$ and $c_2=\frac{\pi^2}{6} + (1-\gamma)^2$ gives:

$$A = \frac{1}{D} \begin{bmatrix} c_2 - k\Sigma d_{i2}^2 & c_2 + k\Sigma d_{i1}d_{i2} \\ c_1 + k\Sigma d_{i1}d_{i2} & 1 - k\Sigma d_{i1}^2 \end{bmatrix}$$

where $D = (1 - k\Sigma d_{i1}^2)(c_2 - k\Sigma d_{i2}^2) - (c_1 + k\Sigma d_{i1}d_{i2})^2$

The Rao-Robson statistic for testing goodness-of-fit to the Extreme Value distribution is then:

$$RR = \frac{k}{n} \Sigma \, (N_i - \frac{n}{k})^2 + \frac{k^2}{n} \sum_{j,k} \{\Sigma_i \, (N_i - \frac{n}{k})u_{ij}\}\{\Sigma_i \, (N_i - \frac{n}{k})u_{ik}\}a_{jk}$$

A copy of the Fortran program for the calculation of this statistic is included in the Appendix along with the required subroutines (provided by Dr. M.A. Stephens) for the calculation of the parameter estimates.

e) The Circular Bivariate Normal Distribution

To test the fit of data to the Circular Bivariate Normal distribution, $f(x,y;\underline{\theta})$, the probability density function, is defined as:

$$f(x,y;\underline{\theta}) = \frac{1}{2\pi\sigma^2} \exp \frac{-1}{2\sigma^2} ((x - \mu_1)^2 + (y - \mu_2)^2) \qquad \sigma^2 > 0$$

The MDL estimates of the unknown parameters $\underline{\theta} = (\mu_1,\mu_2,\sigma^2)$ are:

$$\hat{\mu}_1 = \bar{x}$$
$$\hat{\mu}_2 = \bar{y}$$
$$\hat{\sigma}^2 = \frac{1}{2n}\{ \sum_{j=1}^{n} (x_j - \bar{x})^2 + \sum_{j=1}^{n} (y_j - \bar{y})^2 \} = s^2$$

The covariance matrix of this distribution is:

$$V = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & 4\sigma^2 \end{bmatrix}$$

and $\hat{V}^{-1}$ is:

$$\hat{V}^{-1} = \begin{bmatrix} \frac{1}{s^2} & 0 & 0 \\ 0 & \frac{1}{s^2} & 0 \\ 0 & 0 & \frac{1}{4s^2} \end{bmatrix}$$

For the choice of equally probable cells, define the cell boundaries $c_o,\ldots,c_k$ as:

$$c_o = 0$$
$$c_i = \{-2\log(1 - \frac{i}{k})\}^{\frac{1}{2}} \quad \text{for } i=1,2,\ldots,k-1$$
$$c_k = \infty$$

The cells are then annuli centered at $(\bar{x},\bar{y})$. If we denote the i^{th} cell by I_i, then for i=1,....,k,

$$I_i = \{(x,y): c_{i-1}^2 s^2 < (x-\bar{x})^2 + (y-\bar{y})^2 < c_i^2 s^2\}$$

The u_{ij}'s, for j=1, 2, 3 are:

$$u_{i1} = \iint_{I_i} \frac{\partial f_1}{\partial \mu_1} \, dx_1 dy_1$$

$$u_{i2} = \iint_{I_i} \frac{\partial f_1}{\partial \mu_2} \, dx_1 dy_1$$

and

$$u_{i3} = \iint_{I_i} \frac{\partial f_1}{\partial \sigma^2} \, dx_1 dy_1$$

where $f_1 = f(x_1,y_1;\underline{\theta})$. When evaluated at $\hat{\underline{\theta}}$, u_{i1} and u_{i2} are equal to zero, and u_{i3} becomes:

$$u_{i3} = \frac{1}{s}\{c_{i-1}^2 \exp(-\tfrac{1}{2}c_{i-1}^2) - c_i^2 \exp(-\tfrac{1}{2}c_i^2)\}$$

If we define $d_i = \frac{su_{i3}}{2}$, substitution into the general form of the Rao-Robson statistic yields the statistic for testing fit to the Circular Bivariate Normal distribution:

$$RR = \frac{k}{n} \sum (N_i - \frac{n}{k})^2 + \frac{k^2}{n} \frac{(\sum_{i=1}^{k} N_i d_i)^2}{1 - k \sum d_i^2}$$

A copy of the Fortran program for the calculation of this statistic is contained in the Appendix.

ii) The Dzhaparidze-Nikulin Statistic

Returning once again to Wald's Method, define

$$Q = I - B(B'B)^{-1}B'$$

where B is (as for the statistic RR) the kxs matrix with ij^{th} entry

$$p_i^{-\frac{1}{2}} \frac{\partial p_i}{\partial \theta_j}$$

then $W = V'QV$ is the Dzhaparidze-Nikulin statistic.

K.O. Dzhaparidze and M.S. Nikulin (1974) sought to find a statistic which would, when the MDL estimators or other $n^{\frac{1}{2}}$-consistent estimators are used, have the same asymptotic distribution as Pearson's X^2 using the MML estimators; that is, a statistic asymptotically chi-square distributed with k-s-1 degrees of freedom.

Analogous to Rao and Robson's addition to Pearson's X^2 of a term Y^2 converging in law to $\sum_{i=1}^{s} (1-\lambda_i)y_{k-s-1+i}$, Dzhaparidze and Nikulin derived a term converging to $\sum_{i=1}^{s} -\lambda_i y_{k-s-1+i}$. Whereas the Rao-Robson statistic recovers the partial loss of degrees of freedom resulting when the MDL estimates are used in the calculation of X^2, the Dzhaparidze-Nikulin statistic loses degrees of freedom so that its asymptotic distribution is equivalent to that of X^2 when the MML method of estimation is employed.

Their development given in Dzhaparidze and Nikulin is very brief and is essentially contained in a single theorem.

Simulations indicate the Dzhaparidze-Nikulin (DN) statistic is generally not as powerful as the RR statistic (Moore (N.D.)), and it is therefore preferable to employ RR in tests of fit for best results. However, DN is versatile in that it can be used with any reasonable estimate of $\underline{\theta}$ -- DN is chi-square distributed with k-s-1 degrees of freedom whenever $\hat{\underline{\theta}}$ approaches $\underline{\theta}$ at the usual $n^{\frac{1}{2}}$ rate. It is, further, a useful substitute for RR in cases where $(I-BJ^{-1}B')$ is not invertible, a theoretical requirement of RR. The form of DN is here derived for the case of testing fit to the Double Exponential distribution where RR is not defined for this reason.

a) The Double Exponential Distribution

Suppose we suspect that a random sample of n observations was taken from a population following the Double Exponential distribution. Then, the probability density function f(x) is:

$$f(x) = \frac{1}{2\theta_2} \exp\left(\frac{-|x-\theta_1|}{\theta_2}\right) \qquad -\infty < x < \infty, \quad -\infty < \theta_1 < \infty, \quad \theta_2 > 0$$

The covariance matrix is:

$$\frac{V}{n} = \begin{bmatrix} \frac{\theta_2^2}{n} & 0 \\ 0 & \frac{\theta_2^2}{n} \end{bmatrix}$$

The maximum likelihood estimates of θ_1 and θ_2 are:

$$\hat{\theta}_1 = \text{median } (x_1,.....,x_n)$$

$$\hat{\theta}_2 = \frac{1}{n} \Sigma \, |x_j - \hat{\theta}_1|$$

The natural choice of cell boundaries is then:

$$c_i = \hat{\theta}_1 + a_i \hat{\theta}_2$$

where for $k=2\nu$, $a_0 = -\infty$, $a_\nu = 0$, $a_k = \infty$, and

$$a_{\nu+i} = a_{\nu-1} = -\log(1 - \frac{i}{\nu}) \quad \text{for } i=1,2,.....,\nu-1$$

Recall that the matrix form of the Dzhaparidze-Nikulin statistic is:

$$DN = V'(I-B(B'B)^{-1}B')V$$

The entries of the matrix B are:

$$B_{ij} = \frac{\partial p_i}{\partial \theta_j} \qquad \text{for } i=1,2,.....,k \quad j=1,2$$

Differentiating first with respect to θ_1 gives:

$$B_{i1} = \begin{cases} \frac{-1}{k} \theta_2 & \text{for } i=1,2,.....,\nu \\ \frac{1}{k} \theta_2 & \text{for } i=\nu+1,.....,k \end{cases}$$

Next differentiating with respect to θ_2 gives:

$$B_{i2} = \frac{1}{2\theta_2}\left(c_{i-1}e^{-c_{i-1}} - c_i e^{-c_i}\right)$$

If we define

$$d_i = c_{i-1}e^{-c_{i-1}} - c_i e^{-c_i}$$

then B'B becomes:

$$B'B = \frac{1}{\theta_2^2}\begin{bmatrix} 1 & 0 \\ 0 & \nu\Sigma d_i^2 \end{bmatrix}$$

Hence, the Dzhapardize-Nikulin statistic DN, after some simplification, becomes:

$$DN = V'(I - B(B'B)^{-1}B')V$$

$$= \frac{k}{n}\Sigma\left(N_i - \frac{n}{k}\right)^2 - \frac{k}{n}\cdot\frac{1}{2\Sigma d_i^2}\left\{\Sigma d_i\left(N_{\nu+i} + N_{\nu-i+1}\right)\right\}^2$$

The statistic DN has the chi-square limiting null distribution with k-3 degrees of freedom. A copy of the Fortran program for the calculation of this statistic is included in the Appendix.

2. Other Techniques

In this section several "nonstandard" tests of fit are considered. The Rao-Robson and Dzhaparidze-Nikulin statistics studied earlier are quadratic forms in the normalized cell counts with large-sample theory analogous to that of Pearson's X^2. The tests reviewed in this section differ in one or both of these aspects and for this reason are considered separately here.

i. The Kempthorne Statistic

The paper "The Classical Problem of Inference: Goodness-of-Fit" by O. Kempthorne (1968) is reviewed in Moore (1976), but the statistic presented by Kempthorne is not considered as a serious competitor of the standard chi-square statistics. Moore states that preliminary simulations have shown K to be superior in power to standard chi-square tests only for very short-tailed alternatives and may be quite inferior in other cases.

The asymptotic theory underlying standard chi-square statistics changes radically if the number of cells, k, is allowed to increase with the sample size n at a rate faster than $o(n^{\frac{1}{2}})$. Such is the case of Kempthorne's Statistic K. K is simply Pearson's X^2 with k = n cells, each equiprobable with $p_i = \frac{1}{n}$, for i=1, 2,....., k, under the null hypothesis; that is, the Kempthorne statistic is given by:

$$K = \Sigma\ (N_i - 1)^2$$

For the Simple Hypothesis Case, the N_i are Multinomial and K has a Normal limiting null distribution (see, for example, Morris (1975)). For the Composite Hypothesis Case, the limiting null distribution has not been investigated, but it is suspected that the limiting null distribution of K will remain unchanged.

ii. Dahiya-Gurland Statistic

John Gurland and Ram Dahiya also developed a test statistic free of the complications associated with Pearson's X^2. In particular, the question of how to form class intervals has been removed in their proposed test of fit. Further, their statistic is distributed in the limit exactly as chi-square when the parameters are estimated from the ungrouped data. Their statistic is non-standard, however, in that it does not involve cell counts, and hence the large-sample theory is not associated with that of X^2. The test of fit that Dahiya and Gurland propose is for continuous distributions, but the authors indicate it can be adapted to discrete distributions.

The development of the Dahiya-Gurland statistic is based on sample moments, a review of which follows.

Let $x_1, x_2, \ldots, x_n$ represent a random sample from a certain distribution with probability density function $f(x;\underline{\theta})$ where $\underline{\theta}$ is defined in the usual manner. Denote the j^{th} raw moment by:

$$m_j' = \frac{1}{n} \sum_{j=1}^{n} x_i^j$$

and let $\underline{m}' = (m_1', m_2', \ldots, m_q')$ where q, q s, is a fixed number that remains to be specified (a low value of q is generally desirable due to the large sampling fluctuations of higher order moments). Let $\underline{m}^{*}{}' = (m_1^{*}{}', m_2^{*}{}', \ldots, m_q^{*}{}')$ represent the population counterpart of $\underline{m}'$. Further, let h_i, $i = 1, 2, \ldots, q$, be functions of $\underline{m}'$ such that their population counterparts h_i^{*} are differentiable to the second order with respect to $m_1^{*}{}', m_2^{*}{}', \ldots, m_q^{*}{}'$:

$$\underline{h}' = (h_1, h_2, \ldots, h_q)$$

$$\underline{h}^{*}{}' = (h_1^{*}, h_2^{*}, \ldots, h_q^{*})$$

Further, let $Q = JGJ'$ where J denotes the qxq Jacobian matrix with ij^{th} entry $\frac{\partial h_i^*}{\partial m_j^{*'}}$ and G the matrix with ij^{th} entry $(m_{i+j}^{*'} - m_i^{*'}m_j^{*'})$.

The vector of moments $\underline{m}'$ is Multivariate Normal with mean $\underline{m}^{*'}$ and covariance matrix G ($MVN(\underline{m}^{*'}, G)$). A Taylor expansion of $n^{\frac{1}{2}}\underline{h}$ yields the result that $n^{\frac{1}{2}}(\underline{h} - \underline{h}^*)$ is $MVN(\underline{0}, Q)$. Hence, by the theory of the distribution of quadratic forms,

$$DG^* = n(\underline{h} - \underline{h}^*)'Q^{*-1}(\underline{h} - \underline{h}^*)$$

is asymptotically chi-square distributed with q degrees of freedom.

Thus far it has been assumed that Q^{*-1} is known when in fact it requires estimation. However, if Q is a consistent estimator of Q^* (which is obtained from Q^* on replacing the parameters with maximum likelihood or other consistent estimators) then the asymptotic distribution of

$$DG = n(\underline{h} - \underline{h}^*)'Q^{-1}(\underline{h} - \underline{h}^*)$$

is the same as the asymptotic distribution of DG^* (see Gurland (1948) and Barankin and Gurland (1951)).

If the functions h_i are chosen in such a manner that h_i^*, $i = 1, 2, \ldots, q$, are linear functions of the parameters $\theta_1, \theta_2, \ldots, \theta_s$, then an estimator of $\underline{\theta}$ can be found by minimizing the expression for DG. In particular, letting $\underline{h} = W\underline{\theta}$ where W is a qxs matrix of known constants, then the estimator $\hat{\underline{\theta}}$ is given by:

$$\hat{\underline{\theta}} = (W'Q^{-1}W)^{-1}W'Q^{-1}\underline{h}$$

In this setting, we can view the problem of estimating $\underline{\theta}$ as the linear regression of $\underline{h}$ on the parameters $\underline{\theta}$; the errors are approximately normal so that the technique of generalized least squares was applicable in determining $\hat{\underline{\theta}}$.

As a final step, let:

$$\underline{h} = W\hat{\underline{\theta}}$$

$$\hat{R} = W(W'Q^{-1}W)^{-1}W'Q^{-1}$$

$$\hat{A} = Q^{-1}(I - \hat{R})$$

Then

$$\hat{DG} = n(\underline{h} - \hat{R}\underline{h})'Q^{-1}(\underline{h} - \hat{R}\underline{h})$$

$$= n\underline{h}'(I - \hat{R})'Q^{-1}(I - \hat{R})\underline{h}$$

$$= n\underline{h}'\hat{A}\underline{h}$$

Roughly speaking $\hat{DG}$ may now be viewed as the error sum of squares in the generalized least squares procedure and the conclusions follow immediately. The asymptotic distribution of $n\underline{h}'\hat{A}\underline{h}$ is the same as the asymptotic distribution of $n\underline{h}'A\underline{h}$, where A is obtained by replacing Q by Q^* in $\hat{A}$. Assuming W is of rank s, the null distribution of $n\underline{h}'A\underline{h}$ is chi-square with q-s degrees of freedom (Gurland (1948) and Barankin and Gurland (1951)).

For illustrative purposes, a test of fit for the Normal distribution is derived: The development presented here is that given by Gurland and Dahiya (1972); they provide a clear and easy-to-follow formulation:

Suppose we wish to test the hypothesis that X has pdf

$$f(x;\underline{\theta}) = \frac{1}{(2\pi\theta_2)^{\frac{1}{2}}} \exp\left(-\frac{(x - \theta_1)^2}{2\theta_2}\right)$$

$$-\infty < x < \infty, \quad -\infty < \theta_1 < \infty, \quad \theta_2 > 0$$

Let m_2, m_3, and m_4 denote the second, third, and fourth central sample moments respectively. The statistics b_1, b_2 given by

$$b_1 = \frac{m_3}{m_2^{3/2}}, \quad b_2 = \frac{m_4}{m_2^2}$$

If we define

$$\theta_2^* = \log \theta_2$$

$$\underline{h}^{*\prime} = (m_1^{*\prime}, \log m_2^*, m_3^*, \log (\frac{m_4^*}{3}))$$

then the elements of $\underline{h}^*$ are linear functions of the parameters θ_1 and θ_2^* so that we can now write

$$\underline{h} = W\underline{\theta}^*$$

with

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 2 \end{bmatrix}, \quad \underline{\theta}^* = \begin{bmatrix} \theta_1 \\ \theta_2^* \end{bmatrix}$$

The corresponding h_i functions are given by

$$h_1 = m_1' , \; h_2 = \log m_2 , \; h_3 = m_3 , \; h_4 = \log (\frac{m_4}{3})$$

where m_1' is the sample mean, and m_2, m_3, and m_4 denote the second, third, and fourth central sample moments respectively, as previously indicated.

The transformation from sample raw moments to functions h_i is achieved in two stages, i.e., from m_1', m_2', m_3', m_4' to m_1', m_2, m_3, m_4 and then finally to h_1, h_2, h_3, h_4.

The asymptotic distribution of $n^{\frac{1}{2}}(\underline{h} - \underline{h}^*)$ is $N(\underline{0},Q^*)$ where

$$Q^* = J_2 J_1 G J_1' J_2'$$

$$J_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -3\theta_2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad J_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/\theta_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1/(3\theta_2^2) \end{bmatrix}$$

and

$$G = \begin{bmatrix} \theta_2 & 0 & 3\theta_2^2 & 0 \\ 0 & 2\theta_2^2 & 0 & 12\theta_2^3 \\ 3\theta_2^2 & 0 & 15\theta_2^3 & 0 \\ 0 & 12\theta_2^3 & 0 & 96\theta_2^4 \end{bmatrix}$$

After simplification we obtain:

$$Q^* = \begin{bmatrix} \theta_2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & 6\theta_2^3 & 0 \\ 0 & 4 & 0 & 32/3 \end{bmatrix}$$

Replacing θ_2 by its maximum likelihood estimator m_2 to obtain Q, we have the statistic $\hat{DG}$ for testing normality where:

$$\hat{DG} = n\underline{h}'\hat{A}\underline{h}, \quad \hat{A} = Q^{-1}(I - \hat{R})$$

$$\hat{R} = W(W'Q^{-1}W)^{-1}W'Q^{-1}$$

After simplification

$$\hat{A} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1.5 & 0 & -.75 \\ 0 & 0 & 1/(6m_2^3) & 0 \\ 0 & -.75 & 0 & .375 \end{bmatrix}$$

so that a simplified form of $\hat{DG}$ is given by $\hat{DG} = n\underline{v}'B\underline{v}$ where

$$\underline{v}' = (h_2, h_3, h_4) = (\log m_2, m_3, \log(\frac{m_4}{3}))$$

$$B = \begin{bmatrix} 1.5 & 0 & -.75 \\ 0 & 1/(6m_2^3) & 0 \\ -.75 & 0 & .375 \end{bmatrix}$$

The statistic $\hat{DG} = n\underline{v}'B\underline{v}$ can be easily computed on a desk calculator. Its asymptotic distribution is chi-square with 2 degrees of freedom (here q=4 and s=2). To carry out a test of fit for normality at a particular level of significance, one merely requires the corresponding critical point of the Chi-Square distribution.

Dahiya and Gurland go on to prove that the power of the test based on $\hat{DG} = n\underline{v}'B\underline{v}$ is invariant with respect to the location and scale parameters of the alternative distribution and calculate the power of the statistic for normality against five alternatives.

iii. The Effect of Dependence on Chi-Square Tests of Fit

David Moore (1977b) in his paper entitled "The Effect of Dependence on Chi-Square Tests of Fit" examines the effects on Pearson's chi-square statistic when the assumptions of independent, identically distributed (iid) random variables is not valid. In particular, he examines the case where the data form a Stationary Stochastic Process (SSP).

In practice, it is common to assume that the observations on which a test of fit will be based are iid. This may not always be reasonable as when observations are from a time series. Suppose then that $X_1, X_2, \ldots, X_n$ are observations on a SSP, and that a goodness-of-fit test is to be performed requiring that the data be iid. Moore examines the effect of the dependence on such a test when the null hypothesis is true. From the minimal literature already existing on the subject (Gasser, 1975), it was discovered from a small simulation study including Gaussian autoregressive processes that when iid critical points were used, Pearson's X^2 test rejected normality too often. In addition, the test was more powerful against iid alternatives than against non-Gaussian autoregressive processes.

Moore undertakes a theoretical study of the effects of dependence, again using the Pearson X^2 test. He shows that for a general class of Gaussian SSP's, positive correlation "is confounded with lack of normality" as implied by Gasser's study. Since Moore's formulation makes use of only one property of the Normal laws, the results can be extended to include other distributions as well.

Moore ascertains that if the SSP structure of the data is known, a test statistic for goodness-of-fit can be produced that will have a known limiting null distribution. This area remains open for investigation.

3. Special Applications

In the field of goodness-of-fit testing today, the movement is more and more towards the formulation of specially designed tests and test statistics to meet the needs of precise conditions arising in real-life settings. It is not surprising, therefore, to observe the adaptation of chi-square tests of fit to special conditions as well.

Much of the more recent work of David Moore (1979 to 1981) has been to this end. In this section, some of the findings that have extended the contribution of chi-square techniques to goodness-of-fit testing are examined.

i. Chi-Square Tests of Fit for Type II Censored Data

Type II censored data is data that are "censored on one or both sides at sample percentiles". Data of this type may result from engineering studies. A chi-square goodness-of-fit test can be applied to Type I censored data, or data that are censored at fixed points, as the censored observations fall into one or more fixed cells. By similarly choosing cells as sample percentiles, chi-square tests can also be applied to Type II censored data. This is the objective and goal of Daniel Milhalko and David Moore (1980).

The development Milhalko and Moore present results in goodness-of-fit test statistics that are asymptotically chi-square distributed. For large n, this eliminates the need for separate tabled critical points for each hypothesized family. Since, in addition, they develop a test for the Composite Hypothesis Case (many tests of fit for completely specified distributions have already been proposed), they provide a very useful tool. Due to the dependence of the Type II data, the proofs included in their formulation are analogous to, but quite different from, the usual large sample theory of chi-square statistics provided in Moore and Spruill (1975).

Following is a review of their proposal:

Suppose that from a random sample $x_1, x_2, \ldots, x_n$ we observe only the order statistics:

$$x_{(\{n\alpha\}+1)} < x_{(\{n\alpha\}+2)} < \cdots < x_{(\{n\beta\})}$$

where $0 \leq \alpha < \beta \leq 1$ and $\{x\}$ denotes the greatest integer in x. The k cells are formed by taking the cell boundaries $c_0, c_1, \ldots, c_k$ to be $c_i = x_{(\{n\delta_i\})}$, the δ_i-quantile from $x_1, \ldots, x_n$ with $0=\delta_0<\delta_1< \cdots <\delta_k=1$. To accommodate nontrivial left censoring (that is, $\alpha>0$), right censoring (that is, $\beta<1$), or

both, with a single notation, let $\alpha=\delta_1$ when $\alpha>0$ and otherwise $\alpha=\delta_0=0$. Similarly, $\beta=\delta_{k-1}$ when $\beta<1$ and otherwise $\beta=\delta_k=1$. The observed frequencies N_i, for $i=1, 2,\ldots, k$, are then nonrandom with $N_i = \{n\delta_i\} - \{n\delta_{i-1}\}$.

Suppose we wish to test the composite null hypothesis that the distribution of the x_i is a member of the family of continuous distribution functions $F(X;\underline{\theta})$. The parameter $\underline{\theta}$ must be estimated by an estimator $\hat{\underline{\theta}}$ which is a function of the observed ordered statistics.

Chi-square tests of fit which employ data-dependent cells are constructed by "forgetting" that the cells are functions of the data. Therefore, the probability that an observation will fall into the i^{th} cell is:

$$p_i = F(c_i;\hat{\underline{\theta}}) - F(c_{i-1};\hat{\underline{\theta}}),\ i=1, 2,\ldots, k$$

Since the p_i's depend on the estimates of the parameters $\underline{\theta}$, they are random quantities, unlike the N_i's.

The derivation of the asymptotic normality of the vector of standardized cell frequencies in both the central and noncentral cases for a quite general class of estimators is provided by Milhalko and Moore. It is, however, a rather lengthy procedure and is not detailed here. The approach is to treat the central case first and then to use contiguity methods to obtain corresponding noncentral results.

Based on the results of this derivation, the large-sample behaviour of several chi-square statistics for Type II censored data is discussed. The statistics are Pearson's X^2, the Chernoff-Lehmann statistic (Pearson's X^2 using the MDL method of estimation), the Rao-Robson statistic, and the Dzhaparidze-Nikulin statistic.

For illustrative purposes, the test derived for the Exponential distri-

bution will be reviewed. The test statistic employed is the Rao-Robson statistic RR. The censored sample analog of RR is:

$$RR = X^2 + V'B(K-B'B)^{-1}B'V$$

where X^2 represents Pearson's statistic, V is the covariance matrix of $\underline{\theta}$, B is the matrix with ij^{th} entry:

$$p_i^{\frac{1}{2}} \frac{\partial p_i}{\partial \theta_j}$$

and K is the Fisher information matrix of the ordered data.

Using simplifications that arise in the case where $F(X;\underline{\theta})$ is from a location-scale family, Milhalko and Moore show that:

$$K = \theta_0^{-2}$$

$$B'B = \theta_0^{-2} \Sigma \frac{\nu_i^2}{p_i}$$

where $\nu_i = -(1-\delta_i)\log(1-\delta_i) + (1-\delta_{i-1})\log(1-\delta_{i-1})$ and

$p_i = \delta_i - \delta_{i-1}$

When the cell boundaries are chosen to be the sample δ_i-quantiles, then the form of the Rao-Robson statistic becomes:

$$RR = X^2 + (n\Delta)^{-1}\left(\Sigma \frac{N_i \nu_i}{p_i}\right)^2$$

where $\Delta = 1 - \exp(-x_{(\{n\beta\})}/\hat{\theta}) - \Sigma\nu_i^2/p_i$

RR is asymptotically chi-square distributed with k-1 degrees of freedom.

Test statistics for the Normal family, the two-parameter Uniform family,

and Weibull family are also derived.

ii. Chi-square Tests for Multivariate Normality

Another special distribution that David Moore undertook to develop a chi-square goodness-of-fit test for was the Multivariate Normal distribution. Together with Stubblebine, he applies the theory of chi-square tests with data-dependent cells to this family. When Pearson's X^2 is employed, it has critical points asymptotically bounded between those of the chi-square distribution with k-1 degrees of freedom and k-2 degrees of freedom. It has proven sensitive in the detection of peakedness, broad shoulders, and heavy tails. Since, as noted by Andrews, Gnanadesikan, and Warner (1973) in their summary of proposed methods of assessing Multivariate Normality, it is desirable to have a variety of procedures that are sensitive to some of the possible departures from joint normality, the added fact that it is not sensitive to lack of symmetry is not a serious handicap. In the case where a lack of symmetry is suspected, an alternate test would be appropriate.

In brief, Moore and Stubblebine propose a chi-square test for Multivariate Normality using data-dependent cells bounded by hyperellipses $((\underline{x} - \underline{\bar{x}})'S^{-1}(\underline{x} - \underline{\bar{x}}) = c_i$, for i=1, 2,...,k) with parameters estimated from the data. The hyperellipses are centred at the sample mean $\underline{\bar{x}}$ with their shape being determined by the sample covariance matrix S. This test would fall into the category of Andrews, et al., "tests based on distributional densities".

The advantages of the proposed statistic cited by Moore and Stubblebine are several:

1. Cells having prespecified estimated cell probabilities are easy to choose.

2. The test statistic is relatively easy to evaluate, and the analysis is particularly simple when the cells are equiprobable.

3. The large sample theory of the test is nearly standard and allows use of chi-square critical points to assess the significance of the statistic.

4. The nature of departures from normality is indicated by the observed cell frequencies. Common departures exhibited by peakedness, broad shoulders, and heavy tails are directly apparent in the cell counts.

5. Once the boundaries (c_i = 1=0,1,...,k) are selected, the estimated cell probabilities p_i, i=1,2,...,k are fixed. For the particular choice of c_i equal to the $\frac{i}{k}$ point of the appropriate Chi-Square distribution, these cells are equiprobable.

6. The Pearson statistic is affine invariant, that is, unaffected by affine transformations on the x_j. The relationship between the cells and data implies affine invariance of V, where $X^2 = V'V$, and hence of X^2. Other statistics than Pearson's considered by the authors are also affine invariant.

Typically, chi-square tests are not highly sensitive, and in this multivariate circumstance X^2 must, as in the univariate case, compete for usage on the basis of its ease of application and interpretation.

The test that is investigated is, as mentioned above, the data-dependent cell version of Pearson's X^2, which was studied by Chernoff and Lehmann (1954) and sometimes subsequently referred to as the Chernoff-Lehmann statistic. This statistic is defined in the usual manner:

$$X^2 = V'V = \sum_{i=1}^{k} \frac{(N_i - np_i)2}{np_i}$$

where V is again the vector of standardized cell frequencies $\frac{N_i - np_i}{(np_i)^{\frac{1}{2}}}$ with

the parameters on which the probabilities p_i depend estimated by the Maximum Density Likelihood method from the ungrouped data.

The theory underlying the Chernoff-Lehmann and other statistics in this setting follows from that of V. The results of Moore and Stubblebine's development are given in their Theorem 1 which states:

> Under the null hypothesis of normality the limiting distribution of the Pearson statistic $X^2(\hat{\underline{\theta}})$ for cells defined by $c_{i-1} \leq (\underline{x} - \hat{\underline{\theta}}_1)'S^{-1}(x - \hat{\underline{\theta}}_1) < c_i$, where $\hat{\underline{\theta}}_1 = (\bar{x}_1, \bar{x}_2 \ldots, \bar{x}_v)$ with parameters estimated by the MDL estimators $\underline{\bar{x}}$ and S is distributed asymptotically as the sum of a chi-square variable with k-2 degrees of freedom and a chi-square variable with 1 degree of freedom and coefficient λ where the variables are independent chi-square random variables with the indicated coefficient satisfying $0 \leq \lambda < 1$. When $p_i = \frac{1}{k}$, then

$$\lambda = 1 - 2kv \sum_{i=1}^{k} d_i^2$$

$$\text{where } d_i = (c_{i-1}^{v/2} e^{-c_{i-1}/2} - c_i^{v/2} e^{-c_i/2}) \frac{b_v}{2}$$

$$\text{and } b_v = (v(v-2)\ldots(4)(2))^{-1} \qquad \text{for v even}$$

$$= (\tfrac{2}{\pi})^{\frac{1}{2}}(v(v-2)\ldots(5)(3))^{-1} \qquad \text{for v odd}$$

This implies the previously-mentioned conclusion that the critical points of X^2 are bounded between those of χ^2_{k-2} and χ^2_{k-1}. For practical purposes, unless k is very small, these bounds are sufficient.

Two alternatives that are suggested for this Pearson test, but not pursued in detail due to the acceptability of Theorem 1, are the computation

of the exact asymptotic critical points following the methods of Dahiya and Gurland (1972) and Moore (1971), or the use of the Rao-Robson statistic RR. Except for small k, the improvement realized from the exact asymptotic critical points will probably not be significant, and as far as employing the RR statistic is concerned, it is computationally complex in most instances.

The results of Moore and Stubblebine's development is summarized in their Theorem 2:

When (X_i, Y_i) have density function:

$$f(x,y|\theta) = (2\pi\sigma^2)^{-1} \exp\left(-\frac{1}{2\sigma^2}\{(x-\mu_1)^2+(y-\mu_2)^2\}\right)$$

the statistic

$$R = \frac{k}{n} \sum (N_i - \frac{n}{k})^2 + \frac{1}{n} \frac{(\sum N_i v_i)^2}{4-k\sum v_i^2}$$

has the $\chi^2(k-1)$ asymptotic distribution.

The final undertaking of the authors is to illustrate the use of this adapted Pearson test to the logarithms of common stock prices usually assumed to be Multivariate Normal. For comparative purposes, they also apply this test to simulated Multivariate Normal data.

IV. PRACTICAL APPLICATIONS

To illustrate the use of the Rao-Robson statistic, it is used in the following examples to test fit to:

1. The Normal Distribution
2. The Exponential Distribution
3. The Extreme Value Distribution

The real data sets were chosen to allow for comparison of the results with those of other test procedures. The data sets were drawn from Spinelli (1971) originally having been provided by Dr. W. G. Warren in the first two cases and van Montfort (1973) in the third. Spinelli performs several Regression and EDF tests on each data set.

Following is a summary of the Rao-Robson tests of fit and a comparison of the results with those of Spinelli's procedures.

Example 1: A Test for Normality

Sixty-four values of modulus of elasticity were measured on Douglas Fir and Larch two-by-fours; a sample of 50 of these values was tested for normality. The ordered data points are as follows:

43.19	45.84	49.44	51.55	54.14
55.37	56.93	59.63	60.04	61.07
65.74	67.09	72.24	72.34	73.46
76.52	77.35	78.36	78.47	78.79
82.00	83.57	84.95	86.59	87.96
90.19	91.57	91.74	92.45	94.24
94.54	95.00	98.39	99.74	100.22
103.48	105.54	107.13	108.14	108.64
108.94	109.62	110.81	112.75	116.39
116.79	119.46	120.33	121.16	131.57

The number of cells used in the test was k = 10. The value of the Rao-Robson statistic calculated from the data was 4.75904. Comparison with the exact χ^2_{k-1} or χ^2_9 percentage points shows a significance level of greater than .80. Hence, the Rao-Robson test fails to reject the hypothesis that the data is normally distributed.

This result again agrees with the test results obtained by Spinelli where all 10 test procedures accepted the null hypothesis and concluded normality of the data.

Example 2: A Test for Exponentiality

Thirty-two values of modulus of rupture measured on Douglas Fir and Larch two-by-fours were tested for exponentiality. The ordered data points are as follows:

43.19	49.44	51.55	55.37	56.63	67.27	78.47	86.59
90.63	92.45	94.24	94.35	94.38	98.21	98.39	99.74
100.22	103.48	105.54	105.54	107.13	108.14	108.64	108.94
109.62	110.81	112.75	113.64	116.39	119.46	120.33	131.57

The number of cells used in the test was k = 5. The value of the Rao-Robson statistic calculated from one data was 66.1773. Comparison with the exact x^2_{k-1} or x^2_4 percentage points shows a significance level of less than .005. Hence, the Rao-Robson test rejects the hypothesis that the data came from the Exponential distribution.

This agrees with the test results obtained by Spinelli where all 10 test procedures also strongly rejected the null hypothesis.

Example 3: A Test for the Extreme Value Distribution

Forty-seven values in cubic feet per second of annual maxima of the discharges of the North Saskatchewan River at Edmonton were tested for the Extreme Value distribution. The ordered data points are as follows:

19.885	20.94	21.82	24.888	27.5
28.1	28.6	30.38	31.5	38.1
39.02	40.0	40.0	40.4	44.7
50.33	51.442	58.8	61.2	65.597
66.0	84.1	106.6	121.97	185.56

The number of cells used in the test was k = 5. The value of the Rao-Robson statistic obtained from the data was 18.71547. Comparison with the exact χ^2_{k-1} or χ^2_4 percentage points shows a significance level of less than .005. Hence, the Rao-Robson test rejects the hypothesis that the data came from the Extreme Value distribution.

The summary of test results provided by Spinelli shows that at a .10 significance level, seven of the 10 test procedures rejected the null hypothesis and at a .05 significance level, four of the 10 rejected the null hypothesis.

V. MONTE CARLO POINTS FOR FINITE SAMPLE SIZE

For each of the six distributions to which the theory of the Rao-Robson or Dzhaparidze-Nikulin statistic was applied, Monte Carlo methods were used to simulate the percentage points for the test statistics for finite sample sizes. Three different sample sizes were considered, namely n = 20, 50, and 100; 10,000 samples were generated in each case. In all instances, equiprobable cells were formed, and the parameter values were assumed unknown and were estimated by Maximum Density Likelihood.

Monte Carlo points for the test statistics were also simulated for an additional sample size employing the same number of cells (for example, n=20, k=4, and n=50, k=4). This generated two points from the same distribution (namely the Chi-Square distribution with k-1 or k-s-1 degrees of freedom for the Rao-Robson and Dzhaparidze-Nikulin statistics respectively) which together with the asymptotic distribution point were used for the purpose of "smoothing" the points. The indication in all cases was that smoothing was not warranted, and the actual points derived in the main runs were instead left unadjusted.

For use with the power studies, the Monte Carlo points were determined for Pearson's chi-square statistic X^2 in the same manner. These points are included alongside of the associated Rao-Robson or Dzhapardize-Nikulin points in Tables 1 to 18 following. In addition, the exact asymptotic chi-square points are provided on the right hand side of each table for comparative purposes.

In most instances, convergence of the finite n points to the asymptotic points appears reasonably rapid. Note, however, that in the case of the Exponential distribution, convergence is comparatively slow with very high values still existing at n = 100.

You may note that there are some critical points that do not change from probability level to another. This is due to the fact that the Rao-Robson and Dzhaparidze-Nikulin statistics are discrete statistics; even though the parameters estimated have continuous distributions, terms in the parameters cancel out, and the statistics depend only on the number of observations per cell. Hence, the probability may be concentrated in areas.

TABLE 1

The Normal Distribution

n = 20 k = 4

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_3
.990	7.59	8.10	11.34
.975	5.99	6.24	9.35
.950	5.19	5.44	7.82
.900	3.99	4.04	6.25
.850	2.79	3.30	5.32
.750	2.39	2.81	4.11
.500	1.19	1.27	2.37
.250	0.39	0.56	1.21
.100	0.39	0.41	.58
.050	0.39	0.41	.35
.025	0.07	0.07	.22
.010	0.07	0.07	.12

TABLE 2

The Normal Distribution

$n = 50 \quad k = 10$

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_9
.990	18.78	20.26	21.67
.975	16.38	17.74	19.02
.950	14.38	15.50	16.92
.900	12.38	13.26	14.68
.850	10.78	11.90	13.29
.750	9.18	10.06	11.40
.500	6.78	7.22	8.34
.250	4.78	4.98	5.90
.100	3.18	3.42	4.17
.050	2.38	2.70	3.33
.025	1.98	2.18	2.70
.010	1.58	1.62	2.09

TABLE 3

The Normal Distribution

n = 100 k = 10

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_9
.990	18.97	20.70	21.67
.975	16.37	17.83	19.02
.950	14.38	15.43	16.92
.900	12.38	13.37	14.68
.850	11.18	12.03	13.29
.750	9.38	10,10	11.39
.500	6.58	7.23	8,34
.250	4.58	4,97	5.90
.100	3.18	3.50	4.17
.050	2.58	2.70	3.33
.025	1.98	2.23	2.70
.010	1.58	1.63	2.09

TABLE 4

The Exponential Distribution

n = 20 k = 4

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_3
.990	9.19	13.10	11.34
.975	7.59	10.32	9.35
.950	5.99	8.77	7.82
.900	5.19	7.06	6.25
.850	3.99	6.03	5.32
.750	2.79	4.39	4.11
.500	1.59	2.70	2.37
.250	0.79	1.34	1.21
.100	0.39	0.59	.58
.050	0.39	0.46	.35
.025	0.39	0.46	.22
.010	0.06	0.06	.12

TABLE 5

The Exponential Distribution

n = 50 k = 10

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_9
.990	20.77	33.97	21.67
.975	17.57	29.08	19.02
.950	15.57	25.52	16.92
.900	13.17	21.37	14.68
.850	11.97	19.17	13.29
.750	10.37	16.12	11.39
.500	7.57	11.53	8.34
.250	5.17	7.98	5.90
.100	3.57	5.43	4.17
.050	2.77	4.33	3.33
.025	2.37	3.48	2.70
.010	1.57	2.63	2.09

TABLE 6

The Exponential Distribution

n = 100 k = 10

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_9
.990	20.35	34.43	21.67
.975	17.55	29.10	19.02
.950	15.35	25.17	16.92
.900	13.35	21.30	14.68
.850	11.95	18.97	13.29
.750	10.15	15.90	11.39
.500	7.35	11.43	8.34
.250	5.15	7.97	5.90
.100	3.55	5.57	4.17
.050	2.95	4.37	3.33
.025	2.35	3.57	2.70
.010	1.75	2.70	2.09

TABLE 7

The Double Exponential Distribution

n = 20 k = 4

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Dzhaparidze-Nikulin Statistic	χ^2_1
.990	6.78	5.01	6.64
.975	5.18	4.99	5.02
.950	3.98	3.19	3.84
.900	3.58	3.19	2.71
.850	3.18	1.79	2.07
.750	1.98	0.81	1.32
.500	0.78	0.21	.46
.250	0.38	0.19	.10
.100	0.38	0.06	.016
.050	0.13	0.06	.004
.025	0.13	0.06	.001
.010	0.13	0.06	.000

TABLE 8

The Double Exponential Distribution

n = 50 k = 10

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Dzhaparidze-Nikulin Statistic	χ^2_7
.990	17.97	17.81	18.48
.975	15.58	15.33	16.01
.950	13.58	13.57	14.07
.900	11.98	11.78	12.02
.850	10.78	10.43	10.75
.750	9.18	8.81	9.04
.500	6.38	6.29	6.35
.250	4.38	4.29	4.26
.100	3.18	2.81	2.83
.050	2.38	2.22	2.17
.025	1.98	1.71	1.69
.010	1.58	1.19	1.24

TABLE 9

The Double Exponential Distribution

n = 100 k = 10

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Dzhaparidze-Nikulin Statistic	χ^2_7
.990	18.40	18.12	18.48
.975	15.98	15.78	16,01
.950	14.19	13.93	14.07
.900	11.98	11.73	12.02
.850	10.81	10.53	10.75
.750	8.98	8.88	9.04
.500	6.60	6.33	6.35
.250	4.40	4.18	4.26
.100	2.98	2.78	2.83
.050	2.40	2.18	2.17
.025	1.81	1.68	1.69
.010	1.40	1.18	1.24

TABLE 10

The Circular Bivariate Distribution

$n = 20 \quad k = 4$

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_3
.990	9.99	11.22	11.34
.975	7.59	9.14	9.35
.950	5.99	7.58	7.82
.900	4.79	6.26	6.25
.850	3.99	5.10	5.32
.750	2.79	4.06	4.11
.500	1.59	2.30	2.37
.250	0.79	1.34	1.21
.100	0.39	0.54	.58
.050	0.39	0.42	.35
.025	0.39	0.42	.22
.010	0.07	0.10	.12

TABLE 11

The Circular Bivariate Distribution

n = 50 k = 8

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_7
.990	17.18	18.78	18.48
.975	14.30	15.98	16.01
.950	12.70	13.74	14.07
.900	10.78	11.74	12.02
.850	9.50	10.58	10.75
.750	7.90	8.98	9.04
.500	5.66	6.34	6.35
.250	3.74	4.30	4.26
.100	2.46	2.86	2.83
.050	1.82	2.18	2.17
.025	1.18	1.78	1.69
.010	0.86	1.26	1.24

TABLE 12

The Circular Bivariate Distribution

n = 100 k = 10

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_9
.990	19.97	21.42	21.67
.975	17.58	19.05	19.02
.950	15.58	16.85	16.92
.900	13.38	14.62	14.68
.850	12.18	13.18	13.29
.750	10.38	11.38	11.39
.500	7.58	8.35	8.34
.250	5.18	5.95	5.90
.100	3.58	4.25	4.17
.050	2.98	3.42	3.33
.025	2.38	2.82	2.70
.010	1.78	2.22	2.09

TABLE 13

The Logistic Distribution

n = 20 k = 4

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_3
.990	7.18	10.82	11.34
.975	5.98	9.82	9.35
.950	4.78	7.58	7.82
.900	3.58	6.62	6.25
.850	2.78	5.52	5.32
.750	1.98	4.12	4.11
.500	1.18	2.52	2.37
.250	0.38	1.18	1.21
.100	0.38	0.98	.58
.050	0.08	0.08	.35
.025	0.08	0.08	.22
.010	0.08	0.08	.12

TABLE 14

The Logistic Distribution

n = 50 k = 8

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_9
.990	15.28	18.38	18.48
.975	13.05	15.86	16.01
.950	11.12	13.94	14.07
.900	9.52	11.94	12.02
.850	8.25	10.74	10.75
.750	6.95	8.98	9.04
.500	4.72	6.38	6.35
.250	3.12	4.34	4.26
.100	1.85	2.86	2.83
.050	1.18	2.18	2.17
.025	1.18	1.70	1.69
.010	0.88	1.22	1.24

TABLE 15

The Logistic Distribution

$n = 100 \quad k = 10$

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_9
.990	18.46	21.30	21.67
.975	15.94	18.77	19.02
.950	13.98	16.63	16.92
.900	11.98	14.57	14.68
.850	10.70	13.17	13.29
.750	8.98	11.30	11.39
.500	6.34	8.30	8.34
.250	4.30	5.90	5.90
.100	2.86	4.17	4.17
.050	2.26	3.30	3.33
.025	1.74	2.70	2.70
.010	1.18	2.03	2.09

TABLE 16

The Extreme Value Distribution

n = 20 k = 4

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_3
.990	7.59	10.81	11.34
.975	5.99	9.19	9.35
.950	5.19	7.79	7.82
.900	3.99	6.24	6.25
.850	3.19	5.44	5.32
.750	2.79	3.81	4.11
.500	1.19	2.41	2.37
.250	0.39	1.29	1.21
.100	0.39	0.84	.59
.050	0.39	0.61	.35
.025	0.06	0.06	.22
.010	0.06	0.06	.12

TABLE 17

The Extreme Value Distribution

n = 50 k = 8

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_7
.990	15.26	18.48	18.48
.975	13.34	15.78	16.01
.950	11.42	13.75	14.07
.900	9.50	11.82	12.02
.850	8.54	10.62	10.75
.750	6.94	8.92	9.04
.500	4.70	6.32	6.35
.250	3.10	4.35	4.26
.100	2.14	2.92	2.83
.050	1.50	2.25	2.17
.025	1.18	1.75	1.69
.010	0.86	1.32	1.24

TABLE 18

The Extreme Value Distribution

n = 100 k = 10

	Percentage Points		
$1 - \alpha$	Pearson Statistic	Rao-Robson Statistic	χ^2_9
.990	18.97	21.67	21.67
.975	16.37	19.12	19.02
.950	14.38	16.82	16.92
.900	12.38	14.43	14.68
.850	10.98	13.18	13.29
.750	9.38	11.33	11.39
.500	6.78	8.33	8.34
.250	4.58	5.93	5.90
.100	3.18	4.23	4.17
.050	2.58	3.33	3.33
.025	1.98	2.68	2.70
.010	1.38	2.13	2.09

VI. POWER COMPARISONS

Based on 10,000 samples of sizes 20 and 100, the powers of the Rao-Robson test statistic (as derived for the five distributions previously given) and the Dzhaparidze-Nikulin test statistic (as derived for the Double Exponential distribution) were estimated by Monte Carlo methods. For comparative purposes, the power of the Pearson test statistic was also simulated. In addition, the power of the Anderson-Darling test statistic, A^2, was determined for those distributions for which the percentage points were available; reference to some existing power results for A^2 were also made. The Anderson-Darling statistic is based on formulas provided in Stephens (1974, 1977, and 1979). Specifically, the procedure employed was as follows:

a) estimate parameters by Maximum Density Likelihood

b) calculate $z_i = F(x_i;\underline{\hat{\theta}})$, $i=1,2,\ldots,n$

c) $A^2 = -(\Sigma\ (2i-1)\{\ln z_i + \ln(1-z_{n+1-i})\})/n \ - n$

d) reject the null hypothesis if A^2 exceeds the critical value for a specified significance level.

In all cases, a significance level of $\alpha = .05$ was used. Wherever possible, the alternative distributions were chosen according to those previously employed by others in power simulations (Normal, Extreme Value, Exponential). Otherwise, the alternates were chosen based on their general resemblance to the hypothesized distribution and an attempt to incorporate a variety of departures from the null distribution (such as skewness, heavy tails, broad shoulders, etc.).

For all sample sizes of $n = 20$, $k = 4$ cells were employed; for $n = 100$, $k = 10$ cells were used.

The finite n critical points for the Rao-Robson and Pearson statistics were provided by the Monte Carlo percentage points of the previous chapter.

The finite n critical points for the Anderson-Darling statistic were given in Stephens (1974) for the Normal or Exponential distributions, Stephens (1977) for the Extreme Value distribution and Stephens (1979) for the Logistic distribution. The power comparisons given for A^2 in Table 19 were provided in Stephens (1974).

Dahiya and Gurland produced simulated power results for the Normal distribution. For the case where the alternative is the Logistic distribution, their result at the .05 level of significance and for n = 100 is directly comparable to that obtained for RR. It is clear that the power of the DG statistic is far superior to that of the Rao-Robson statistic, at least in this instance. DG rejected successfully 65.4% of the time versus RR's 10% success.

In general, for small n, the powers of the Rao-Robson and Dzhaparidze-Nikulin statistics are low, surpassed, where considered, by the EDF Anderson-Darling statistic. This is not unexpected given the grouping inherent in chi-square tests. In cases where the alternative distribution is very close to the null distribution (such as in the Normal versus Logistic case above) the improvement of A^2 over RR and X^2 is not significant, with all tests considered having low power. However, the comparisons performed are relatively common, and other, more powerful statistics (some specially designed) are readily applicable and therefore preferable.

TABLE 19

Power Comparisons

Normal Distribution

Alternative	Sample Size	Test Statistic RRStat	X^2	A^{2*}
Logistic	n = 20	.08	.08	--
	n = 100	.10	.09	--
t_1	n = 20	.50	.50	--
	n = 100	1.00	1.00	--
t_4	n = 20	.15	.14	.23
	n = 100	.34	.32	$.69_1$
Lognormal	n = 20	.55	.52	.91
	n = 100	1.00	1.00	1.00_2

* Provided by Stephens (1974).

1. Based on n = 90.

2. Based on n = 50.

TABLE 20

Power Comparisons

Exponential Distribution

Alternative	Sample Size	Test Statistic RRStat	χ^2	A^2
χ^2_6	n = 20	.66	.69	.88
	n = 100	1.00	1.00	1.00
Half-Normal	n = 20	.10	.13	.17
	n = 100	.52	.45	.77
χ^2_8	n = 20	.86	.89	.99
	n = 100	1.00	1.00	1.00

TABLE 21

Power Comparisons

Double Exponential Distribution

Alternative	Sample Size	Test Statistic DNStat	X^2
Logistic	n = 20	.12	.07
	n = 100	.23	.26
Normal	n = 20	.11	.06
	n = 100	.14	.15
t_4	n = 20	.12	.05
	n = 100	.11	.11

TABLE 22

Power Comparisons

Circular Bivariate Distribution

Alternative	Sample Size	Test Statistic RRStat	Test Statistic X^2
Uniform	n = 20	.22	.21
	n = 100	.95	.94
Bivariate Normal	n = 20	.05	.07
	n = 100	.92	.90

TABLE 23

Power Comparisons

Logistic Distribution

Alternative	Sample Size	RRStat	Test Statistic X^2	Test Statistic A^2
Normal	n = 20	.04	.06	.04
	n = 100	.09	.07	.09
t_4	n = 20	.04	.06	.05
	n = 100	.14	.08	.18
χ^2_8	n = 20	.07	.07	.16
	n = 100	.45	.47	.85

TABLE 24

Power Comparisons

Extreme Value Distribution

Alternative	Sample Size	RRStat	Test Statistic X^2	Test Statistic A^2
Beta (1,4)	n = 20	.12	.09	.15
	n = 100	.57	.32	.87
χ^2_6	n = 20	.05	.00	.04
	n = 100	.08	.06	.11
χ^2_4	n = 20	.07	.06	.08
	n = 100	.21	.12	.40

VII. CONCLUSIONS

Clearly, there has been a good deal of attention centred on chi-square goodness-of-fit techniques, and interest in the area is continuing. As long as the versatility and ease of computation are retained, developing improvements to Pearson's test statistic X^2 is a worthwhile endeavour.

The review of some of the modern methods indicates the areas into which interest has evolved, namely in overcoming the handicaps of Pearson's test by producing alternate quadratic forms, by adapting a chi-square test to other than full sample data where competitive procedures are unavailable, and by tailoring a chi-square test to, for example, a multivariate distribution where, again, other tests are not applicable. There is a wide area open for future study.

The results of the previous sections indicate that implementation of the Rao-Robson (or Dzhaparidze-Nikulin in cases where RR is undefined) statistic in favour of X^2 is recommended. The obvious advantage is that the asymptotic distribution is exactly that of chi-square, and the convergence of the finite n percentage points is rapid enough to justify its use; particularly for small values of k, there is potentially large error inherent in the X^2 test when the MDL estimates are used. The disadvantage of the new test procedures is the necessity of deriving the particular form of the statistic for each hypothesized distribution. This, as shown, is usually not difficult. For some specific distributions, such as the Beta distribution, the integrals involved must be solved numerically for the Rao-Robson statistic, but in many cases, the derivation is quite simple.

As far as the general application of chi-square tests is concerned, it is still recommended that, where the conditions of the test allow, more powerful

statistics such as the EDF statistics be used. To emphasize a frequently made point, the basis for preference of a chi-square test remains to be its wide-spread applicability and ease of calculation.

A P P E N D I X

i. Subroutine RRNORM

```
C
C     THIS SUBROUTINE TAKES THE INPUT DATA X AND CALCULATES THE
C     RAO ROBSON STATISTIC RRSTAT.  ASSUMING WE WISH TO TEST FOR
C     NORMALITY, N IS THE NO. OF DATA POINTS AND K IS THE NUMBER
C     OF CELLS.
C
      SUBROUTINE RRNORM(X,N,K,RRSTAT)
      INTEGER OBS(K),NUM(K)
      DIMENSION C(K),D(K,2),BOUND(K)
      DIMENSION X(N),DIFF(N),A(2,2),V(K,2)
C
C     REQUIRE ONE X(I) IN ASCENDING ORDER
C
      CALL VSRTA (X,N)
C
C     CALCULATE THE PARAMETER ESTIMATES
C
      SUM1=0.0
      XN=N
      DO 1 I=1,N
    1 SUM1=SUM1+X(I)
      TH1=SUM1/XN
      SUM2=0.0
      DO 2 I=1,N
      DIFF(I)=X(I)-TH1
    2 SUM2=SUM2+(DIFF(I)*DIFF(I))
      TH2=SQRT(SUM2/(XN-1.0))
C
C
C     THE K+1 CLASS BOUNDARIES ARE OF THE FORM TH1+C(I)TH2.
C     WE CAN OBTAIN THE C(I)'S EASILY FROM THE STANDARD
C     NORMAL TABLE.  THE FOLLOWING ARE GIVEN FOR K=10.
C
C
C     CZERO=-999999.0
C     C(10)=-CZERO
C     C(1)=-1.28
C     C(9)=-C(1)
C     C(2)=-.84
C     C(8)=-C(2)
C     C(3)=-.52
C     C(7)=-C(3)
C     C(4)=-.255
C     C(6)=-C(4)
C     C(5)=0.0
C
C
C     FOR K=4, THE C(I)'S ARE:
C
      CZERO=-999999.0
      C(1)=-.675
      C(3)=-C(1)
      C(2)=0.0
      C(4)=-CZERO
```

```
C
C     CALCULATE THE BOUNDARIES.
C
      BDZERO=TH1+(CZERO*TH2)
      DO 3 I=1,K
    3 BOUND(I)=TH1+(C(I)*TH2)
C
C     TO DETERMINE THE NUMBER OF OBSERVATIONS PER CELL, FIRST CALCULATE
C     THE NUMBER OF OBSERVATIONS LESS THAN OR EQUAL TO BOUND(I),I=1,M
C
      IS=1
      DO 4 J=1,K
      TEMP=BOUND(J)
      DO 5 I=IS,N
      IF(X(I).GT.TEMP) GO TO 50
    5 CONTINUE
      IF((I.EQ.N).AND.(X(I).LT.TEMP)) GO TO 51
   50 NUM(J)=I-1
      IS=I
      GO TO 4
   51 NUM(J)=N
      IS=N
    4 CONTINUE
C
C     NOW CALCULATE THE NUMBER OF OBSERVATIONS PER CELL.
C
      OBS(1)=NUM(1)
      DO 6 I=2,K
    6 OBS(I)=NUM(I)-NUM(I-1)
C
C     DETERMINE THE D(I)'S, V(I)'S, AND A(I,J)'S REQUIRED TO CALCULATE
C     THE TEST STATISTIC.
C
      XN=N
      XK=K
      D(1,1)=0.0
      Q1=SQRT(2.0*3.1416)
      DO 7 I=2,K
      Q2=-(C(I-1)*C(I-1)/2.0)
      Q3=-(C(I)*C(I)/2.0)
      D(I,1)=(1.0/(TH2*(Q1)))*(EXP(Q2)-EXP(Q3))
    7 CONTINUE
      D(1,2)=(1.0/(2.0*TH2*TH2*Q1))*(-C(1)*EXP(Q3))
      DO 9 I=2,K
      D(I,2)=(1.0/(2.0*TH2*TH2*Q1))*(C(I-1)*EXP(Q2)-C(I)*EXP(Q3))
    9 CONTINUE
      DO 10 I=1,K
      V(I,1)=TH2*D(I,1)
      v(I,2)=TH2*TH2*D(I,2)
   10 CONTINUE
C
C     THE A(I,J)'S ARE THE ENTRIES OF THE MATRIX (V+ - J)+ WHERE +
C     DENOTES INVERSE.
```

```
C
      SUM3=0.0
      DO 11 I=1,K
   11 SUM3=SUM3+V(I,2)*V(I,2)
      SUM4=0.0
      SUM5=0.0
      DO 12 I=1,K
      SUM4=SUM4+V(I,1)*V(I,2)
      SUM5=SUM5+V(I,1)*V(I,1)
   12 CONTINUE
      Q4=(XK*SUM5-1.0)*(XK*SUM3-2.0)-(XK*XK*SUM4*SUM4)
C
C     READY TO CALCULATE A(I,J)'S.
C
      A(1,1)=-(XK*2.0*SUM3-1.0)/Q4
      A(1,2)=(XK*SUM4)/Q4
      A(2,1)=A(1,2)
      A(2,2)=-(XK*SUM5-1.0)/Q4
C
C     DETERMINE THE SUMS REQUIRED.
C
      SUM6=0.0
      SUM7=0.0
      SUM8=0.0
      XNK=XN/XK
      DO 14 I=1,K
      XOBS=OBS(I)
      SUM6=SUM6+(XOBS-XNK)*(XOBS-XNK)
      SUM7=SUM7+((XOBS-XNK)*V(I,1))
      SUM8=SUM8+((XOBS-XNK)*V(I,2))
   14 CONTINUE
C
      XKN=XK/XN
      Q5=SUM7*SUM7
      Q6=SUM8*SUM8
      Q7=SUM7*SUM8
C
C     READY TO CALCULATE THE RAO-ROBSON STATISTIC.
C
      RRSTAT=XKN*SUM6+XKN*XK*Q5*A(1,1)+2.0*XKN*XK*Q7*A(1,2)+XKN*XK*Q6*A(2,2)
      RETURN
      END
```

ii. Subroutine RREXP

```
C
C     THIS SUBROUTINE TAKES THE INPUT DATA X AND CALCULATES THE
C     RAO ROBSON STATISTIC RRSTAT.  ASSUMING WE WISH TO TEST FOR
C     EXPONENTIALITY, N IS THE NO. OF DATA POINTS AND K IS THE
C     NUMBER OF CELLS.
C
      SUBROUTINE RREXP(X,N,K,RRSTAT)
      INTEGER OBS(K), NUM(K)
      DIMENSION C(K).D(K).V(K),BOUND(K)
      DIMENSION X(N)
C
C     WANT X(I) IN ASCENDING ORDER
C
      CALL VSRTA (X,N)
C
C     CALCULATE THE PARAMETER ESTIMATES
C
      SUM1=0.0
      XN=N
      XK=K
      DO 1 I=1,N
      SUM1=SUM1+X(I)
    1 CONTINUE
      TH1=SUM1/XN
C
C     THE K+1 CLASS BOUNDARIES ARE OF THE FORM XBAR*C(I-1)
C
      CZERO=0.0
      C(K)=999999.0
      KK=K-1
      DO 2 I=1,KK
      XI=I
      C(I)=-ALOG(1.0-(XI/XK))
    2 CONTINUE
C     CALCULATE THE D(I)'S:
      D(1)=(1.0/TH1)*(-C(1)*EXP(-C(1)))
      DO 3 I=2,K
      D(I)=(1.0/TH1)*(C(I-1)*EXP(-C(I-1))-C(I)*EXP(-C(I)))
    3 CONTINUE
C     CALCULATE THE V(I)'S:
      DO 4 I=1,K
      V(I)=TH1*D(I)
    4 CONTINUE
C     DETERMINE THE BOUNDARIES,
      BDZERO=0.0
      DO 5 I=1,K
      BOUND(I)=TH1*C(I)
    5 CONTINUE
C     DETERMINE THE NUMBER OF OBS. LESS THAN BOUND(J)
      IS=1
      DO 6 J=1,K
      TEMP=BOUND(J)
```

```
      DO 7 I=IS,N
      IF(X(I).GT.TEMP) GO TO 50
    7 CONTINUE
      IF((I.EQ.N).AND.(X(I).LT.TEMP)) GO TO 51
   50 NUM(J)=I-1
      IS=I
      GO TO 6
   51 NUM(J)=N
      IS=N
    6 CONTINUE
C     DETERMINE THE NUMBER OF OBS. IN EACH CELL
      OBS(1)=NUM(1)
      DO 8 I=2,K
      OBS(I)=NUM(I)-NUM(I-1)
    8 CONTINUE
C     READY TO CALCULATE THE RAO-ROBSON STATISTIC
      SUM2=0.0
      SUM3=0.0
      SUM4=0.0
      XNK=XN/XK
      DO 9 I=1,K
      XOBS=OBS(I)
      SUM2=SUM2+(XOBS-XNK)*(XOBS-XNK)
      SUM3=SUM3+((XOBS-XNK)*V(I))*((XOBS-XNK)*V(I))
      SUM4=SUM4+V(I)*V(I)
    9 CONTINUE
      XKN=XK/XN
C
      RRSTAT=XKN*SUM2+(XK*XKN)*(SUM3/(1.0-XK*SUM4))
      RETURN
      END
```

iii. Subroutine DNDEXP

```
C
C     THIS SUBROUTINE TAKES THE INPUT DATA X AND CALCULATES THE
C     DZHAPARIDZE-NIKULIN STATISTIC TO TEST FOR DOUBLE EXPONENTIALITY.
C     N IS THE NO. OF DATA POINTS AND K IS THE NUMBER OF CELLS.
C
      SUBROUTINE DNDEXP(X,N,K,DNSTAT)
      DIMENSION DIFF(N),X(N),C(K),D(K),BOUND(K),XOBS(K)
      INTEGER OBS(K),NUM(K)
C
C     K MUST BE CHOSEN AS EVEN FOR THE FOLLOWING:
C
      NU=K/2
C
C     WANT THE DATA POINTS IN ASCENDING ORDER:
C
      CALL VSRTA (X,N)
C
C     NOW CALCULATE THE PARAMETER ESTIMATES:
C
C
C     IF N IS ODD, TH1 IS X(N+1)/2.  OTHERWISE, TH1 IS
C     (X(N/2)+X(N/2+1))/2.
C
      NN=N/2
C     NN=(N+1)/2
      NN1=NN+1
C
C     IF N IS EVEN, USE:
C
      TH1=(X(NN)+X(NN1))/2.0
C
C
C     IF N IS ODD, USE:
C
C     TH1=X(NN)
C
      SUM1=0.0
      DO 1 I=1,N
      DIFF(I)=X(I)-TH1
      IF(DIFF(I).LT.0.0) DIFF(I)=-DIFF(I)
      SUM1=SUM1+DIFF(I)
    1 CONTINUE
      XN=N
      XK=K
      TH2=SUM1/XN
C
C     CELL BOUNDARIES ARE CHOSEN SUCH THAT P(I)=1/K.
C     THE ITH BOUNDARY IS TH1+ OR -C(I)TH2.  NOW FIND THE
C     C(I), SETTING CZERO AND C(NU) SEPARATELY.
C
      CZERO=0.0
      C(NU)=999999.0
```

```
      NUU=NU=1
      XNU=NU
      DO 2 I=1,NUU
      XI=I
      C(I)=-ALOG(1.0-(XI/XNU))
    2 CONTINUE
C
C     NOW CALCULATE THE D(I)'S:
C
      D(1)=-C(1)*EXP(-C(1))
      SUM2=D(1)*D(1)
      DO 3 I=2,NU
      D(I)=C(I-1)*EXP(-C(I-1))-C(I)*EXP(-C(I))
      SUM2=SUM2+D(I)*D(I)
    3 CONTINUE
C
C     NOW CALCULATE THE BOUNDARIES:
C
      BDZERO=-999999.0
      BOUND(NU)=TH1
      DO 4 I=1,NU
    4 BOUND(NU+I)=TH1+C(I)*TH2
      DO 5 I=1,NUU
    5 BOUND(NU-I)=TH1-C(I)*TH2
C
C     NOW CALCULATE THE NO. OF OBSERVATIONS LESS THAN OR
C     EQUAL TO BOUND(J).
C
      IS=1
      DO 6 J=1,K
      TEMP=BOUND(J)
      DO 7 I=IS,N
      IF(X(I).GT.TEMP) GO TO 50
    7 CONTINUE
      IF((I.EQ.N).AND.(X(I).LT.TEMP)) GO TO 51
   50 NUM(J)=I-1
      IS=I
      GO TO 6
   51 NUM(J)=N
      IS=N
    6 CONTINUE
C
C     NOW CALCULATE THE NUMBER OF OBSERVATIONS PER CELL:
C
      OBS(1)=NUM(1)
      DO 8 I=2,K
      OBS(I)=NUM(I)-NUM(I-1)
    8 CONTINUE
C
C     READY TO CALCULATE THE STATISTIC.
C
C     SUM2 IS CALCULATED ABOVE.  SUM3 RUNS FROM 1 TO N.  BUT SUM4 FROM 1
C     TO NU, SO DETERMINED SEPARATELY.
```

```
C
      SUM3=0.0
      SUM4=0.0
      XNK=XN/XK
      DO 9 I=1,K
      XOBS(I)=OBS(I)
      SUM3=SUM3+(XOBS(I)-XNK)*(XOBS(I)-XNK)
    9 CONTINUE
      DO 11 I=1,NU
      XOBS(I)=OBS(I)
      SUM4=SUM4+D(I)*(XOBS(NU+I)+XOBS(NU-I+1))
   11 CONTINUE
      SUM4=SUM4*SUM4
      XKN=XK/XN
C
C     CALCULATE THE D-N STATISTIC.
C
      DNSTAT=XKN*SUM3-(XKN/(2.0*SUM2))*SUM4
      RETURN
      END
```

iv. Subroutine RRCB

```
C
C     THIS SUBROUTINE TAKES THE INPUT DATA X AND CALCULATES THE
C     RAO ROBSON STATISTIC RRSTAT TO TEST FOR THE CIRCULAR
C     BIVARIATE DISTRIBUTION.  N IS THE NO. OF DATA POINTS AND
C     K IS THE NO. OF CELLS.
C
      SUBROUTINE RRCB(X,Y,N,K,RRSTAT)
      INTEGER OBS(K),NUM(K)
      DIMENSION C(K),D(K),V(K),BOUND(K),X(N)
      DIMENSION Y(N),DIFF(2N),RADIUS(N)
C
C     PLACE DATA IN ASCENDING ORDER
C
C
C     CALCULATE THE THREE PARAMETER ESTIMATES.
      XN=N
      XK=K
      SUM1=0.0
      SUM2=0.0
      SUM3=0.0
      SUM4=0.0
      DO 1 I=1,N
      SUM1=SUM1+X(I)
      SUM2=SUM2+Y(I)
    1 CONTINUE
      TH1=SUM1/XN
      TH2=SUM2/XN
      DO 2 I=1,N
      SUM3=SUM3+(X(I)-TH1)*(X(I)-TH1)
      SUM4=SUM4+(Y(I)-TH2)*(Y(I)-TH2)
    2 CONTINUE
      TH3=(SUM3+SUM4)/(2.*XN)
      TH3=SQRT(TH3)
C
C
C     CELLS ARE CENTRED AT (XBAR,YBAR) WITH SUCCESSIVE RADII C(I)TH3.
C     FIRST REQUIRE THE C(I)'S.
C
C
      CZERO=0.0
      KK=K-1
      DO 3 I=1,KK
      XI=I
      C(I)=SQRT(-2.0*ALOG(1.0-(XI/XK)))
    3 CONTINUE
      C(K)=999999.0
      V(1)=-(C(1)*C(1))*(EXP(-.5*(C(1)*C(1))))
      DO 4 I=2,KK
      QTY1=(C(I-1)*C(I-1))*(EXP(-.5*(C(I-1)*C(I-1))))
      QTY2=(C(I)*C(I))*(EXP(-.5*(C(I)*C(I))))
      V(I)=QTY1-QTY2
    4 CONTINUE
```

```
      V(K)=(C(KK)*C(KK))*(EXP(-.5*(C(KK)*C(KK))))
C
C     CALCULATE THE D(I)'S
C
      DO 5 I=1,K
      D(I)=V(I)/2.0
    5 CONTINUE
C
C     FIND THE BOUNDARIES BY DETERMINING THE RADII C(I)TH3.
C
      DO 6 I=1,K
      BOUND(I)=C(I)*TH3
    6 CONTINUE
C
C
C     IF THE DISTANCE FROM (XBAR,YBAR) TO (X(I),Y(I)) IS LESS THAN
C     RADIUS(I) THEN THE NUMBER OF OBS. IN CELL(I) INCREASES BY
C     ONE.  ALTERNATIVELY, WE CAN CALCULATE ALL THE DISTANCES
C     (XBAR,YBAR) TO (X(I),Y(I)) AND THEN FIND THE NUMBER LESS THAN
C     RADIUS(I).
C
C
      DO 7 I=1,N
      DIFF(I)=X(I)=TH1
      NN=N+I
      DIFF(NN)=Y(I)-TH2
      IF(DIFF(I).LT.0.0) DIFF(I)=DIFF(I)*(-1.0)
      IF(DIFF(NN).LT.0.0) DIFF(NN)=DIFF(NN)*(-1.0)
    7 CONTINUE
      NN=2*N
      DO 8 I=1,N
      RADIUS(I)=SQRT((DIFF(I)*DIFF(I))+DIFF(N+I)*DIFF(N+I)))
    8 CONTINUE
      CALL VSRTA (RADIUS,N)
C
C
C     NOW FIND THE NUMBER OF OBSERVATIONS LESS THAN BOUND(I).
C
C
C     FIRST REQUIRE THAT THE DISTANCES FROM XBAR,YBAR BE IN
C     ASCENDING ORDER.
C
C
      IS=1
      DO 9 J=1,K
      TEMP=BOUND(J)
      DO 10 I=IS,N
      IF (RADIUS(I).GT.TEMP) GO TO 50
   10 CONTINUE
      IF((I.EQ.N).AND.(RADIUS(I).LT.TEMP)) GO TO 51
   50 NUM(J)=I-1
      IS=I
```

```
      GO TO 9
   51 NUM(J)=N
      IS=N
    9 CONTINUE
C
C     NOW CAN FIND THE NUMBER OF OBS. PER CELL.
C
      OBS(1)=NUM(1)
      DO 11 I=2,K
      OBS(I)=NUM(I)-NUM(I-1)
   11 CONTINUE
  209 FORMAT (/2X,5I10)
C
C
C     READY TO CALCULATE THE RAO-ROBSON STATISTIC.
C
C
      SUM5=0.0
      SUM6=0.0
      SUM7=0.0
      XNK=XN/XK
      DO 12 I=1,K
      XOBS=OBS(I)
      SUM5=SUM5+(XOBS-XNK)*(XOBS-XNK)
      SUM6=SUM6+XOBS*D(I)
      SUM7=SUM7+(D(I)*D(I))
   12 CONTINUE
      XKN=XK/XN
C
      RRSTAT=XKN*SUM5+(XK*XKN*(SUM6*SUM6))/(1.0-XK*SUM7)
      RETURN
      END
```

v. Subroutine RRLOG

```
C
C     THIS SUBROUTINE TAKES THE INPUT DATA X AND CALCULATES THE
C     RAO ROBSON STATISTIC RRSTAT.  TO TEST FOR THE LOGISTIC
C     DISTRIBUTION N IS THE NO. OF DATA POINTS AND K IS THE NUMBER
C     OF CELLS.
C
      SUBROUTINE RRLOG(X,N.K,RRSTAT)
      DIMENSION C(K),D(K,2),V(K,2),A(2,2),Q(K)
      DIMENSION X(N),Z(N)
      INTEGER OBS(K),NUM(K)
C
C     REQUIRE X(I) IN ASCENDING ORDER:
C
      CALL VSRTA (X,N)
C
C     CALCULATE THE PARAMETER ESTIMATES:
C
C     FIRST CONSIDER THE CASE WHERE BOTH PARAMETERS ARE UNKNOWN.
C     THE MEAN AND STANDARD DEVIATION*PI/SQRT3 WILL BE THE INITIAL
C     USED IN THE ITERATIVE NUMERICAL SOLUTION:
C
      XN=N
      XK=K
      SUM1=0.0
      SUM2=0.0
      DO 1 I=1,N
      SUM1=SUM1+X(I)
    1 CONTINUE
      TH1=SUM1/XN
      DO 2 I=1,N
      SUM2=SUM2+(X(I)-TH1)*(X(I)-TH1)
    2 CONTINUE
      STD=SQRT(SUM2/XN)
      PI=3.14159265
C     TH2=(PI/SQRT(3.0))*STD
      TH2=STD
C
C     BEGIN THE ITERATIVE PROCESS TO FIND TH1,TH2:
C
      COUNT=0.0
   51 COUNT=COUNT+1.0
      CALL FUNS(X,N,TH1,TH2,FXY,GXY)
      CALL DERIV(X,N,TH1,TH2,FPX,FPY,GPX,GPY)
      XS=((GXY*FPY)-(FXY*GPY))/((FPX*GPY)-(FPY*GPX))
      YS=((GXY*FPX)-(FXY*FPX))/((FPY*GPX)-(FPX*GPY))
      TH1S=TH1+XS
      TH2S=TH2+YS
      DIF1=ABS(XS)
      DIF2=ABS(YS)
      IF (COUNT.GT.30.) GO TO 52
      IF ((DIF1.LT..0001).AND.(DIF2.LT..0001)) GO TO 50
      TH1=TH1S
```

```
      TH2=TH2S
      GO TO 51
   52 WRITE (6,294)
  294 FORMAT(//2X,'ITERATION DOES NOT CONVERGE')
C
   50 CONTINUE
C
C     THE FINAL PARAMETER ESTIMATES ARE:
C
      TH2=(SQRT(3.0)/PI)*TH2
C
C
C     TRANSFORM THE 'X' POINTS TO STANDARD 'Z' POINTS:
C
      DO 3 I=1,N
      Z(I)=(X(I)-TH1)/TH2
    3 CONTINUE
C
C     PREPARE TO CALCULATE THE STATISTIC.  FIRST REQUIRE THE
C     STARDARDIZED BOUNDARIES, C(I)'S:
C
      CZERO=-999999.0
      KM1=K-1
      DO 91 I=1,KM1
      XK=K
      XI=I
      C(I)=-ALOG((XK/XI)-1.0)
   91 CONTINUE
      C(K)=999999.0
C
C     TO DETERMINE THE NUMBER OF OBSERVATIONS PER CELL, FIRST
C     CALCULATE NUMBER OF OBSERVATIONS LT C(I)
C
      IS=1
      DO 4 J=1,K
      TEMP=C(J)
      DO 5 I=IS,N
      IF(Z(I).GT.TEMP) GO TO 96
    5 CONTINUE
      IF((I.EQ.N).AND.(Z(I).LT.TEMP)) GO TO 95
   96 NUM(J)=I-1
      IS=I
      GO TO 4
   95 NUM(J)=N
      IS=N
    4 CONTINUE
C
C     CALCULATE NUMBER OF OBSERVATIONS PER CELL
C
      OBS(1)=NUM(1)
      DO 6 I=2,K
    6 OBS(I)=NUM(I)-NUM(I-1)
```

```
C
C     DETERMINE D(I,J)'S (HAVE STANDARD Z POINTS AND PARAMETERS TH1, TH2)
C
      DO 7 I=1,KM1
      Q(I)=EXP(-C(I))/((1.0+EXP(-C(I)))**2)
    7 CONTINUE
      KM1=K-1
      D(1,1)=-Q(1)
      D(1,2)=-C(1)*Q(1)
      D(K,1)=Q(KM1)
      D(K,2)=C(KM1)*Q(KM1)
      DO 77 I=2,KM1
      D(I,1)=(Q(I-1)-Q(I))
      D(I,2)=-C(I)*Q(I)+C(I-1)*Q(I-1)
   77 CONTINUE
C
C     DETERMINE A(I,J)'S
C
      SUM1=0.0
      SUM2=0.0
      SUM3=0.0
      DO 8 I=1,K
      SUM1=SUM1 + D(I,1)*D(I,1)
      SUM2=SUM2 + D(I,2)*D(I,2)
      SUM3=SUM3 + D(I,1)*D(I,2)
    8 CONTINUE
      PI=3.1415926536
      G1=(3.0+PI**2)/9.0
      DD=(G1-XK*SUM2)*(1.0/3.0-XK*SUM1)-((XK*SUM3)**2)
      CC=1.0D0/DD
C
      A(1,1)=CC*(G1-XK*SUM2)
      A(1,2)=CC*XK*SUM3
      A(2,1)=A(1,2)
      A(2,2)=CC*(1.0/3.0-XK*SUM1)
C
C     READY TO CALCULATE RRSTAT
C
      SUM4=0.0
      SUM5=0.0
      SUM6=0.0
      XNK=XN/XK
      XKN=XK/XN
      DO 9 I=1,K
      XOBS=OBS(I)
      SUM4=SUM4 + (XOBS-XNK)*(XOBS-XNK)
      SUM5=SUM5 + (XOBS-XNK)*D(I,1)
      SUM6=SUM6 + (XOBS-XNK)*D(I,2)
    9 CONTINUE
      Q5=SUM5*SUM5
      Q6=SUM6*SUM6
      Q7=SUM5*SUM6
```

```
C
C
C     SUBROUTINE TO CALCULATE THE VALUES OF F(A,B) AND G(A,B):
C     SUBROUTINE FUNS(X,N,A,B,FXY,GXY)
      DIMENSION X(N)
      XN=N
      SUM1=0.0
      SUM2=0.0
      PI=3.14159265
      DO 100 I=1,N
      C=EXP((PI*(X(I)-A))/(B*SQRT(3.)))
      D=1./(1.+C)
      SUM1=SUM1+D
      E=(1.-C)/(1.+C)
      F=(X(I)-A)/B
      D2=E*F
      SUM2=SUM2+D2
  100 CONTINUE
      FXY=SUM1/XN
      GXY=SUM2/XN
      FXY=FXY-.5
      GXY=FXY+(SQRT(3.)/PI)
      RETURN
      END
C
C     SUBROUTINE TO CALCULATE THE DERIVATIVES OF F(A,B) AND G(A,B):
C     SUBROUTINE DERIV(X,N,A,B,FPX,FPY,GPX,GPY)
      DIMENSION X(N)
      PI=3.14159265
      CON=PI/(SQRT(3.)*B)
      XN=N
      SUM1=0.0
      SUM2=0.0
      SUM1A=0.0
      SUM2A=0.0
      DO 100 I=1,N
      W=DON*(X(I)-A)
      EW=EXP(W)
      SUM1=SUM1+EW/(1.+EW)**2
      DWDB=DON*((A/B)-(X(I)/B))
      SUM1A=SUM1A+(EW/(1.+EW)**2)*DWDB
      P1=W-W*EW
      P2=1./(1.+EW)**2
      P3=1.-W*EW-EW
      P4=1./(1.+EW)
      PROD=-1.*P1*P2*EW+P3*P4
      SUM2=SUM2+PROD
      SUM2A=SUM2A+(PROD*DWDB)
  100 CONTINUE
      FPX=(CON/XN)*SUM1
      FPY=SUM1A/XN*(-1.)
      GPX=SUM2*(-1./(B*XN))
```

```
GPY=SUM2A*(SQRT(3.)/(XN*PI))
RETURN
END
```

vi. Subroutine RREVD

```
C
C     THIS SUBROUTINE TAKES THE INPUT DATA X AND CALCULATES THE
C     RAO ROBSON STATISTIC RRSTAT.  TO TEST FOR THE EXTREME VALUE
C     DISTRIBUTION N IS THE NO. OF DATA POINTS AND K IS THE NUMBER
C     OF CELLS.
C
      SUBROUTINE RREVD(X,N,K,RRSTAT,PEAR,TH1,TH2)
      DIMENSION X(N),XX(N),Z(N)
      DIMENSION C(K),D(K,2),A(2,2)
      DIMENSION Q(K)
      INTEGER OBS(K),NUM(K)
C
C     REQUIRE X(I) IN ASCENDING ORDER
C
      CALL VSRTA (X,N)
C
C     CALCULATE THE PARAMETER ESTIMATES.
C
      VAR=0.0
      SUM=0.0
      XN=N
      DO 33 I=1,N
      SUM=SUM + X(I)
   33 CONTINUE
      TH1=SUM/XN
      DO 22 I=1,N
      XX(I)=X(I)-TH1
      VAR=VAR+(XX(I)*XX(I))
   22 CONTINUE
      STD=SQRT(VAR/(XN-1.0))
      TH2=STD/1.3
      CALL THETA(X,N,TH2)
      CALL ZI(X,N,TH2,TH1)
C
C     STANDARDIZE THE X VALUES: Z=(X-TH1)/TH2
C
      DO 1 I=1,N
      Z(I)=(X(I)-TH1)/TH2
    1 CONTINUE
C
C     REQUIRE STANDARDIZED CELL BOUNDARIES C(I)'S
C
      CZERO=-999999.0
      KM1=K-1
      DO 2 I=1,KM1
      XI=I
      XK=K
      C(I)+-ALOG(-ALOG(XI/XK))
    2 CONTINUE
      C(K)=999999.0
C
```

```
C     TO DETERMINE THE NUMBER OF OBSERVATIONS PER CELL, FIRST
C     TO CALCULATE THE NUMBER OF OBSERVATION LT C(I)
C
      IS=1
      DO 4 J=1,K
      TEMP=C(J)
      DO 5 I=IS,N
      IF(Z(I).GT.TEMP) GO TO 50
   5  CONTINUE
      IF((I.EQ.N).AND.(Z(I).LT.TEMP)) GO TO 51
  50  NUM(J)=I-1
      IS=I
      GO TO 4
  51  NUM(J)=N
      IS=N
   4  CONTINUE
C
C     NOW CALCULATE THE NUMBER OF OBSERVATIONS PER CELL
C
      OBS(1)=NUM(1)
      DO 6 I=2,K
   6  OBS(I)=NUM(I)-NUM(I-1)
C
C     DETERMINE D(I,J)'S.  HAVE STANDARDIZED DATA POINTS AND PARAMETERS
C     TH1 AND TH2.
C
      Q(K)=0.0
      DO 7 I=1,KM1
      Q(I)=EXP(-C(I)-EXP(-C(I)))
   7  CONTINUE
      D(1,1)=-Q(1)
      D(1,2)=-C(1)*Q(1)
      D(K,1)=Q(KM1)
      D(K,2)=C(KM1)*Q(KM1)
      DO 77 I=2,KM1
      D(I,1)=Q(I-1)-Q(I)
      D(I,2)=C(I-1)*Q(I-1)-C(I)*Q(I)
  77  CONTINUE
C
C     PREPARE TO CALCULATE A(I,J)'S:
C
      EC=.577215664
      PI=3.1415926536
      SUM1=0.0
      SUM2=0.0
      SUM3=0.0
      DO 8 I=1,K
      SUM1=SUM1 + D(I,1)*D(I,1)
      SUM2=SUM2 + D(I,2)*D(I,2)
      SUM3=SUM3 + D(I,1)*D(I,2)
   8  CONTINUE
      C1=PI*PI/6.0
```

```
      C2=1.0-EC
      C3=C1+C2*C2
      T1=C3-XK*SUM2
      T2=1.0-XK*SUM1
      T3=C2+XK*SUM3
      DD=T1*T2-(T3*T3)
      CC=1.0D0/DD
C
C     CALCULATE A(I,J)'S:
C
      A(1,1)=T1*CC
      A(1,2)=T3*CC
      A(2,1)=A(1,2)
      A(2,2)=T2*CC
C
C     PREPARE TO CALCULATE THE INPUTS FOR RRSTAT
C
      SUM4=0.0
      SUM5=0.0
      SUM6=0.0
      XNK=XN/XK
      XKN=XK/XN
      DO 9 I=1,K
      XOBS=OBS(I)
      SUM4=SUM4 + (XOBS-XNK)*(XOBS-XNK)
      SUM5=SUM5 + ((XOBS-XNK)*D(I,1))
      SUM6=SUM6 + ((XOBS-XNK)*D(I,2))
    9 CONTINUE
      Q5=SUM5*SUM5
      Q6=SUM6*SUM6
      Q7=SUM5*SUM6
C
C     READY TO CALCULATE RRSTAT:
C
      RRSTAT=XKN*SUM4+XKN*XK*Q5*A(1,1)+2.0*XK*XKN*Q7*A(1,2)+XKN*XK*Q6*A(2,2)
      RETURN
      END
C
      SUBROUTINE THETA (X,N,T)
      DIMENSION X(N)
C     WRITE(6,100)
  100 FORMAT(7X,'T',11X,'T1',5X,'COUNT')
C     WRITE(6,200)
      S1=0.0
      COUNT=0.0
      IF(T.GT.1.8) T=T/1.5
      DO 10 I=1,N
      S1=S1+X(I)
   10 CONTINUE
      S1=S1/N
   15 S2=0.0
      S3=0.0
```

```
      COUNT=COUNT+1.0
      DO 20 I=1,N
      E=EXP(-X(I)/T)
      S2=S2+X(I)*E
      S3=S3+E
   20 CONTINUE
      T1=S1-S2/S3
C     WRITE (6,200) T,T1,COUNT
      T1=(T+T1*2.0)/3.0
C     FORMAT (2X,3F12.6)
      Z=ABS(T-T1)
      T=T1
      IF (COUNT.GT.20.0) GO TO 31
      IF (Z.LT.0.001) GO TO 30
      GO TO 15
   30 CONTINUE
C     WRITE (6,300) T
  300 FORMAT(/2X,'T =',F12.6)
      RETURN
   31 T=1.0
      WRITE (6,700)
  700 FORMAT (/2X,'ITERATION DID NOT CONVERGE')
      RETURN
      END
C
      SUBROUTINE ZI(X,N,T,Z)
      DIMENSION X(N)
      S=0.0
      DO 10 I=1,N
      S=S+EXP(-X(I)/T)
   10 CONTINUE
      Z=-T*ALOG(S/N)
C     WRITE (6,100) Z
  100 FORMAT(2X,'ZI=',F12.6)
      RETURN
      END
```

BIBLIOGRAPHY

Andrews, D.F., Gnanadesikan, R. and Warner, J.L. (1973). "Methods for Assessing Multivariate Normality" in Multivariate Analysis III (P.R. Krishnaiah, Editor). New York: Academic Press, 95-116.

Barankin, E.D. and Gurland, J. (1951). "On Asymptotically Normal Efficient Estimators." Univ. Calif. Publ. Stat 1, 89-129.

Chernoff, H. and Lehmann, E.L. (1954). "The Use of Maximum-Likelihood Estimates in Chi-Square Test of Goodness of Fit." The Annals of Mathematical Statistics, 25, 579-586.

Cochran, William G. (1952). "The Chi-Square Test of Goodness of Fit." The Annals of Mathematical Statistics, 23, 315-345.

Cochran, William G. (1954). "Some Methods for Strengthening the Common Chi-Square Tests." Biometrics, 12, 417-451.

Dahiya, Ram. C. and Gurland, John. (1972a). "Pearson Chi-Square Test of Fit with Random Intervals." Biometrika, 59, 147-153.

Dahiya, Ram. C. and Gurland, John. (1972b). "A Test of Fit for Continuous Distributions Based on Generalized Minimum Chi-Square." in Statistical Papers in Honor of George W. Snedecor (T.A. Bancroft, editor). The Iowa State University Press, 115-128.

Dahiya, Ram. C. and Gurland, John. (1973). "How Many Classes in the Pearson Chi-Square Test?" Journal of the American Statistical Association, 707-712.

Davis, A.W. (1977). "A Differential Equation Approach to Linear Combinations of Independent Chi-Squares." Journal of the American Statistical Association, 72, 212-214.

Dzhaparidze, K.O. and Nikulin, M.S. (1974). "On a Modification of the Standard Statistics of Pearson." Theor. Probability Appl., 19, 851-853.

Gasser, T. (1975). "Goodness-of-Fit Tests for Correlated Data." Biometrika, 62, 563-570.

Gurland, J. (1948). "Best Asymptotically Normal Estimates." Unpublished PhD Thesis, Univ. Calif., Berkeley.

Johnson, N.L. and Kotz, S. (1970). Distributions in Statistics: Continuous Univariate Distributions, Volume I. Boston: Houghton Mifflin.

Johnson, N. L. and Kotz, S. (1970). Distributions in Statistics: Continuous Univariate Distributions, Volume 2. Boston: Houghton Mifflin.

Kempthorne, O. (1968). "The Classical Problem of Inference: Goodness of Fit." Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press.

Kendall, Maurice G. and Stuart, Alan. (1969). The Advanced Theory of Statistics, Volume I, Distribution Theory. New York: Hafner Publishing

Kendall, Maurice G. and Stuart, Alan. (1963). The Advanced Theory of Statistics, Volume II, Inference and Relationship. New York: Hafner Publishing Company.

Mann, H.B. and Wald, A. (1942). "On the Choice of the Number of Class Intervals in the Application of the Chi-Square Test." Annals of Mathematical Statistics, 13, 306-317.

Mihalko, Daniel P. and Moore, David S. (1980). "Chi-Square Tests of Fit for Type II Censored Data." The Annals of Statistics, Vol. 8, No. 3, 635-644.

Moore, David S. (1971). " A Chi-Square Statistic with Random Cell Boundaries." The Annals of Mathematical Statistics, Vol. 42, No. 1, 147-256.

Moore, David S. (1976). "Recent Developments in Chi-Square Tests for Goodness of Fit." Department of Statistics, Division of Mathematical Sciences, Purdue University, Mimeograph Series #459.

Moore, David S. (1977a). "Generalized Inverses, Wald's Method, and the Construction of Chi-Squared Tests of Fit." Journal of the American Statistical Association, Volume 72, Number 357, 131-137.

Moore, David S. (1977b). "The Effect of Dependence on Chi-Square Tests of Fit." Department of Statistics, Division of Mathematical Sciences, Purdue University, Mimeograph Series #505.

Moore, David S. (N.D.) "Chi-Square Techniques." Chapter Three (unpublished).

Moore, David S. and Spruill, M.C. (1975). "Unified Large-Sample Theory of General Chi-Squared Statistics for Tests of Fit." The Annals of Statistics, Vol. 3, No. 3, 599-616.

Moore, David s. and Stubblebine, John D. (1981). "Chi-Square Tests for Multivariate Normality with Application to Common Stock Prices." Commun. Statist.-Theor. Meth., A10(8), 713-738.

Morris, Carl. (1975). "Central Limit Theorems for Multinomial Sums." The Annals of Mathematical Statistics, Vol. 3, No. 1, 165-188.

Rao, K.C. and Robson, Douglas S. (1974). "A Chi-Square Statistic for Goodness-of-Fit Tests with the Exponential Family." Communications in Statistics, 3(12), 1139-1153.

Roscoe, J.T. and Byars, J.A. (1971). "The Investigation of the Restraints with Respect to a Sample Size Commonly Imposed on the Use of the Chi-Square Statistic." Journal of the American Statistical Association, 66, 755-759.

Spinelli, J. (1981). "Regression and EDF Tests of Fit." Unpublished M.Sc. Thesis, Simon Fraser University.

Stephens, M.A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons." Journal of the American Statistical Association, Volume 69, Number 347, 730-737.

Stephens, M.A. (1977). "Goodness of Fit for the Extreme Value Distribution." Biometrika, 64, 3, 583-588.

Stephens, M.A. (1979). "Tests of Fit for the Logistic Distribution Based on the Empirical Distribution Function." Biometrika, 66, 3, 591-595.

Watson, Geoffrey S. (1958). "On Chi-Square Goodness-of-Fit Tests for Continuous Distributions." Roy. Statist. Soc. Ser. B, 20, 44-61.

Watson, Geoffrey S. (1959). "Some Recent Results in Chi-Square Goodness-of-Fit Tests." Biometrics, 15, 440-468.