1994

The goodness of fit publications of J.C.W. Rayner

J. C.W. Rayner

University of Wollongong

Recommended Citation

NOTE

This online version of the thesis may have different page formatting and pagination from the paper copy held in the University of Wollongong Library.

UNIVERSITY OF WOLLONGONG

COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.
The Goodness of Fit Publications

of

J.C.W. Rayner

Submitted in fulfilment of the requirements for the award of the degree

DOCTOR OF PHILOSOPHY

from the

University of Wollongong

Department of Applied Statistics

1994
Declaration

None of the published works in the list following, the Goodness of Fit Publications of J.C.W. Rayner, have been submitted for a degree or diploma or other qualification at any other institution.

J.C.W. Rayner

February, 1994
Foreword

This thesis has two parts. The first is an overview, “The Development of the Smooth Tests of Goodness of Fit”, of the submitted work. This overview is relatively brief, being approximately 5000 words long. The second part consists of the submitted publications themselves. These are listed below by year of publication, and alphabetically within year.

The bulk of the thesis has necessitated it being bound in two volumes. The first volume consists of the overview and publications [1] to [18]. The second volume consists of publications [19] to [25].

The Goodness of Fit Publications of J.C.W. Rayner


The Goodness of Fit Publications

of

J.C.W. Rayner

Volume 1
The Development of the Smooth Tests of Goodness of Fit

1. The Content

The following summary focuses on goodness of fit research, done mainly in conjunction with John Best. Subsequently by "we", I mean John Rayner and John Best. In as much as it is possible to divide our contributions, the theoretical derivations have been primarily done by me, and John Best has provided the bulk of the applications and the programming. In all our joint papers, the ordering of the authors is significant. The first author contributed most to the conception of the project and wrote it up.

Our early work considered mainly the properties of Pearson's $X^2$ goodness of fit test. The later work developed "smooth tests" that are available for any distribution, that are convenient to use, and have excellent power. That material is reviewed in [21]. The remainder of this section gives a brief introduction to goodness of fit testing, Pearson's $X^2$ test, and Neyman's $\Psi_k^2$ smooth test. A full account of the majority of this material is contained in [19]. It is timely to extract from there a description of the goodness of fit problem and its outcomes.

"What then is a goodness of fit test? According to David (1966, p.399)

'A goodness of fit procedure is a statistical test of a hypothesis that the sampled population is distributed in a specific way ... for example, that the sampled population is normal.'

This is the one sample problem; the corresponding S-sample problem assesses whether or not S independent random samples come from the same population.

Subsequently we shall mainly be concerned with one sample tests for goodness of fit. Formally, given a random sample $X_1, X_2, \ldots, X_n$, we test the null hypothesis that the sampled
population has cumulative distribution function $F(x; \theta), \theta \in \Theta$, against the alternative hypothesis that the cumulative distribution function is $G(x; \omega), \omega \in \Omega$. All of $X, \Theta$ and $\Omega$ may be multidimensional. Frequently the alternative is simply 'not the null hypothesis'.

What do we get from having applied a goodness of fit test? Firstly a compact description of the data. Saying that the data are binomial with parameters $n = 15$ and $p = 0.51$ is a valuable abbreviation of the available information. Secondly, powerful parametric procedures, such as the tests in the analysis of variance, are valid if the data are consistent with normality. And thirdly, light may be shed on the mechanisms generating the data. For example, if the data cannot be viewed as a Poisson process, then at least one of the axioms sufficient for a Poisson process has failed. If lifetimes for cancer patients from the onset of 'standard' treatment have been exponentially distributed with mean 36 months in the past, and this distribution no longer holds under a new treatment, what has changed? It could be that either the mean, or the distribution has changed. In the latter case perhaps treatment is less effective than the standard treatment for some, and they die sooner than under the standard treatment; and the treatment is apparently effective for others, who survive longer than previously."

Perhaps the most widely known test in statistical inference and certainly the first goodness of fit test, is Pearson's $X^2_p$ test. It is not widely known that Pearson's test is a smooth test, but this was demonstrated in [16] and [18, Theorem 5.1.2]. An informal definition of $X^2_p$ follows. Suppose observations may fall into $m$ non-overlapping classes or cells. We hypothesise the cells should contain respectively $E_1, \ldots, E_m$ observations, but the observed cell counts are $O_1, \ldots, O_m$. Now define the Pearson test statistic by

$$X^2_p = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j}.$$
If this is larger than the $100\alpha\%$ point of the $\chi^2_{k-1}$ distribution then the hypothesised expectations can be rejected at the $100\alpha\%$ level of significance. Pearson's test is applicable for testing discrete data when there are no parameters that need to be estimated. The expansion of the methodology to cover more practical situations has occupied statisticians almost continuously since Karl Pearson introduced this test in Pearson (1900).

The concept of smooth goodness of fit tests was introduced in Neyman (1937). Neyman's $\Psi^2_k$ smooth test was developed to overcome presumed deficiencies in Pearson's $X^2_k$ test. The test was called "smooth" because it was constructed to have good power against alternatives whose probability density functions depart "smoothly" from that specified by the null hypothesis. For example the null hypothesis may specify the normal distribution with zero mean and unit variance, while the alternative may specify the normal distribution with unit mean and unit variance. Smooth changes include shifts in mean, variance, skewness and kurtosis. See [19, Figures 1.5 and 2.3], in which normal probability density functions and smooth alternatives to normality are fitted to data.

Suppose we have a random sample from a continuous distribution with completely specified cumulative distribution function $F(x)$. Applying the probability integral transformation, the null hypothesis $H_0$ specifies that $Y = F(X)$ is uniformly distributed on $(0, 1)$. Neyman's smooth alternative of order $k$ to the uniform distribution on $(0, 1)$ has probability density function

$$h_k(y, \theta) = \exp\left(\sum_{i=1}^{k} \theta_i \pi_i(y) - K(\theta)\right), 0 < y < 1,$$

in which $k = 1, 2, \ldots, \theta^T = (\theta_1, \ldots, \theta_k)$, $K(\theta)$ is a normalising constant, and the $\pi_i(y)$ are orthonormal polynomials related to the Legendre polynomials. The first five such polynomials are:

- $\pi_0(y) = 1$, $\pi_1(y) = \sqrt{3} (2y - 1)$, $\pi_2(y) = \sqrt{5} (6y^2 - 6y + 1)$,
- $\pi_3(y) = \sqrt{7} (20y^3 - 30y^2 + 12y - 1)$, and
- $\pi_4(y) = 3(70y^4 - 140y^3 + 90y^2 - 20y + 1)$.
The \( \pi_i(y) \) are constructed so that \( \pi_i(y) \) is of degree \( i \) and the \( \pi_i(y) \) are orthonormal using the uniform \((0, 1)\) weight function. To test the null hypothesis \( H_0: \theta_1 = \ldots = \theta_k = 0 \), we use the Neyman statistic, given by

\[
\Psi_k^2 = \sum_{i=1}^{k} U_i^2 \text{ in which } U_i = \sum_{j=1}^{n} \pi_i(Y_j)/\sqrt{n}.
\]

The \( U_i \) are the components into which \( \Psi_k^2 \) is partitioned.

Neyman's conception for his smooth test was that it should be constructed to be locally most powerful, unbiased and of size \( \alpha \) for testing for uniformity against the order \( k \) alternatives given by \( h_k(y, \theta) \). Its power function was also constrained to be symmetric, depending on \( \theta \) only through \( \theta_1^2 + \ldots + \theta_k^2 \). Neyman (1937) noted that his solution is only approximate; only asymptotically is the test of size \( \alpha \), unbiased, and most powerful.

2. Early Papers: Properties of Pearson's Test

The Honours class in Mathematical Statistics at Sydney University in 1967 had a very strong flavour of what became Lancaster (1969), and therefore demanded some facility with orthonormal functions. During that year I developed an interest in goodness of fit in general and Pearson's \( X^2 \) test in particular. Subsequently I read the following passage in Kendall and Stuart (1967, p. 436), relating the power \( P \) of the equiprobable \( X^2 \) goodness of fit test to its size \( \alpha \).

"Thus \( P \geq \alpha \) near \( H_0 \) and the equiprobable test is locally unbiased, which is a recommendation of this class-formation procedure, since no such result is known to hold for the \( X^2 \) test in general."

I began to explore the unbiasedness of the \( X^2 \) test in general, and took sabbatical leave from May 1979 to May 1980, principally to work with John Best, who had also been a student of Lancaster. This was the beginning of
a long and fruitful collaboration. About this time we wrote two criticisms, [1] and [3], and laid the foundations for much subsequent work.

We showed in [5] that there are unbiased $X^2$ tests apart from the equiprobable test. But there is no consistent power advantage in constructing such tests, and it is certainly inconvenient to do so. Of course the fact that a test is biased is not necessarily damning. In [9] a brief review of some of the literature concerning bias and the $X^2$ test, now resolved to our satisfaction, was given.

The problem of class construction for the $X^2$ test is classically allied with the choice of the number of classes. Consideration of the number of equiprobable classes was given briefly in [5], and in more detail later in [10]. The result, that two classes may be more powerful than more classes, was so surprising to us that we wrote a more discursive paper, [2], mainly to publicise the result. However that paper also included a conjecture as to why fewer classes may produce more power. In fact if testing for uniformity when the alternative is exponential ($\lambda$), it is interesting to construct a sequence of $X^2$ tests based on $k$ classes and to calculate the power for each. Typically the power is either greatest for just two classes, or increases from two classes to $k_0$ classes (where typically $k_0$ is small, such as two or four) and then declines monotonically. Hence using a large number of classes, as for example Mann and Wald (1942) recommended, can lead to a considerable power loss.

In [4] we began a consideration of the components into which test statistics of interest may be conveniently partitioned. This approach permeated much of our subsequent goodness of fit work. We reconsidered Lancaster’s (1953) partitioning of Pearson’s equiprobable $X^2$ test. Its components are asymptotically independent and asymptotically standard normally distributed. The $r$th component detects a discrepancy between the $r$th moment of the data and that of the hypothesized normal distribution. While the $X^2$ test is an omnibus test, with moderate power in all directions in the parameter space, its components are directional, typically with good power only in only one direction. This distinction is worth making because the $X^2$ test had frequently been unfairly compared
in power studies with essentially directional tests. Fairer comparisons would involve either comparing the $X^p$ test with other omnibus tests, or comparing the components of the $X^p$ test with other directional tests. See for example Quesenberry and Miller (1977), who compare the omnibus $X^p$ tests with 10 and 20 classes with directional goodness of fit tests. In subsequent power studies, such as that in [5], we made the point that competitor tests such as those based on the empirical distribution function are neither completely directional nor truly omnibus. It would seem to be more helpful to apply an omnibus test, such as the $X^p$ test, and use the components in a data analytic fashion. The major difficulty with the $X^p$ test is that a large number of classes would seem to be needed to avoid loss of information due to overcategorisation. The test is then very omnibus, in that detects very general alternatives but with only moderate to weak power. On the other hand, the first few components of an $X^p$ test with many classes, combined through their sum of squares, provide an excellent omnibus test with good power against alternatives likely to occur in practice and little loss of information by overcategorisation. These ideas were further developed in [5] and [7], and subsequent papers.

In [7] we applied our knowledge of components and related ideas to the choice of the number of classes in the application of $X^p$ and Neyman's $\Psi^2_k$ (and in fact other tests also). Alternative hypotheses can be considered in terms of the Fourier coefficients of their probability density functions with respect to a given orthonormal system. Many alternatives of interest in the literature are well described by relatively few coefficients, and hence can be considered to have low order: $k$ "small" in the definition of $h_k(y, \theta)$. Low order alternatives are optimally detected by the sum of the squares of relatively few components. (The meaning of "optimal" and, subsequently in this summary, "weakly optimal," is discussed in [19, section 3.4].) This suggested the notion of the effective order of the alternative. An alternative has effective order $k_0$, roughly, if greatest power of the Neyman smooth tests using $k$ components, results from using $k = k_0$ components. This notion is very robust to the choice of $X^p$ or $\Psi^2_k$ for the test detecting the alternative, and to the test size used. So the choice of the number of classes $m$ for $X^p$ and components $k$ for $\Psi^2_k$ depends on the alternatives one wishes to best detect. In fact much practical data can be best described by distributions involving relatively few parameters,
and hence require low order appropriately chosen tests to best detect departures from the hypothesised distribution.

Papers [6], [10] and [11] all built on student projects. It is fair to say that the students contribution was mainly in the computing. For [6] and [10] John Best checked the calculations and made other contributions to the methodology. Size and power studies are an integral part of our approach. In [6] we assessed calculating size and power of the $X^2$ test by several techniques including simulation, an Edgeworth series approximation and a three moment chi-squared approximation. Exact calculation is prohibitive as either the number of multinomial cells or sample size increases. We recommended simulation and the three moment chi-squared approximation.

In [11] score tests were developed to assess the improvement of the fit of certain generalised tests over their non-generalised counterparts. Specifically, the quasi-binomial and generalised Poisson, gamma and negative binomial tests were considered. The tests developed essentially answer the question, "do the extra parameters of the generalized distribution significantly improve the fit of the data to the model?" An extension of this approach was extensively used in the development of the smooth tests.

[13] is a review, primarily of the Pearson $X^2$ tests and its generalisations and their properties.

3. The Smooth Tests of Goodness of Fit

At this point the inter-relationship of several factors led to a formulation of a wide class of tests. It was apparent from the work in [7] that several test statistics have similar structure in terms of their components. We had examined Lancaster's test of normality in [8]. It is defined as the sum of the squares of components that asymptotically are independent with the standard normal distribution. A class of tests with this structure would be very useful and convenient. We then became aware of Kopecky and Pierce (1979) and Thomas and Pierce (1979). They had generalized Neyman (1937) to permit nuisance parameters, and hence the cases of most importance in applications.
A probability density function $f_X(x; \beta)$, such as the normal with unspecified mean and variance, can be generalised to a smooth alternative of order $k$ as follows. If $\beta$ is a vector of $q$ say nuisance parameters, $f_X(x; \beta)$ is embedded in an order $k$ alternative

$$g_k(x; \theta, \beta) = C(\theta, \beta) \exp\left(\sum_{i=1}^{k} \theta_i h_i(x; \beta)\right) f_X(x; \beta).$$

Here $C(\theta, \beta)$ is a normalising constant ensuring the probability density function $g_k(x; \theta, \beta)$ integrates to one. We test $H: \theta = (\theta_i) = 0$ against $K: \theta \neq 0$ to answer the question "does the smooth alternative of order $k$ significantly improve the fit of the data to the hypothesized model $f_X(x; \beta)$?" As in [11] we do this optimally by using the score statistic.

Neyman (1937) took $f_X(x; \beta)$ to be uniform on $(0, 1)$ and the $\{h_i(x; \beta)\}$ to be the Legendre polynomials. Pierce and his co-workers preferred to take

$$h_i(x; \beta) = F^i(x; \beta),$$

the $i$th power of the cumulative distribution function. Their test statistics require the calculation of tables of constants and their components are not asymptotically independent. A considerable gain in convenience is achieved if we take the $\{h_i(x; \beta)\}$ to be orthonormal on $f_X(x; \beta)$, although other choices can lead to useful and powerful tests. Choosing orthonormal functions leads to the structure desired.

In [14] and [18] a general structure for smooth tests of goodness of fit is given. Smooth tests of goodness of fit based on orthonormal functions for location-scale families were introduced in the former. The latter extends this class of tests to 'regular' families. The extension preserves the desirable properties of the class, such as weak optimality, accessible components and convenient distribution theory.

In [14] we introduced a class of smooth goodness of fit tests that can test for any location-scale family of distributions provided a set of complete
orthonormal functions on the standardised distribution is available. Since functions orthonormal on the uniform may be transformed to functions orthonormal on the hypothesised distribution, there are usually many orthonormal systems from which to choose. In fact, the system may be chosen to give good power against specified alternatives. The test statistics have a simple form and convenient asymptotic distribution, and since the statistics are score statistics the tests are weakly optimal. Moreover, asymptotically independent components of the test statistic can be calculated routinely to assess particular aspects of the fit. Applications include Lancaster’s test discussed in [8], and tests for exponentiality discussed in [14].

The experience of [8] and [14] was that in the particular cases examined there, these smooth tests are, in terms of their simulated powers, very competitive with existing tests of fit. The aim of [18] was to extend the families for which such smooth goodness of fit tests exist, from location-scale families to families whose probability density functions (or probability functions) depend on nuisance parameters and satisfy mild regularity conditions. This meant it was now possible to analyse in detail tests for such distributions as the gamma, the Poisson and the binomial, that were inaccessible using the previous theory. Some detail about these papers is now given since it is central to the exposition.

Consider a random sample $X_1, \ldots, X_n$ from the distribution with probability density function $g_k(x; \theta, \beta)$. Subsequently by "regular" we mean that derivatives of the logarithm of the likelihood with respect to elements of $\theta$ and $\beta$ exist up to second order, as do their expectations. Furthermore, we assume that the integral over the reals of $g_k(x; \theta, \beta)$ may be differentiated under the integral; such differentiations with respect to the $\theta_i$ are permitted by Lehmann (1986, Theorem 9, p.59). Analogously, in the discrete case, differentiation and summation may be exchanged. In specific cases these assumptions must be checked.

In all cases the technique is to calculate the score statistic for the smooth model, or a simple transformation of it. The modification enables us to handle covariance matrices of less than full rank and also generalisations such as focusing and overlapping that are discussed below. Typically the
sum of squares of the components provides an omnibus test that is weakly optimal for \( g_k(x; \theta, \beta) \), while the \( r \)th component is a directional test and weakly optimal for detecting

\[
C_r(\theta, \beta) \exp\{\theta_r h_r(x; \beta)\} f_X(x; \beta),
\]

in which \( C_r(\theta, \beta) \) is a normalizing constant ensuring the probability density function integrates to one. That the components are asymptotically independent follows ultimately from the orthonormality, while the use of score statistics leads to asymptotic normality of the components and to the omnibus statistic having an asymptotic \( \chi^2 \) distribution.

Theorem 3.1 of [18] gives a flavour of the technical results. The score statistic for testing \( \theta = 0 \) with the regular model \( g_k(x; \theta, \beta) \) is

\[
S(\hat{\beta}) = \left( \sum_{j=1}^{n} h(X_j ; \hat{\beta}) / \sqrt{n} \right)^T \hat{M}^{-1} \left( \sum_{j=1}^{n} h(X_j ; \hat{\beta}) / \sqrt{n} \right),
\]

in which

\[
M = I_k - \text{cov}_{\theta_0} \left( \frac{\partial \log f_X}{\partial \theta} \right) \left( \text{cov}_{\theta_0} \left( \frac{\partial \log f_X}{\partial \theta}, \frac{\partial \log f_X}{\partial \theta} \right) \right)^{-1} \text{cov}_{\theta_0} \left( \frac{\partial \log f_X}{\partial \theta}, h \right),
\]

with \( h = (h_i(x; \beta)) \), provided \( \hat{M} \) is non-singular. The hats indicate that the maximum likelihood estimator \( \hat{\beta} \) has replaced \( \beta \). The covariances are evaluated under the null hypothesis \( \theta = 0 \).

A simple structure for the test statistic results when \( \hat{\beta} \) is a linear combination at most the first \( k \) \( h_i(x; \beta) \).

When the general theory is applied, it is desirable to assess the adequacy of the asymptotic null distributions of the test statistics and their components. This usually leads to so-called "Bartlett corrections", adjustments to the critical points of the test statistics so that the actual size is closer to the nominal size. It is also useful to assess the power of the
smooth tests in comparison with the known and often entrenched tests. In some of our papers we have only given the test statistics and left the size-power study for other workers and/or other times. So to a greater or lesser extent the distributions we have considered are the discrete ([15]), and continuous ([7]) uniform, the univariate ([8] and [18]), bivariate ([17]) and multivariate ([18]) normal, the exponential ([14]), binomial ([18]), gamma ([18]), univariate Poisson ([18]) and geometric ([18] and [20]). In an as yet unpublished paper, the bivariate Poisson has been examined. In all the power studies we have conducted, the smooth tests have been competitive with existing tests.

In many cases the first one or two components of our test statistics are already known and widely used goodness of fit tests. The first two components of the smooth test for the normal are standardised versions of the skewness and kurtosis statistics. For the exponential our first component is a linear translation of the statistic given by Greenwood (1946). The binomial and Poisson first components are standardised versions of R.A. Fisher's indices of dispersion, while the first component of the smooth test for the geometric is clearly related to Vit's (1974) test. The higher order components are therefore analogues of these entrenched tests, and usually surprising ones at that; for these entrenched tests were in almost all cases proposed on grounds quite different from the smooth approach.

Like [13], [21] is a review. It focused on the Rayner and Best smooth work, most of which was published after [13]. The intention was to stimulate interest in [19], that unified our previous goodness of fit research. In fact the book also stimulated more research. The structure desired for the book was centred on a double dichotomy: smooth tests for categorized and uncategorised data with null probability density functions both with and without nuisance parameters. We found that more work needed to be done on the categorized models, and in particular when there are nuisance parameters present. This led to [16] and [22].

In the categorized situation the underlying distribution could be continuous, depending on a vector \( \beta \) of unspecified parameters. For example, the distribution could be normal, depending on the unknown mean and variance. Observations are categorised into \( m \) known classes.
The cell probabilities \( p_j = p_j(\beta), j = 1, \ldots, m \), are found by integrating the probability density function over the classes. Alternatively the model could be genuinely discrete or categorised. In [22], as an alternative to the null hypothesis, we took the \( m \) cell probabilities to be

\[
\pi_j(\beta) = C(\theta, \beta) \exp \left( \sum_{i=1}^{k} h_{ij}(\beta) \right) p_j(\beta), j = 1, \ldots, m.
\]

Here \( \beta = (\beta_1, \ldots, \beta_Q)^T \) is a vector of nuisance parameters, \( \theta = (\theta_1, \ldots, \theta_K)^T \) is a vector of real parameters and \( C(\theta, \beta) \) is a normalising constant ensuring that \( \sum_{j=1}^{m} \pi_j(\beta) = 1 \). The \( h_{ij}(\beta) \) are functions of \( \beta \) to be specified by the statistician. Different choices of these functions emphasize different alternatives and result in different tests. The selection of these functions is analogous to the choice of \( \{h_i(x; \beta)\} \) in the uncategorised smooth models. Again the technique used is to calculate the score statistic for the smooth model, or a simple transformation of it. As before, the modification enables us to handle covariance matrices of less than full rank and other generalisations.

The model corresponding to \( \pi_j(\beta) \) but without nuisance parameters was discussed in [16]. A particular case of the full rank score statistic is Pearson’s \( \chi^2 \). This formulation enables a variety of different components to be calculated. Use of the Chebyshev orthogonal polynomials results in the \( rth \) component reflecting deviations of the data from the \( rth \) moment of the hypothesised distribution. Another useful particular case is a generalisation of Hall’s (1985) idea of overlapping. If cell expectations are low, the asymptotic \( \chi^2 \) distribution will not be a good approximation. Pooling loses information. Overlapping combines cells with low expectation, possibly with cells with high expectation, until the combined cell expectation is at least “moderate”. By overlapping a particular cell in different ways, the information in that cell can be retained. Simulations suggest that this technique yields test statistics for which the \( \chi^2 \) approximation is generally adequate. Another subclass of tests in this formulation enables focusing on particular cells, to assess, for example, if the data disagree with the hypothesized model in particular cells, such as those in the tail.
In [22] the results of [16] are generalised to the composite situation. The analogue of $X^2_F$ is $X^2_F$, which is shown to be a smooth test. Components of $X^2_F$ were calculated, the $r$th component $\hat{V}_r$ being a contrast in the differences observed minus expected for each cell, standardised to be asymptotically independent and standard normal. Such components are not unique. Those $\hat{V}_r$ constructed involved the first $r + q + 1$ cells only, with $r = 1, \ldots, m - q - 1$. If such a statistic were judged to be significantly large, the interpretation would be that the $r$th cell differs from its predecessors in the comparison of the cell expectations and observations. In the sense that other cells could have been focused on, the construction is not unique.

In [19, section 7.4] a comparison of the Pearson-Fisher $X^2_F$ and Rao-Robson $X^2_R$ tests was given. The latter is a smooth test, and can be derived from the class of uncategorised smooth tests by using indicator functions. In as much as components can be derived for all tests in this class, they can be derived for the $X^2_R$ test. But it is argued in [19, section 7.4] that the $X^2_R$ tests can rarely be appropriate. They require uncategorised maximum likelihood estimators, and if these are available, the uncategorised smooth tests should be available, and these will be more convenient and frequently more powerful. In any case the hypothesis that for uncategorised data, the $X^2_R$ tests are superior to the $X^2_F$ tests, is not sustainable.

An essential part of the Rayner and Best work has been its applicability. In [23] and [24] medical and ecological applications are discussed. In the latter, in the same way that contrasts are used in the analysis of variance, components may be used with Pearson's goodness of fit test. The use of these components to make the Pearson test more informative was demonstrated.

Finally in [25] S-sample smooth goodness of fit tests are introduced. These tests are omnibus tests and have directional components, all based on score tests which thus are weakly optimal for specifiable alternatives. Testing may proceed on a hierarchical basis. We test if the first two samples are consistent with each other, and then if the $s$th, $s = 3, \ldots, S$, sample is consistent with its predecessors. The test statistics are given in
terms of the one sample components for each sample. Then, if all samples are consistent with having come from the same population, we can conveniently perform a one-sample test for that distribution. If the samples are not consistent, an LSD-type analysis can be performed on the one-sample components to identify where the differences occur. Our procedure requires the specification of a target distribution. The subjectivity of the choice of this distribution is balanced by the wealth of information the procedure provides.

While the theoretical derivations have been primarily done by me, John Best has provided most of the applications and the programming. In [12] the programmes to implement our smooth tests for the uniform, normal and exponential distributions are given. These programmes are extended to other distributions in an appendix in [19].

Although relatively recent, the smooth approach to assessing models is having a noticeable effect on the literature. Kallenberg et al. (1985), Kallenberg (1985) and Oosterhoff (1985) all discussed the choice of the number of classes for the $X^2$ test, building on and greatly extending our early work. Hutchinson and Lai (1990) reported mainly on [17]. Daniel (1990) referred to the smooth tests for both univariate and bivariate normality: [8] and [17]. The school of goodness of fit testing based on the empirical distribution function (EDF) has yet to focus on testing for discrete distributions, something that is done routinely by the smooth approach. Research in this area is currently being done by John Spinelli and Michael Stephens. In private correspondence between John Best and John Spinelli, the latter gave power comparisons between some EDF and smooth tests for the binomial and Poisson distributions. The smooth tests are so widely applicable that proposers of new tests almost necessarily must give comparisons between their new tests and the smooth competitors. Bickel and Ritov (1992) studied "a new approach ... analytically". This approach is essentially that given in [19]. LaRiccia (1992) also gave a new approach: he modelled quantiles smoothly and mirrored the approach taken in [14] to obtain a new class of tests for location-scale families. I would expect this class to be competitive with those introduced by John Best and myself; neither could be expected to be uniformly superior.
Although the published material covers many pages, the questions of interest have not all been answered. Work is in progress on what happens when a smooth test is used to reject the null hypothesis. Contrary to one of the classical criticisms of goodness of fit testing, an alternative model is suggested. The bivariate Poisson work already alluded to involves new theory as well as new smooth tests. An algorithm for constructing an orthonormal system to most powerfully detect departure from the null hypothesis is being assessed. As well as further goodness of fit work, John Best and I are engaged in applying our smooth methods to contingency table analysis and to constructing partially parametric tests. These tests adjust for failure of the parametric model when some of the components are significantly large, but take advantage of the fact that typically several components are not significant, so that the original parametric model is partially acceptable.
References


Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philos. Mag.* 5th ser., 50, 157-175.


A Note on Mineo's Grouping Method for the Chi-Square Test of Goodness of Fit

D. J. BEST and J. C. W. RAYNER

DMS, CSIRO, North Ryde, Australia and University of Otago, New Zealand

Received October 1980, in final form February 1981

ABSTRACT. This note comments on some claims and computations given by Mineo (1979) in support of a new method of grouping for the classical Chi-Square test of goodness of fit.

Key words: chi-square test, goodness of fit, grouping, Monte Carlo simulation

Are Two Classes Enough for the $X^2$ Goodness of Fit Test?

by D.J. Best* and J.C.W. Rayner**

Summary The chi-squared goodness of fit test (also known as the PEARSON $X^2$ test) is often used to test whether data are consistent with a specified continuous distribution. The value of $X^2$ (and hence its associated probability level) can be altered by the choice of (i) the number of classes and (ii) the class probabilities. The effect on the power of the $X^2$ test of varying the number of classes when class probabilities are chosen to be equal is investigated.

On Hühn's Notion of Optimal Sample Size in the $\chi^2$ Goodness of Fit Test

J. C. W. Rayner and D. J. Best

Abstract

Hühn's notion of optimal sample size for the $\chi^2$ goodness of fit test is discussed. It involves minimizing the achievable level of the test subject to a given nominal level. This ignores the power of the competing tests and can lead to an 'optimal' test that is uniformly less powerful than all the competing tests considered.

Key words: Optimal sample size, $\chi^2$ goodness of fit test.
PARTITIONING THE EQUIPROBABLE $X^2$ STATISTIC FOR TESTING UNIFORMITY

D. J. Best
CSIRO Division of Mathematics and Statistics, North Ryde, NSW, Australia

J. C. W. Rayner
Mathematics Department, University of Otago, New Zealand

1. INTRODUCTION
The equiprobable $X^2$ test is often used to test for uniformity. Suppose that the $n$ observed points have been partitioned into $k$ classes, chosen to be equiprobable under the uniformity hypothesis. If $X_i, i = 1(1)k$, is the number of points falling into the $i$th class, the familiar $X^2$ statistic is

$$X^2 = \sum_{i=1}^{k} \frac{(X_i - n/k)^2}{(n/k)} = (k/n) \sum_{i=1}^{k} X_i^2 - n.$$  

This statistic is also, of course, used to test for the multinomial distribution with equal probabilities (the symmetric multinomial). Note, too, the well known result (see, for example, Kendall and Stuart (1973, section 30.36)) that by use of the probability integral transformation a test of goodness of fit for any continuous distribution with known parameters can be converted to a test for uniformity. Other interesting but not so well known results are outlined (i) in Quesenberry et al (1976) where testing for normality of several samples is converted to a test for uniformity, and (ii) in Michael and Schucany (1979) where testing for goodness of fit for censored samples is also converted to a test for uniformity.

A drawback in the use of $X^2$ is that it is an omnibus test. Thus, if an appropriate critical value is exceeded there is no indication of the type of alternative distribution which may hold and further, the $X^2$ test may not be powerful for some alternatives. A solution to this problem is the partition of $X^2$ for the symmetric multinomial suggested by Lancaster (1953). Partition or component analysis has also been advocated by other writers for other goodness of fit tests (see, for example, the introduction of Durbin and Knott (1972)).

This note will give an example of the computation involved in obtaining the partition of $X^2$ and illustrate how the components so derived supplement the $X^2$ test for uniformity.

2. PARTITIONING $X^2$
Suppose now that an $X^2$ test has been applied. If the test statistic is significant we would often like to know more about the nature of the departure from the null hypothesis. If the test statistic is not significant it may be that $k$ has not been chosen well. For example, Best and Rayner (1981) show that smaller $k$ may provide a more powerful $X^2$ test than larger $k$.

To deal with these problems consider the partition of $X^2$ suggested by Lancaster (1953) for testing the symmetric multinomial.

Define

$$X^2 = \sum_{i=1}^{k-1} U_i^2.$$  

29
where

\[ U_i = \sum_{j=0}^{k-1} f(j) X_{i-j} / \sqrt{(n_s)} \]

in which

\[ n = \sum_{i=1}^{k} X_i, \quad \text{and} \quad s_i = \sum_{j=0}^{k-1} f_i^2(j)/k. \]

The \( f_i \) are Chebyshev orthogonal polynomials of degree \( i \) with discrete range which may be defined by the recurrence relation

\[ f_{i+1}(j) = f_i(j) f_i(j) - \alpha_i f_{i-1}(j), \quad f_0(j) = 1, \quad f_1(j) = j - (k - 1)/2 \]

where \( \alpha_i = i^2(k^2 - i^2)/(4i^2 - 1) \). Tables of the \( f_i(j) \) and \( \sum_{j=0}^{k-1} f_i^2(j) \) (apart from scaling factors) are given in Fisher and Yates (Table XXIII, 1963) or Pearson and Hartley (Table 47, 1970). These tables are convenient for hand calculation; for computer work the recurrence relation given above may be used.

It is easily shown that the \( U_i \) are uncorrelated standardised (\( E(U_i) = 0, \ E(U_i^2) = 1 \)) variables which are asymptotically independent standard normal. Computer simulation studies, which are to be reported elsewhere, indicate (i) that the normal approximation is quite reasonable for \( n \) as small as 20 and (ii) that a significant \( U_i \) value can indicate deviations from uniformity in the \( i \)th moment (note that Durbin and Knott (p. 302, 1972) found their first two components detected location and scale differences respectively). Thus significance of a particular \( U_i \) gives a clue to the type of non-uniformity. It is possible for a \( U_i \) to be highly significant when \( \chi^2 \) itself is not significant. Significant \( U_i \) values indicate departures in mean (Lancaster (1953)).

Often the first three or four components make up most of the \( \chi^2 \) value and higher components will not need to be computed. The value of \( \chi^2 - \sum_{i=1}^{4} U_i^2 \) which is approximately distributed as \( \chi^2 \) with \( k - 5 \) degrees of freedom would make this clear.

3. EXAMPLES

Example 1. Katti (p. 435, 1973) refers to an example where crabs are free to enter any of 6 chambers and it is required to test the null hypothesis that the decision to enter a given chamber is made randomly. The frequencies of entry in an experiment using 16 crabs were 3, 2, 5, 2, 1, 3 for which \( \chi^2 = 3.5 \). The \( U_i \) are \(-.44, -.07, +.87, +1.27, -.96\). Combined with the non significant \( \chi^2 \) value these values give little evidence for rejection of the null hypothesis.

Example 2. Suppose we wish to make a preliminary check on a newly-proposed algorithm for the computer generation of pseudo-random uniform (0.1) values. The algorithm generated the following 20 values (rounded to 2 decimal places):

\[ .31, .27, .86, .06, .08, .86, .02, .38, .12, .11, .36, .31, .41, .01, .21, .23, .34, .48, .74, .29. \]

After grouping these into \( k = 4 \) classes which would be equiprobable if the null uniformity hypothesis were true, one obtains \( X_1 = 8, X_2 = 9, X_3 = 1, X_4 = 2 \) and thus \( \chi^2 = 10.00 \). Using the tables of Katti (1973) we see that \( \chi^2 \) exceeds the exact upper 5% critical value, which is 8.00. The
2.5% critical value of 9.348 from the usual approximating \( X^2 \) distribution with 3 degrees of freedom is also exceeded. Table 1 illustrates the calculation of \( U_1, U_2 \) and \( U_3 \).

### Table 1: Calculation of \( U_i \) for Example 2

<table>
<thead>
<tr>
<th>( j )</th>
<th>( x )</th>
<th>( f_1 )</th>
<th>( f_2 )</th>
<th>( f_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
<td>-2</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>-1</td>
<td>-1</td>
<td>+3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>+1</td>
<td>-1</td>
<td>-3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>+3</td>
<td>+1</td>
<td>+1</td>
</tr>
</tbody>
</table>

\[
\sum x = 20 \\
\sum f_1 x = -26 \\
\sum f_2 x = 0 \\
\sum f_3 x = 18
\]

\[
\begin{align*}
s_1 &= 5 \\
s_2 &= 1 \\
s_3 &= 5
\end{align*}
\]

\[
\begin{align*}
U_1 &= -2.60 \\
U_2 &= 0.00 \\
U_3 &= +1.80
\end{align*}
\]

The \( f_j(x) \) and \( \sum_{j=0}^{k-1} f_j^2(x) \) values used here can be read from either of the two books of tables referenced above. The large negative value of \( U_1 \)—remember that the \( U_i \) are approximately standard normal—indicates the mean may be less than expected under the null uniformity hypothesis. This is in agreement with the sample histogram although in other examples the \( U_i \) values may highlight non-uniformity not so evident from the histogram.

Example 3. Suppose the data of example 2 are grouped into 10 classes instead of 4. Then the class frequencies are 4,2,4,5,2,0,1,2,0 and \( X^2 \) is 15.00. The 5% critical value of \( X^2 \) with 9 degrees of freedom is 16.919 and Katti’s (1973) exact tables show that \( \Pr(X^2 > 15.0) = 0.0965 \). Even though \( X^2 \) might not now be judged significant it is instructive to calculate the first 5 \( U_i \) which are -2.6469, 0.2462, 0.8855, -0.6611, -2.3799. \( U_1 \) is still rather low indicating, as before, a mean less than expected. On the basis of this preliminary or exploratory analysis on a sample of size 20 one might be inclined to test a sample of larger size with directional tests for degree 1 and 5 alternatives—one such possibility would be to use \( U_1^2 + U_2^2 \) and consider this to have a \( X^2 \) distribution with 2 degrees of freedom.

### 4. SUMMARY

Two common difficulties with the \( X^2 \) test are firstly, that if the test is significant there may be no indication of the type of alternative causing significance; and secondly, the test may have low power for many particular alternatives. To help remedy these difficulties \( X^2 \) may be partitioned and its components examined.

Having partitioned \( X^2 \), one must be careful in interpreting the results. There is a danger that effectively many tests are being applied simultaneously, and false significance levels may be inferred. Notwithstanding, the suggested decomposition of \( X^2 \) is a useful technique that deserves to be more widely known.

### 5. REFERENCES


THE CHOICE OF CLASS PROBABILITIES AND NUMBER OF CLASSES FOR THE SIMPLE $\chi^2$ GOODNESS OF FIT TEST

By J. C. W. RAYNER

University of Otago, New Zealand

and

D. J. BEST

CSIRO, Australia

SUMMARY. Stephens (1974), Quesenberry and Miller (1977), and others make power comparisons between their favoured goodness of fit tests and the classical equiprobable $\chi^2$ test. This symmetric test has been recommended for many years as the $\chi^2$ test to use when testing goodness of fit of continuous distributions. Cohen and Sackrowitz (1975) showed this test to be unbiased and type D for testing multinomial probabilities. We show that there are other unbiased $\chi^2$ tests; further, it is shown by example, that for testing goodness of fit the type D property is lost. As regards the number of classes we show by example that just two classes may give the greatest power. In contrast to many previous authors we give exact, rather than approximate, power calculations.
Summary:

A number of approximations for the power of Pearson's chi-squared test are considered. A three-moment $\chi^2$ approximation is shown to behave well.

Keywords: Power, Monte Carlo, Non-central chi-squared, Normal approximation, Goodness of fit test.

1. Introduction

Suppose there are $n$ independent observations of a random variable $Z$ and that the sample space of $Z$ has been partitioned into $k$ classes so that the probability of an observation lying in the $i$th class is $\pi_i$. The well known Pearson or chi-squared goodness of fit statistic for testing the simple null hypothesis $H_0$: $\pi_i=p_i$, $i=1, ..., k$ is

$$X^2 = \sum (x_i - np_i)^2/(np_i)$$

where $x_i$ is the number of the $n$ observations falling in the $i$th class.

The power function of the test based on $X^2$ is useful for a number of reasons including, (i) comparisons, like those of Stephens (1974), of the $X^2$ test with other tests for $H_0$, (ii) determining the optimum partition of $Z$ and the optimum choice of $k$, as in Best and Rayner (1981), and (iii) calculating the sample size needed to produce a given power for a specified alternative. However, the complexity of the power function prohibits its exact computation for all but small $n$ and $k$. Power approximations are thus necessary.

Broffitt and Randles (1977) compare the non-central $\chi^2$ approximation suggested by Eisenhart (1938) with a normal approximation. West (1971) and Prosini (1976) examine gamma approximations. Our contribution will be to develop and evaluate two new approximations based on an Edgeworth series and a three-moment $\chi^2$ approximation. These approximations will use the moments of $X^2$ for general alternatives given by Kochler (1979).
2. Existing Methods of Power Calculation

A number of methods have been suggested for calculating the power, \( \beta \), of the \( \chi^2 \) test. These include the exact method, the non-central \( \chi^2 \) approximation, Monte Carlo approximation, the normal approximation and gamma approximations.

The exact method will be considered first. The \( n \) observations may be allocated to \( k \) classes in \( \frac{(n+k-1)!}{(n!)(k-1)!} \) ways (see, for example, Riordan (1958), p. 7). The probability of obtaining a given allocation \( x_1, x_2, \ldots, x_k \) is

\[
P(x) = n! \prod_{i=1}^{k} \left( \frac{\pi_i}{\pi_i} \right)^{x_i}.
\]

To actually perform the calculation of \( P(x) \) it is best to first compute the natural logarithm of \( P(x) \) and then exponentiate. This avoids overflow problems and can be done efficiently by storing the natural logarithms up to \( n \) and the natural logarithms of the \( \pi_i \) in arrays. The exact power, \( \beta \), is then just \( \beta = \sum P(x) \) where the summation is over all partitions such that \( \chi^2 > C \) where \( C \) is a critical value such that \( \beta = 0.05 \) when \( n = 1 \), \( n = 2, \ldots \). Unfortunately, \( N \) increases rapidly with \( n \) or \( k \). For \( n=10, k=2, N=11 \) while for \( n=20, k=4, N=1,771 \). All \( N \) partitions can be generated using routines such as \( \text{NEXCOM} \) (Nijenhuis and Wilf (1975)).

Note that this exact method is in agreement with the "Exact \( \chi^2 \) test" method of Radlow and Alf (1975) but not in agreement with the method of a number of authors including Neyman and Pearson (1931) or, more recently, Tate and Hyer (1973).

The Monte Carlo method was used by Slakter (1968) to obtain powers and involves obtaining random samples from the alternative distribution. This involves generating random samples from a multinomial distribution with parameters \( \pi_1, \ldots, \pi_k \) and \( n \). This is easily done on a computer. Let \( u_1, \ldots, u_n \) be \( n \) uniform \((0,1)\) pseudo-random numbers. For each \( u_i \) find which of the \( k \) intervals \([0, \pi_{i-1}), [\pi_{i-1}, \pi_i), \ldots \) covers \( u_i \) and add one to the count for that interval. The vector of counts formed after all the \( u_i \) have been considered is the required random sample. Suppose \( N \) such samples are generated and \( M \) of them give \( \chi^2 > C \) then \( \beta \approx M/N \) and using standard binomial theory the standard error of this estimate is approximately \( \sqrt{(M - M^2/N)} \).

Two other power approximations are based on the non-central \( \chi^2 \) and normal distributions. If tables of these distributions are available approximate powers may be found without recourse to a computer. However, these two approximations can be considerably in error. Let \( \chi^2 \{k, \lambda \} \) be a non-central \( \chi^2 \) random variable with \( k \) degrees of freedom and non-centrality parameter \( \lambda \). Further, let \( \Phi \) denote the cumulative distribution function of the standard normal distribution. The non-central \( \chi^2 \) approximation and the normal approximation are then

\[
\beta = \Pr(\chi^2 \{k-1, n \sum (\pi_i - \pi_{i-1})^2 / \pi_{i-1} \} > C) \quad \text{and} \quad \beta = 1 - \Phi(\{C - \lambda \chi^2 \}/\sqrt{\text{var}(\chi^2)}).
\]
Slakter (1968) compared the non-central $\chi^2$ approximation with Monte Carlo estimates of $\beta$ and found reasonable agreement. Broffitt and Randies (1977) compared the normal and the non-central $\chi$ approximations with Monte Carlo estimates and again found reasonable agreement although the normal approximation seemed closer to the Monte Carlo estimates for high $\beta$ while the non-central $\chi$ approximation was closer to the Monte Carlo estimates for smaller $\beta$. Best and Rayner (1981) made limited comparisons of the non-central $\chi^2$ approximation and exact powers, concluding "there was a difference of less than 0.06". West (1971) gives quite extensive comparisons for small $n$ and $k$ between the exact power, the non-central $\chi^2$ approximation and two approximations based on the gamma distribution. The better of these is

$$\beta \triangleq \Pr(\chi^2_r > 2\gamma C)$$

where $r/\gamma = E(\chi^2)$, $r/\gamma^2 = \text{var}(\chi^2)$ and $\chi^2_r$ is a central $\chi^2$ random variable with $k$ degrees of freedom. This method will be called West's approximation. Frosini (1976) also looks at this approximation.

3. Two New Power Approximations

(a) A Three-Moment $\chi^2$ Approximation

As the power of $\chi^2$ can be approximated by a non-central $\chi^2$ integral which in turn can be approximated by a three-moment $\chi$ approximation (see Pearson (1959)) it seems reasonable to attempt approximating the power of $\chi^2$ by a $\chi$ distribution. Thus if $C_0 + C_1X^2$ is to be $\chi^2_v$, say, then on equating the first three moments

$$C_1 = 4\mu_2/3, \quad v = C_1/\mu_2, \quad C_0 = v - C_1^2/2$$

where $\mu_1$, $\mu_2$, and $\mu_3$ are the moments of $X^2$ under the alternative hypothesis. Note that $v$ is not always an integer. The first two such moments $\mu_1 = E(X^2)$ and $\mu_2 = \text{var}(X)$ were given by Patnaik (1949) while more recently higher moments including $\mu_3$ were given by Koehler (1979). The three-moment $\chi$ approximation is thus

$$\beta \triangleq \Pr(\chi^2_v > C_0 + C_1X).$$

(b) An Edgeworth Series Approximation

As the power of $\chi^2$ can be approximated by a normal integral (as in Broffitt and Randies (1977)) it seems likely that an Edgeworth approximation of the form

$$\beta \triangleq 1 - \Phi(x) + \Phi'(x)[H_2\gamma_1/6 + H_3\gamma_2/24 + H_5\gamma_1^2/72]$$

may do better. In this approximation $x = (X^2 - \mu_1)/\sqrt{\mu_2}$ while the $H_i$, $\gamma_i$ and $\gamma_1$ are Hermite orthogonal polynomials and the usual standardized skewness and kurtosis measures.
4. Comparisons

The results were obtained by computing the exact and approximate powers of the \( X^2 \) test for a small but hopefully representative range of alternatives to uniformity. In Tables 1, 2, 3 and 4 critical values are taken from the \( X^2 \) tables. All the Monte Carlo power approximations given in this section are based on samples of size 2000.

**Table 1**

Errors in power approximations and exact power, \( \beta \), for alternative \( h(z) = 2 - 2z \) for \( n = 10, 20, 30 \) and \( k = 2, 3, 4, 5, 6 \). The nominal test size is 10%.

(a) \( n = 10 \)

<table>
<thead>
<tr>
<th>( k )</th>
<th>( \beta )</th>
<th>( \Delta_{MC} )</th>
<th>( \Delta_{NC} )</th>
<th>( \Delta_N )</th>
<th>( \Delta_F )</th>
<th>( \Delta_C )</th>
<th>( \Delta_W )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.5260</td>
<td>-0.0065</td>
<td>-0.0508</td>
<td>0.0550</td>
<td>-0.0100</td>
<td>-0.0001</td>
<td>-0.0510</td>
</tr>
<tr>
<td>3</td>
<td>0.4574</td>
<td>-0.0006</td>
<td>-0.0175</td>
<td>0.0504</td>
<td>-0.0279</td>
<td>-0.0279</td>
<td>-0.0398</td>
</tr>
<tr>
<td>4</td>
<td>0.4104</td>
<td>-0.0149</td>
<td>-0.0059</td>
<td>0.0403</td>
<td>-0.0419</td>
<td>-0.0404</td>
<td>-0.0375</td>
</tr>
<tr>
<td>5</td>
<td>0.3546</td>
<td>-0.0151</td>
<td>0.0211</td>
<td>0.0537</td>
<td>-0.0275</td>
<td>-0.0231</td>
<td>-0.0136</td>
</tr>
<tr>
<td>6</td>
<td>0.2893</td>
<td>0.0022</td>
<td>0.0632</td>
<td>0.0868</td>
<td>0.0082</td>
<td>0.0154</td>
<td>-0.0130</td>
</tr>
</tbody>
</table>

max | \( \Delta \) | 0.0151 | 0.0632 | 0.0868 | 0.0419 | 0.0404 | 0.0510 |

av | \( \Delta \) | 0.0079 | 0.0317 | 0.0572 | 0.0231 | 0.0214 | 0.0310 |

(b) \( n = 20 \)

<table>
<thead>
<tr>
<th>( k )</th>
<th>( \beta )</th>
<th>( \Delta_{MC} )</th>
<th>( \Delta_{NC} )</th>
<th>( \Delta_N )</th>
<th>( \Delta_F )</th>
<th>( \Delta_C )</th>
<th>( \Delta_W )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.7858</td>
<td>-0.0003</td>
<td>-0.0570</td>
<td>0.0017</td>
<td>-0.0224</td>
<td>-0.0081</td>
<td>-0.0055</td>
</tr>
<tr>
<td>3</td>
<td>0.7643</td>
<td>0.0122</td>
<td>-0.0676</td>
<td>-0.0048</td>
<td>-0.0249</td>
<td>-0.0261</td>
<td>-0.0281</td>
</tr>
<tr>
<td>4</td>
<td>0.6873</td>
<td>-0.0028</td>
<td>-0.0278</td>
<td>0.0321</td>
<td>-0.0003</td>
<td>-0.0055</td>
<td>0.0009</td>
</tr>
<tr>
<td>5</td>
<td>0.6441</td>
<td>-0.0089</td>
<td>-0.0194</td>
<td>0.0363</td>
<td>-0.0078</td>
<td>-0.0143</td>
<td>-0.0097</td>
</tr>
<tr>
<td>6</td>
<td>0.6030</td>
<td>-0.0015</td>
<td>-0.0091</td>
<td>0.0419</td>
<td>-0.0110</td>
<td>-0.0174</td>
<td>-0.0090</td>
</tr>
</tbody>
</table>

max | \( \Delta \) | 0.0122 | 0.0676 | 0.0419 | 0.0224 | 0.0261 | 0.0280 |

av | \( \Delta \) | 0.0051 | 0.0362 | 0.0234 | 0.0133 | 0.0143 | 0.0106 |

(c) \( n = 30 \)

<table>
<thead>
<tr>
<th>( k )</th>
<th>( \beta )</th>
<th>( \Delta_{MC} )</th>
<th>( \Delta_{NC} )</th>
<th>( \Delta_N )</th>
<th>( \Delta_F )</th>
<th>( \Delta_C )</th>
<th>( \Delta_W )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.8943</td>
<td>-0.0053</td>
<td>-0.0314</td>
<td>-0.0114</td>
<td>-0.0056</td>
<td>0.0026</td>
<td>0.0317</td>
</tr>
<tr>
<td>3</td>
<td>0.8936</td>
<td>0.0039</td>
<td>-0.0432</td>
<td>-0.0156</td>
<td>0.0036</td>
<td>0.0059</td>
<td>0.0136</td>
</tr>
<tr>
<td>4</td>
<td>0.8724</td>
<td>-0.0014</td>
<td>-0.0483</td>
<td>-0.0158</td>
<td>0.0009</td>
<td>0.0012</td>
<td>0.0013</td>
</tr>
<tr>
<td>5</td>
<td>0.8321</td>
<td>-0.0011</td>
<td>-0.0355</td>
<td>-0.0003</td>
<td>0.0092</td>
<td>0.0075</td>
<td>0.0058</td>
</tr>
<tr>
<td>6</td>
<td>0.7957</td>
<td>-0.0032</td>
<td>-0.0255</td>
<td>0.0110</td>
<td>0.0123</td>
<td>0.0086</td>
<td>0.0075</td>
</tr>
</tbody>
</table>

max | \( \Delta \) | 0.0053 | 0.0463 | 0.0158 | 0.0123 | 0.0086 | 0.0317 |

av | \( \Delta \) | 0.0030 | 0.0368 | 0.0108 | 0.0063 | 0.0043 | 0.0120 |

\( \Delta_{MC}, \Delta_{NC}, \Delta_N, \Delta_F, \Delta_C \) and \( \Delta_W \) are the errors for the Monte Carlo results, the non-central \( X^2 \), normal, Edgeworth, three-moment \( X \) and West approximations.

Table 1 shows exact and approximate powers for various values of the parameters involved for the alternative density function \( h(z) = 2 - 2z \). This non-symmetric alternative is a member of

From Table 1 it can be seen that the Monte Carlo power approximations differ from $\beta$ by less than 0.01. Such accuracy was of course expected since with Monte Carlo samples of size 2000 the estimated standard error of the power approximations obtained is less than 0.0112.

For $n=10$ both the normal and the non-central $\chi^2$ approximations give power estimates which commonly differ from $\beta$ by over 0.05. Thus, not unexpectedly, we might conclude that $n=10$ is too small for these power approximations based on large sample theory to give reasonable agreement with $\beta$. When larger values of $n$ are used both methods do on occasions give close agreement with the true power and on others differ markedly from it. The normal approximation appears to improve with $n$ more quickly than the non-central $\chi^2$ approximation but it must be remembered (as shown in Table 2 below and by Broffitt and Randies) that the normal approximation is best for large $\beta$ values. The non-central and normal $\chi^2$ approximations are clearly inferior to the Edgeworth and three-moment $\chi^2$ approximations. These latter approximations improve with $n$ until for $n=30$ they are almost as good as the Monte Carlo method.

Broffitt and Randies (1977) found that for $\beta > 0.7$ the normal approximation was superior to the non-central $\chi^2$ approximation while the reverse was the case for $\beta < 0.7$. To investigate this claim further exact and approximate powers were computed for the following asymmetric distribution (Stephens (1974), uniformity alternative B)

$$f(z) = m(2z)^{m-1} 0 < z < 0.5, \ m > 0$$

$$f(z) = m(2(1-z))^{m-1} 0.5 < z < 1, \ m > 0.$$  

By varying the value of $m$ we can obtain alternatives at various distances from the null distribution. The more distant an alternative is from the null distribution the higher the power will be. Thus by varying $m$ we can examine the accuracy of the normal and non-central $\chi^2$ power approximations to various values of $\beta$. This was done for $n=20$ and $k=4$ with a test size of 10%; the results obtained are shown in Table 2.

From the tabulated errors in Table 2 the following points are apparent. The normal approximation over-estimates $\beta$ markedly for 0.2 < $\beta$ < 0.7. The errors rise steadily as $\beta$ increases, change sign at about $\beta = 0.1$, and peak at around 0.08 when $\beta$ is about 0.4. This pattern has been noticed in nearly all the cases studied. Unpresented results suggest that when $\beta < 0.1$ the normal approximation tends to under-estimate $\beta$. Also the pattern revealed in Table 2 for the non-central $\chi^2$ approximation has been almost universally present in all the cases considered. For high values of $\beta$ the non-central $\chi^2$ approximation tends to under-estimate the true power quite markedly whilst for middle and lower values of $\beta$ it consistently (if only slightly) tends to over-estimate $\beta$. The latter finding gives some support to the conclusion drawn by Slakter (page 917, 1968) that the actual power is usually less than the non-central $\chi^2$ power approximation.
Table 2

Errors in power approximations and exact power, $\beta$, for Stephen's uniformity alternative B for $n = 20$, $k = 4$, nominal size 10% and various $m$.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$\beta$</th>
<th>$\Delta_{MC}$</th>
<th>$\Delta_N$</th>
<th>$\Delta_F$</th>
<th>$\Delta_C$</th>
<th>$\Delta_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.0934</td>
<td>0.0066</td>
<td>-0.0068</td>
<td>-0.0127</td>
<td>0.0032</td>
<td>0.0032</td>
</tr>
<tr>
<td>1.30</td>
<td>0.1444</td>
<td>0.0223</td>
<td>-0.0459</td>
<td>0.0057</td>
<td>0.0146</td>
<td>0.0140</td>
</tr>
<tr>
<td>1.35</td>
<td>0.1760</td>
<td>0.0125</td>
<td>0.0375</td>
<td>-0.0012</td>
<td>0.0038</td>
<td>0.0031</td>
</tr>
<tr>
<td>1.42</td>
<td>0.2085</td>
<td>0.0144</td>
<td>0.0512</td>
<td>0.0020</td>
<td>0.0041</td>
<td>0.0013</td>
</tr>
<tr>
<td>1.49</td>
<td>0.2451</td>
<td>0.0160</td>
<td>0.0632</td>
<td>0.0049</td>
<td>0.0048</td>
<td>0.0001</td>
</tr>
<tr>
<td>1.74</td>
<td>0.3975</td>
<td>0.0160</td>
<td>0.0831</td>
<td>0.0112</td>
<td>0.0110</td>
<td>-0.0024</td>
</tr>
<tr>
<td>1.91</td>
<td>0.5081</td>
<td>0.0089</td>
<td>0.0762</td>
<td>0.0156</td>
<td>0.0162</td>
<td>0.0003</td>
</tr>
<tr>
<td>2.00</td>
<td>0.5652</td>
<td>0.0029</td>
<td>0.0682</td>
<td>0.0176</td>
<td>0.0183</td>
<td>0.0031</td>
</tr>
<tr>
<td>2.20</td>
<td>0.6913</td>
<td>-0.0133</td>
<td>0.0461</td>
<td>0.0201</td>
<td>0.0207</td>
<td>0.0119</td>
</tr>
<tr>
<td>2.40</td>
<td>0.7771</td>
<td>-0.0296</td>
<td>0.0253</td>
<td>0.0200</td>
<td>0.0199</td>
<td>0.0210</td>
</tr>
<tr>
<td>2.50</td>
<td>0.8166</td>
<td>-0.0367</td>
<td>0.0170</td>
<td>0.0191</td>
<td>0.0186</td>
<td>0.0244</td>
</tr>
<tr>
<td>2.60</td>
<td>0.8506</td>
<td>-0.0426</td>
<td>0.0105</td>
<td>0.0179</td>
<td>0.0169</td>
<td>0.0267</td>
</tr>
<tr>
<td>3.20</td>
<td>0.9635</td>
<td>-0.0532</td>
<td>-0.0047</td>
<td>0.0077</td>
<td>0.0077</td>
<td>0.0199</td>
</tr>
</tbody>
</table>

max $|\Delta|$ | 0.0532 | 0.0831 | 0.0201 | 0.0207 | 0.0267

av $|\Delta|$ | 0.0212 | 0.0412 | 0.0119 | 0.0123 | 0.0101

The errors of the Edgeworth, West and three-moment $\chi^2$ approximations are considerably less than those of the other approximations shown and very similar to each other.

Table 3 shows errors for the five approximate methods for some alternatives used by Slakter. The results of Table 3 are in agreement with those of Table 1 and 2, again demonstrating the superiority of the Edgeworth and three-moment $\chi^2$ approximations.

Table 3

Errors in power approximations and exact power, $\beta$, for Slakter's alternatives when $n = 10$, $k = 10$ and nominal size is 0.05.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$\Delta_{MC}$</th>
<th>$\Delta_N$</th>
<th>$\Delta_F$</th>
<th>$\Delta_C$</th>
<th>$\Delta_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0395</td>
<td>-0.0005</td>
<td>0.0105</td>
<td>-0.0149</td>
<td>0.0101</td>
<td>0.0070</td>
</tr>
<tr>
<td>0.0782</td>
<td>0.0008</td>
<td>0.0216</td>
<td>-0.0020</td>
<td>-0.0030</td>
<td>0.0117</td>
</tr>
<tr>
<td>0.1519</td>
<td>0.0141</td>
<td>0.0481</td>
<td>0.0322</td>
<td>-0.0046</td>
<td>0.0135</td>
</tr>
<tr>
<td>0.2288</td>
<td>0.0137</td>
<td>0.0712</td>
<td>0.0566</td>
<td>-0.0014</td>
<td>0.0125</td>
</tr>
<tr>
<td>0.3239</td>
<td>-0.0014</td>
<td>0.0761</td>
<td>0.0840</td>
<td>0.0031</td>
<td>0.0063</td>
</tr>
<tr>
<td>0.3999</td>
<td>-0.0018</td>
<td>0.1001</td>
<td>0.1002</td>
<td>0.0049</td>
<td>0.0050</td>
</tr>
<tr>
<td>0.4914</td>
<td>0.0011</td>
<td>0.1086</td>
<td>0.0877</td>
<td>0.0038</td>
<td>0.0009</td>
</tr>
<tr>
<td>0.5684</td>
<td>-0.0009</td>
<td>0.1386</td>
<td>0.0807</td>
<td>0.0080</td>
<td>0.0074</td>
</tr>
<tr>
<td>0.6789</td>
<td>-0.0084</td>
<td>0.1211</td>
<td>0.0478</td>
<td>0.0038</td>
<td>0.0066</td>
</tr>
<tr>
<td>0.7596</td>
<td>-0.0031</td>
<td>0.1404</td>
<td>0.0278</td>
<td>0.0056</td>
<td>0.0181</td>
</tr>
</tbody>
</table>

max $|\Delta|$ | 0.0189 | 0.1404 | 0.1002 | 0.0101 | 0.0181 | 0.0416

av $|\Delta|$ | 0.0063 | 0.0836 | 0.0534 | 0.0048 | 0.0089 | 0.0227
So far all the comparisons have been concerned with the equiprobable $\chi^2$ test using $\chi^2$ critical values. Further similar comparisons for the non-central $\chi^2$ and West approximations are given in Tables 7-14 of Frosini (1976). The equiprobable test is unbiased, tables of exact critical values are available (see Smith et al (1979)) and the usual $\chi^2$ approximation is known to be good (see Slakter (1968)). Thus if a choice of classes was available the equiprobable test would be the one to use. However, other $\chi^2$ tests need to be examined for those cases when there is no choice. Table 4 briefly examines the case $k = 3$ when the null probabilities are $\frac{1}{15}$, $\frac{1}{3}$ and $\frac{3}{5}$.

Table 4

Errors in power approximations and exact power, $\beta$, for unequal null probabilities when $k = 3$, $n = 10$, 20 and nominal test size is 0.05.

<table>
<thead>
<tr>
<th>n</th>
<th>$\pi_1$, $\pi_2$, $\pi_3$</th>
<th>$\beta$</th>
<th>$\Delta_{HC}$</th>
<th>$\Delta_H$</th>
<th>$\Delta_X$</th>
<th>$\Delta_C$</th>
<th>$\Delta_W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>.2, .4, .4</td>
<td>0.407</td>
<td>-0.0437</td>
<td>0.1434</td>
<td>-0.0152</td>
<td>-0.0145</td>
<td>0.0015</td>
</tr>
<tr>
<td>10</td>
<td>.2, .3, .5</td>
<td>0.3498</td>
<td>-0.0410</td>
<td>0.1675</td>
<td>-0.0019</td>
<td>0.0015</td>
<td>0.0111</td>
</tr>
<tr>
<td>10</td>
<td>.33, .33, .33</td>
<td>0.7405</td>
<td>0.1385</td>
<td>0.0245</td>
<td>-0.0140</td>
<td>0.0047</td>
<td>0.0075</td>
</tr>
<tr>
<td>20</td>
<td>.07, .33, .6</td>
<td>0.0494</td>
<td>0.0006</td>
<td>-0.0218</td>
<td>-0.0007</td>
<td>0.0047</td>
<td>0.0049</td>
</tr>
<tr>
<td>20</td>
<td>.1, .2, .7</td>
<td>0.2091</td>
<td>-0.0087</td>
<td>0.0686</td>
<td>-0.0762</td>
<td>-0.0157</td>
<td>0.0020</td>
</tr>
<tr>
<td>20</td>
<td>.2, .3, .5</td>
<td>0.5164</td>
<td>0.0471</td>
<td>0.1186</td>
<td>0.0017</td>
<td>0.0031</td>
<td>0.0027</td>
</tr>
<tr>
<td>20</td>
<td>.2, .4, .4</td>
<td>0.6113</td>
<td>0.0397</td>
<td>0.0739</td>
<td>-0.0090</td>
<td>-0.0105</td>
<td>-0.0062</td>
</tr>
</tbody>
</table>

$\max |\Delta|$, $0.1385$, $0.1675$, $0.0762$, $-0.0145$, $0.0111$

$\text{av} |\Delta|$, $0.0456$, $0.0884$, $0.0170$, $0.0078$, $0.0051$

If anything, Table 4 indicates the errors in the non-central $\chi^2$ and normal approximations are so gross that use of these approximations could not be recommended. Once again the Edgeworth and three-moment $\chi^2$ approximations are superior. Further, as in Table 2 but not in Tables 1 and 3 the West approximation performs well.

West (1971) compares approximations using different critical values to the $\chi^2$ points considered so far. Most users would probably use the $\chi^2$ points and this is the main justification for our choice but it is also of interest to extend West's comparisons using critical values which the $\chi^2$ statistic actually attains. Table 5 extends his Table X and shows that for the approximations considered either West's approximation or, as before, the three-moment $\chi^2$ approximation are preferred.

For $n = 10$ all the approximations shown in Table 5 occasionally have quite large errors and if possible either the Monte Carlo or exact methods should be used. Table 6 extends West's Table XI. The Edgeworth, West and three-moment $\chi^2$ approximations do best of those compared.
### Table 5

Errors in power approximations and exact power, $\beta$, for West's alternatives when $n=10$, $k=4$, $p_1=0.05$, $p_2=0.15$, $p_3=0.35$, $p_4=0.45$.

<table>
<thead>
<tr>
<th>$\pi_1$, $\pi_2$, $\pi_3$, $\pi_4$</th>
<th>C</th>
<th>Size</th>
<th>$\beta$</th>
<th>$\Delta_{NC}$</th>
<th>$\Delta_{N}$</th>
<th>$\Delta_{E}$</th>
<th>$\Delta_{C}$</th>
<th>$\Delta_{M}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.15, .35, .2, .3</td>
<td>14.1270</td>
<td>.0100</td>
<td>.2885</td>
<td>-.1330</td>
<td>.0583</td>
<td>-.0421</td>
<td>-.0275</td>
<td>-.0199</td>
</tr>
<tr>
<td>7.1429</td>
<td>.0502</td>
<td>.5811</td>
<td>-.0246</td>
<td>.0864</td>
<td>-.0509</td>
<td>-.0004</td>
<td>.0075</td>
<td></td>
</tr>
<tr>
<td>5.3651</td>
<td>.1053</td>
<td>.7391</td>
<td>-.0369</td>
<td>.0010</td>
<td>-.0290</td>
<td>-.0397</td>
<td>-.0384</td>
<td></td>
</tr>
<tr>
<td>.2, .1, .25, .45</td>
<td>14.1270</td>
<td>.0100</td>
<td>.2663</td>
<td>-.1483</td>
<td>.0744</td>
<td>-.0391</td>
<td>-.0397</td>
<td>-.0278</td>
</tr>
<tr>
<td>7.1429</td>
<td>.0502</td>
<td>.4147</td>
<td>.0749</td>
<td>.1871</td>
<td>.0336</td>
<td>.0368</td>
<td>.0537</td>
<td></td>
</tr>
<tr>
<td>5.3651</td>
<td>.1053</td>
<td>.5848</td>
<td>.0575</td>
<td>.0810</td>
<td>-.0452</td>
<td>-.0382</td>
<td>-.0260</td>
<td></td>
</tr>
<tr>
<td>.15, .1, .25, .5</td>
<td>14.1270</td>
<td>.0100</td>
<td>.1459</td>
<td>.1075</td>
<td>.0102</td>
<td>-.0836</td>
<td>-.0273</td>
<td>-.0149</td>
</tr>
<tr>
<td>7.1429</td>
<td>.0502</td>
<td>.2771</td>
<td>.0002</td>
<td>.2012</td>
<td>.0142</td>
<td>.0208</td>
<td>.0614</td>
<td></td>
</tr>
<tr>
<td>5.3651</td>
<td>.1053</td>
<td>.4560</td>
<td>.0316</td>
<td>.1192</td>
<td>-.0663</td>
<td>-.0656</td>
<td>-.0220</td>
<td></td>
</tr>
<tr>
<td>.1, .2, .3, .4</td>
<td>14.1270</td>
<td>.0100</td>
<td>.0619</td>
<td>.0525</td>
<td>-.0335</td>
<td>-.0206</td>
<td>-.0074</td>
<td>-.0070</td>
</tr>
<tr>
<td>7.1429</td>
<td>.0502</td>
<td>.1786</td>
<td>.0502</td>
<td>.1372</td>
<td>-.0282</td>
<td>.0151</td>
<td>.0474</td>
<td></td>
</tr>
<tr>
<td>5.3651</td>
<td>.1053</td>
<td>.3146</td>
<td>.0774</td>
<td>.1390</td>
<td>-.0529</td>
<td>-.0372</td>
<td>.0102</td>
<td></td>
</tr>
<tr>
<td>.05, .1, .45, .4</td>
<td>14.1270</td>
<td>.0100</td>
<td>.0094</td>
<td>.0028</td>
<td>-.0094</td>
<td>.0014</td>
<td>-.0003</td>
<td>-.0040</td>
</tr>
<tr>
<td>7.1429</td>
<td>.0502</td>
<td>.0384</td>
<td>.0672</td>
<td>.0412</td>
<td>-.0296</td>
<td>.0480</td>
<td>.0532</td>
<td></td>
</tr>
<tr>
<td>5.3651</td>
<td>.1053</td>
<td>.0789</td>
<td>.1258</td>
<td>.1456</td>
<td>-.0159</td>
<td>.0795</td>
<td>.1033</td>
<td></td>
</tr>
<tr>
<td>.05, .15, .3, 5</td>
<td>14.1270</td>
<td>.0100</td>
<td>.0099</td>
<td>-.0064</td>
<td>-.0099</td>
<td>-.0023</td>
<td>-.0014</td>
<td>-.0053</td>
</tr>
<tr>
<td>7.1429</td>
<td>.0502</td>
<td>.0572</td>
<td>.0195</td>
<td>.0060</td>
<td>.0572</td>
<td>.0186</td>
<td>.0232</td>
<td></td>
</tr>
<tr>
<td>5.3651</td>
<td>.1053</td>
<td>.1542</td>
<td>-.0071</td>
<td>.0409</td>
<td>-.1370</td>
<td>-.0160</td>
<td>-.0060</td>
<td></td>
</tr>
</tbody>
</table>

| max | $\Delta$ | .1483 | .2012 | .1370 | .0795 | .1033 |
| av  | $\Delta$ | .0569 | .0768 | .0416 | .0289 | .0296 |

### Table 6

Errors in power approximations and exact power, $\beta$, for West's alternatives when $n=20$, $k=4$, $p_1=.05$, $p_2=.15$, $p_3=.35$, $p_4=.45$.

<table>
<thead>
<tr>
<th>$\pi_1$, $\pi_2$, $\pi_3$, $\pi_4$</th>
<th>C</th>
<th>Size</th>
<th>$\beta$</th>
<th>$\Delta_{NC}$</th>
<th>$\Delta_{N}$</th>
<th>$\Delta_{E}$</th>
<th>$\Delta_{C}$</th>
<th>$\Delta_{M}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.15, .35, .2, .3</td>
<td>11.7778</td>
<td>.0100</td>
<td>.6132</td>
<td>.0105</td>
<td>.0596</td>
<td>.0068</td>
<td>.0001</td>
<td>-.0003</td>
</tr>
<tr>
<td>7.3333</td>
<td>.0502</td>
<td>.8151</td>
<td>-.0347</td>
<td>-.0029</td>
<td>.0044</td>
<td>.0053</td>
<td>.0059</td>
<td></td>
</tr>
<tr>
<td>5.3668</td>
<td>.1014</td>
<td>.8885</td>
<td>.0376</td>
<td>-.0306</td>
<td>-.0011</td>
<td>.0085</td>
<td>.0108</td>
<td></td>
</tr>
<tr>
<td>.2, .1, .25, .45</td>
<td>11.7778</td>
<td>.0100</td>
<td>.4537</td>
<td>.0542</td>
<td>.1370</td>
<td>.0173</td>
<td>.0176</td>
<td>.0234</td>
</tr>
<tr>
<td>7.3333</td>
<td>.0502</td>
<td>.6507</td>
<td>.1341</td>
<td>.0641</td>
<td>.0058</td>
<td>.0007</td>
<td>.0027</td>
<td></td>
</tr>
<tr>
<td>5.3668</td>
<td>.1014</td>
<td>.7403</td>
<td>.1426</td>
<td>.0223</td>
<td>-.0133</td>
<td>.0061</td>
<td>.0036</td>
<td></td>
</tr>
<tr>
<td>.15, .1, .25, .5</td>
<td>11.7778</td>
<td>.0100</td>
<td>.2547</td>
<td>-.0506</td>
<td>.1315</td>
<td>.0021</td>
<td>.0077</td>
<td>.0268</td>
</tr>
<tr>
<td>7.3333</td>
<td>.0502</td>
<td>.4588</td>
<td>.0206</td>
<td>.1286</td>
<td>-.0139</td>
<td>-.0167</td>
<td>.0076</td>
<td></td>
</tr>
<tr>
<td>5.3668</td>
<td>.1014</td>
<td>.5897</td>
<td>.0545</td>
<td>.0814</td>
<td>-.0254</td>
<td>-.0282</td>
<td>-.0114</td>
<td></td>
</tr>
<tr>
<td>.1, .2, .3, .4</td>
<td>11.7778</td>
<td>.0100</td>
<td>.1013</td>
<td>-.0564</td>
<td>.0224</td>
<td>.0465</td>
<td>.0092</td>
<td>.0154</td>
</tr>
<tr>
<td>7.3333</td>
<td>.0502</td>
<td>.2375</td>
<td>.0528</td>
<td>.1346</td>
<td>.0092</td>
<td>.0087</td>
<td>.0322</td>
<td></td>
</tr>
<tr>
<td>5.3668</td>
<td>.1014</td>
<td>.3444</td>
<td>-.0210</td>
<td>.1698</td>
<td>.0111</td>
<td>.0115</td>
<td>.0423</td>
<td></td>
</tr>
<tr>
<td>.05, .1, .45, .4</td>
<td>11.7778</td>
<td>.0100</td>
<td>.0135</td>
<td>.0151</td>
<td>-.0097</td>
<td>.0356</td>
<td>.0110</td>
<td>.0089</td>
</tr>
<tr>
<td>7.3333</td>
<td>.0502</td>
<td>.0460</td>
<td>.0916</td>
<td>.0727</td>
<td>.0338</td>
<td>.0688</td>
<td>.0716</td>
<td></td>
</tr>
<tr>
<td>5.3668</td>
<td>.1014</td>
<td>.0836</td>
<td>.1762</td>
<td>.2136</td>
<td>.1159</td>
<td>.1387</td>
<td>.1494</td>
<td></td>
</tr>
<tr>
<td>.05, .15, .3, 5</td>
<td>11.7778</td>
<td>.0100</td>
<td>.0121</td>
<td>.0002</td>
<td>-.0115</td>
<td>.0237</td>
<td>.0032</td>
<td>.0002</td>
</tr>
<tr>
<td>7.3333</td>
<td>.0502</td>
<td>.0736</td>
<td>.0060</td>
<td>-.0137</td>
<td>-.0329</td>
<td>.0036</td>
<td>.0046</td>
<td></td>
</tr>
<tr>
<td>5.3668</td>
<td>.1014</td>
<td>.1707</td>
<td>.0029</td>
<td>.0340</td>
<td>-.0729</td>
<td>-.0128</td>
<td>-.0009</td>
<td></td>
</tr>
</tbody>
</table>

| max | $\Delta$ | .1762 | .2136 | .1159 | .1387 | .1494 |
| av  | $\Delta$ | .0534 | .0744 | .0199 | .0233 |       |
5. Summary

From the evidence we have presented it is suggested that either the Monte Carlo (with 2,000 simulations) or three-moment $\chi^2$ approximations are preferred methods of quickly estimating the power, $\beta$. It is recommended that the non-central $\chi^2$ or normal approximations not be used. Further, the Edgeworth approximation is no more accurate than the three-moment approximation and is more complicated. Thus we do not recommend it either. Computer costs could be a problem with the Monte Carlo method, particularly as $n$ increases. This leaves the three-moment $\chi^2$ approximation as our preferred approximation. A FORTRAN routine which implements the three-moment $\chi^2$ approximation is available on request from the first author.

References

UNIFORMITY TESTING WHEN ALTERNATIVES HAVE LOW ORDER

By D. J. BEST

CSIRO Division of Mathematics and Statistics, Australia

and

J. C. W. RAYNER

University of Otago, New Zealand

SUMMARY. In recent years a variety of statistics have been suggested for testing uniformity. Many of these statistics, and sometimes their components, perform well in power studies. Here we explain these power performances in terms of the effective order of the alternative hypotheses. The components of Pearson's $X^2$ are revived and shown to be both effective, and convenient in applications.
LANCASTER'S TEST OF NORMALITY

D.J. BEST

C.S.I.R.O., Division of Mathematics and Statistics, P.O. Box 52, North Ryde, N.S.W., Australia

J.C.W. RAYNER

Mathematics Department, University of Otago, Dunedin, New Zealand

Received 26 October 1984; revised manuscript received 14 March 1985
Recommended by M.L. Puri

Abstract: A test of normality proposed by Lancaster (1969) is shown to have power comparable to that of other normality tests. Approximate formulae for critical values are given and the relation with a test given in Thomas and Pierce (1979) is discussed.

AMS Subject Classification: 62F03, 62F04.

Key words: Goodness of fit; Power comparisons; Neyman smooth test.

Article removed for copyright reasons, please see print copy for the article:
THE CONSTRUCTION OF THE SIMPLE \( X^2 \) AND NEYMAN SMOOTH GOODNESS OF FIT TESTS

by J.C.W. Rayner, * D.J. Best † and K.G. Dodds ‡

Abstract Suppose we wish to test whether data are consistent with a completely specified continuous distribution against a general alternative. Familiar omnibus tests are PEARSON'S \( X^2 \) test and NEYMAN'S smooth test. Fundamental problems in the application of these tests are the construction and number of classes to use for \( X^2 \), and the choice of the order of the NEYMAN smooth test. This paper examines these questions.

Key words: Goodness of fit tests, \( X^2 \) tests, smooth tests, class construction, number of components.

Article removed for copyright reasons, please see print copy for the article: Rayner, J.C.W., Best, D.J. and Dodds, K.G. (1985). The construction of the simple \( X^2 \) and Neyman smooth goodness of fit tests. Statistica Neerlandica, 39, 35-50.
Use of the Score Statistic for Testing Goodness-of-Fit of Some Generalised Distributions

J. C. W. Rayner and R. L. McIntyre

Mathematics Department, University of Otago, New Zealand

Abstract

The use of the score statistic to test whether a generalised distribution gives an improved fit over a non-generalised distribution is recommended. The score statistic for a generalised exponential family is derived. Several specific examples are given.

Key words: Score statistic, goodness-of-fit, exponential family.
GOODNESS OF FIT FOR THE UNIFORM, NORMAL
AND EXPONENTIAL DISTRIBUTIONS

by

D.J. Best,
Division of Mathematics and Statistics, CSIRO,
P.O. Box 52, North Ryde, 2113, Australia

and

J.C.W. Rayner,
Department of Mathematics and Statistics,
University of Otago, New Zealand

Language

Fortran 77

Description and Purpose

Given a sample of n (≤10) independent observations \(x_1, x_2, \ldots, x_n\) the routine below tests goodness of fit for any one of the standard uniform, normal or exponential distributions. The test for uniformity is just the smooth test based on the Legendre polynomials suggested by Neyman (1937) and discussed, for example, in Kendall and Stuart (1973, pp. 460-6). Normality and exponentiality are tested in the manner suggested by Lancaster (1969, pp. 121-3) using analogues of the Neyman test. The appropriate orthogonal polynomials in these cases are the
Hermite and the Laguerre. Power comparisons which indicate that Neyman's test is a good test of uniformity are given in Miller and Quesenberry (1979). Similarly, power comparisons also indicate the advantages of the normality and exponentiality tests given here. Details of these comparisons are given in Best and Rayner (1985) and Rayner and Best (1986). The test provided by the algorithm should be used in conjunction with a graphical technique such as a probability plot or histogram. For $n < 10$, where the approximate critical values may err, a $P$ value could be found, if necessary, by Monte Carlo methods as in Rayner and Best (§ 4 , 1986).

**Numerical Method**

(a) **Uniformity**

The statistic calculated is $S = \sum_{i=1}^{n} U_i^2$ where

$$U_i = ((2i+1)/n)^{1/2} \sum_{j=1}^{n} \pi_i(z_j)$$

and $\pi_1(z) = z$, $\pi_2(z) = (2z^2-1)$, $r\pi_r(z) = (2r-1)z\pi_{r-1}(z)-(r-1)\pi_{r-2}(z)$ for $r > 2$, with $z = (2x-1)$. To a good approximation, when $n \geq 10$, the upper 5% critical value of $S$ is the $\chi^2_4$ value 9.49 while the upper 10% critical value is the adjusted $\chi^2_4$ value given by $7.78(1-0.06n^{-1/2})$. The $U_i$, here and below, are approximately standard $N(0,1)$ and Solomon and Stephens (1983) note that $U_1$ and $U_2$ detect changes in mean and variance.
(b) Normality

In this case \( S = \sum_{i=3}^{6} U_i^2 \) where

\[
U_i = (n-i!)^{-\frac{1}{2}} \sum_{j=1}^{n} \pi_i(Z_j)
\]

and \( \pi_1(z) = z, \pi_2(z) = z^2 - 1, \pi_r(z) = z \pi_{r-1}(z) - (r-1) \pi_{r-2}(z) \) for \( r > 2 \), with \( z = (x-a)/b, a = \bar{x}, b = \sum_{i=1}^{n} (x_i - \bar{x})^2/n \). For \( n \geq 10 \) adjusted \( \chi^2 \) values for the 5% and 10% upper critical values were found to be 9.49(1-1.6n^{-\frac{1}{2}}) and 7.78(1-1.8n^{-\frac{1}{2}}). Table 1 shows actual approximate tail areas based on 1,000 samples each of size \( n \) for the adjusted and unadjusted \( \chi^2 \) critical values.

<table>
<thead>
<tr>
<th>( n )</th>
<th>5%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>.05 (.00)</td>
<td>.12 (.01)</td>
</tr>
<tr>
<td>15</td>
<td>.05 (.02)</td>
<td>.09 (.02)</td>
</tr>
<tr>
<td>20</td>
<td>.05 (.01)</td>
<td>.09 (.02)</td>
</tr>
<tr>
<td>30</td>
<td>.05 (.03)</td>
<td>.08 (.04)</td>
</tr>
<tr>
<td>50</td>
<td>.05 (.03)</td>
<td>.08 (.05)</td>
</tr>
<tr>
<td>70</td>
<td>.06 (.04)</td>
<td>.10 (.06)</td>
</tr>
<tr>
<td>90</td>
<td>.06 (.05)</td>
<td>.09 (.06)</td>
</tr>
<tr>
<td>200</td>
<td>.05 (.05)</td>
<td>.09 (.06)</td>
</tr>
</tbody>
</table>
U_3 and U_4 are now just the usual standardized sample skewness and kurtosis coefficients. Best and Rayner (1985) give (i) power comparisons which show S competes well with other normality tests, (ii) examples of the use of the components, U_i and (iii) details of how the adjusted \( \chi^2 \) values were obtained.

(c) Exponentiality

\[ S = \sum_{k=2}^{5} U_k^2 \]

where

\[ U_i = (n_i)^{-\frac{1}{2}} \sum_{j=1}^{n} \pi_i(z_j) \]

and

\[ \pi_1(z) = 1 - z, \quad \pi_2(z) = 1 - 2z + z^2/2, \]

\[ r\pi_r(z) = (2r-1-z)\pi_{r-1}(z)-(r-1)\pi_{r-2}(z) \]

for \( r > 2 \), with \( z = x/\bar{x} \). For \( n \geq 10 \) adjusted \( \chi^2 \) values for the 5% and 10% upper critical values were found to be 9.49(1-1.5n^{-\frac{1}{2}}) and 7.78(1-1.8n^{-\frac{1}{2}}). The component U_2 is related to an exponentiality test discussed by Pettitt (1978).
Suppose it is desired to test the adequacy of the fit of the log-normal distribution to the annual maximum 24 hourly rainfall for 47 years for Turramurra (northern suburb of Sydney, Australia). The data (in mm) are:

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1468</td>
<td>909</td>
<td>841</td>
<td>475</td>
<td>846</td>
<td>452</td>
<td>3830</td>
<td>1397</td>
<td></td>
</tr>
<tr>
<td>556</td>
<td>978</td>
<td>1715</td>
<td>747</td>
<td>909</td>
<td>2002</td>
<td>1331</td>
<td>1227</td>
<td></td>
</tr>
<tr>
<td>2543</td>
<td>2649</td>
<td>1781</td>
<td>1717</td>
<td>2718</td>
<td>584</td>
<td>1859</td>
<td>1138</td>
<td></td>
</tr>
<tr>
<td>2675</td>
<td>1872</td>
<td>1359</td>
<td>1544</td>
<td>1372</td>
<td>1334</td>
<td>955</td>
<td>1849</td>
<td></td>
</tr>
<tr>
<td>719</td>
<td>1737</td>
<td>1389</td>
<td>681</td>
<td>1565</td>
<td>701</td>
<td>994</td>
<td>1188</td>
<td></td>
</tr>
<tr>
<td>962</td>
<td>1564</td>
<td>1800</td>
<td>580</td>
<td>1106</td>
<td>880</td>
<td>850</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

After taking logarithms subroutine GOF may be used with K=1 to obtain $S = 0.59$ ($U_3 = 0.01$, $U_4 = -0.67$, $U_5 = 0.25$, $U_6 = 0.28$). The $S$ value is well below the approximate 5% critical value of $C = 7.3$ ($D = 5.7$) and so the lognormal fit is tenable.
References


subroutine gof(x,n,k,s,u,c,d,ifault)
dimension ::(n),u(6),h(6),f(4),g(4)
data f/2.449490,4.898979,10.95445,26.83282/
data g/1.732051,2.236068,2.645751,3.0/
goodness of fit for uniform, normal and exponential distribut

ifault=1
if (n.lt.10) return
ifault=0
do 1 j=1,6
u(j)=0
continue
1
e=sqrt(float(n))
s=0

after initializing perform appropriate test

first, uniformity test

if (k.eq.0) then
do 4 i=1,n
z=2*x(i)-1
h(1)=z
h(2)=(3*z+z-1)/2
do 2 k=3,4
h(k)=((2*k-1)*z*h(k-1)-(k-1)*h(k-2))/k
continue
do 3 j=1,4
u(j)=u(j)+g(j)*h(j)/e
continue
4
c=9.49
d=7.78*(1-0.06/e)

normality test

else if (k.eq.1) then
a=0
do 5 i=1,n
a=a+x(i)
continue
6
a=a/n
b=0
do 6 i=1,n
t=x(i)-a
b=b+t*t
b=sqrt(b/n)
do 9 i=1,n
z=(x(i)-a)/b
h(1)=z
h(2)=z*z-1
do 7 k=3,6
h(k)=z*h(k-1)-(k-1)*h(k-2)
continue
do 8 j=3,6
u(j)=u(j)+h(j)/(e*f(j-2))
continue
c=9.49*(1-1.4/a)


d=7.78*(1-1.8/e) 

c 

e: exponentiality test

c 

else if (k.eq.2) then 

    a=0
    do 10 i=1,n
    a=a+x(i)

10 continue

    a=a/n
    do 13 i=1,n
    z=x(i)/a
    h(1)=1-z
    h(2)=1-2*z+z*z/2
    do 11 k=3,5
    h(k)=((2*k-1-z)*h(k-1)-(k-1)*h(k-2))/k

11 continue

    do 12 j=2,5
    u(j)=u(j)+h(j)/e

12 continue

13 continue

c=9.49*(1-1.5/e)

d=7.78*(1-1.8/e)

else 
    ifault=2 
    return 
endif 

c 

calculate s

c 

    do 14 j=1,6
    s=s+u(j)*u(j)

14 continue

return 
end
The development of the smooth goodness of fit tests is reviewed. Since the $X^2$ tests are included in this class, there is much to consider. Pearson’s original contribution is discussed in detail, as is the question, “What is a goodness of fit test?”

1. INTRODUCTION

The book D’Agostino and Stephens (1986) gives a good introduction to many of the standard approaches to goodness of fit: via the empirical distribution function, regression and correlation methods, graphical methods and more. However it only briefly mentions the smooth tests, which are powerful and convenient, and deserve to be better known. The Pearson $X^2$ test is a smooth test, and its presentation as such is valuable, because it provides alternatives that may be used if the null hypothesis is rejected. This review examines the foundations of goodness of fit, the paper by Pearson (1900), and overviews the subsequent development of goodness of fit. But to begin, what is goodness of fit, and why is it important?

2. GOODNESS OF FIT: THE PROBLEM DEFINED

Important decisions are based on distributional assumptions. For example, quality control rules for the import and export of various foods often assume a normal distribution. Tolerance limits for extreme rainfall used by engineers involved in road-building sometimes assume a lognormal distribution. Failure of the distributional assumptions means failure of the model, thereby invalidating the conclusions of the model. This possibility of failure can be assessed by a goodness of fit test.

According to David (1966, p.399)

“A goodness of fit procedure is a statistical test of the hypothesis that the sampled population is distributed in a specific way ... for example, that the sampled population is normal.”

This is the one sample problem; the corresponding k-sample test assesses whether or not k independent random samples come from the same population. Related problems in estimation include forming a confidence interval for a population distribution function. Here only the one sample problem is discussed.

Formally the one sample test for goodness of fit is as follows: given a random sample $X_1, X_2, ..., X_n$, we wish to test the null hypothesis that the sample comes from a population with cumulative distribution function $F(x; \theta), \theta \in \Theta$ against the alternative that the cumulative distribution function is $G(x; \omega), \omega \in \Omega$. All of $X, \Theta$ and $\Omega$ may be multidimensional. Frequently the alternative is merely
What do we get from having applied a goodness of fit test?

(1) **A compact description of the data:** Saying that the data is binomial with parameters \( n = 15 \) and \( p = 0.51 \) is a valuable abbreviation of the available information.

(2) **Availability of powerful parametric procedures:** For example, the analysis of variance becomes available if the data are consistent with normality.

(3) **Light may be shed on the mechanisms generating the data:** For example if lifetimes for cancer patients from the onset of "standard" treatment have been exponentially distributed with mean 36 months in the past, and this distribution no longer holds under a new treatment, what has changed? It could be that either the mean or the distribution has changed. In the latter case perhaps treatment is relatively ineffective for some, and they die sooner than under the standard treatment, and the treatment is apparently effective for others, who survive longer than previously.

What a goodness of fit test tells us is important, but so is what it doesn't tell us! Geary (1947) says that "Normality is a myth; there never was, and never will be a normal distribution."

Strongly put perhaps, but given enough observations of virtually any generating mechanism we could reject any specified hypothesis. As for normality, we do not observe arbitrarily large (or small) data; and as all data are rounded, we should ultimately be able to reject any continuous model. But although a distributional model may not hold precisely, it may hold sufficiently well for the purposes outlined above. The important question is, *are our data sufficiently well approximated by the distribution for which we test?*

Of course a passive attitude may be adopted: a particular powerful parametric procedure will be used unless there is strong evidence to the contrary. A "weak" goodness of fit test fails to find such evidence, and so with much relief, the parametric procedure is used. Why the relief? Because there may be no suitable nonparametric procedure to replace the parametric one; or if there is, it may be inconvenient or have known disadvantages, such as tables not readily available, or difficulties in handling ties.

This of course is hardly good science. Nor is it sufficient to say that the procedure is "robust". This is rarely all the story. Rounding may not degrade least squares, but it seriously affects a test for exponentiality (see Cox and Lewis (1966, p.162)). Another procedure may be seriously distorted if the sample values are correlated, but impervious to so called "normal" observations being heavily skewed. We do not have the detail of knowledge to know, say, that a procedure is adequate if a test for normality against skewness alternatives is not significant at the 15% level, and if a test for normality against kurtosis alternatives is not significant at the 3% level. The best we can do is to use our most powerful goodness of fit tests to eliminate unsuitable models.

It is hardly a solution to avoid the goodness of fit test by avoiding the parametric procedure. Non-parametric or distribution-free procedures have their place and their use will sometimes be highly desirable. However, if parametric procedures, such as the analysis of variance and regression tests, are appropriate, (that is, if the distributional and other necessary assumptions hold) then these parametric procedures are, in some senses, superior to their non-parametric competitors. Superior may mean greater asymptotic relative efficiency or small sample power, or sometimes more accessible software, or readier access to necessary tables, or easier hand computation. If at all possible, these superior procedures should be used!

3. **FOUNDATIONS**

Before Karl Pearson introduced the \( \chi^2 \) goodness of fit test in Pearson (1900), model assessment was a chancy business. Pearson himself, in Pearson (1897), could only conclude
"Considering the rough nature of cloudiness observations, the agreement must be considered fairly good, and very probably the smooth results of the theory are closer to the real facts of the case than the irregular observations".

See Plackett (1983, p.61). The "considering" here is purely subjective. Three years later, in his hallmark paper, Pearson criticises this sort of subjectivity:

"But the comparison of observation and theory in general amounts to a remark—based on no quantitative criterion—of how well theory and practice really do fit"

See Pearson (1900, p.171). The need for objectivity in model assessment is no less today than it was a century ago, but thankfully we now have the tools. That we do have the tools is very much due to Karl Pearson. In reviewing this remarkable paper it is as well to put Pearson’s achievement in context. This is done in detail in Plackett (1983) who, in part, says

"During the period 1983-9, ... He developed his system of frequency curves, laid the foundations of multiple correlation and regression, and obtained the probable errors and correlations of estimated coefficients in large samples. Everything was directed towards the study of problems in heredity and evolution. Among these manifold activities is included his presentation of the multivariate normal distribution."

Pearson (1900) starts directly and somewhat abruptly into a "preliminary proposition", now sometimes called Pearson’s lemma. In modern notation, if $X$ is multivariate normal with mean zero and nonsingular covariance matrix $V$, than $Q = X^T V^{-1} X$ has a central chi-squared distribution. The probability density function is given.

In the second section of the paper, integration by parts is used to develop a power series expansion for the $X^2$ tail probabilities. These were calculated with the help of Miss Alice Lee. Probabilities are given to 6 decimal places, for degrees of freedom from 2 to 19 and give

"the probability of a system of deviations as great or greater than the outlier in question."

The $X^2$ test is derived in the third section. Observations are grouped into $n + 1$ categories, so that there are $n$ algebraically independent random variables. The multinomial variances and covariances are cited, and then a polar transformation used to convert the covariance matrix to apply the previously mentioned $X^T V^{-1} X$ result. Of course, this assumes the multinomial approaches the multivariate normal, and that the sample size is sufficiently large. Now in this situation the quadratic form $Q$ turns out to be $X^2$, so we have the desired test statistic.

Pearson now concludes that this result is of very great simplicity, and very easily applicable.

He then outlines how, in three stages, to apply the result.

Then in section 5 the vexed question of estimating parameters is addressed. Unfortunately Pearson concludes that estimation makes no difference, and this question was only properly resolved by Fisher in a series of papers in the 1920’s.

Next follows a section with three illustrations where estimation is not required, and a section with five illustrations where estimation is required, and finally a brief conclusion. Of the illustrations it is informative to quote Cochran (1952, p.319).

"Finally, the paper contains eight numerical applications of the new technique. In two of these he pokes fun at Sir George Airy and Professor Merriman. They had both published series of observations which they claimed to be good illustrations of variates that follow the normal distribution. In the absence of any test of goodness of fit, Airy and Merriman could judge this question only by eye inspection. Pearson showed that the significance probability for Airy’s data was 0.014, although the data from which Pearson calculated $X^2$ had already been
smoothed by Airy. Merriman fared worse, his probability being 1½ parts in a million. These examples show the weak position in which the scientist was placed when he had to judge goodness or badness of fit in the absence of an objective test of significance.”

For more detail on Pearson, the period, and the early development of the $X^2$ test, the reader is referred especially to Lancaster (1969, Chapters I and V) and Plackett (1983). Even today, Pearson’s $X^2$ test could be claimed to be one of the most widely known tests of statistical inference. As Plackett (1983) concludes “Pearson’s 1900 paper on chi squared is one of the great monuments of twentieth century statistics”.

4. THE PEARSON $X^2$ TEST AND ITS PROPERTIES

In the Pearson $X^2$ test for a simple null hypothesis it is desired to test the null hypothesis that a random sample $X_1, ..., X_n$ of size $n$ comes from a population with completely specified cumulative distribution function $F(x)$, against a general alternative, not $F(x)$. The sample space is partitioned into $k$ classes, and $N_i$ is defined as the number of observations from the sample that fall into the $i$th class. If $p_i$ is the probability of falling into the $i$th class when $F(x)$ holds, the Pearson $X^2$ statistic is defined by

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

The asymptotic null distribution of $X^2$ is $\chi^2_{k-1}$ chi-squared with $k-1$ degrees of freedom. Thus if $X^2 > a(\alpha)$, the 100$\alpha$% point of the $\chi^2_{k-1}$ distribution, then the null hypothesis that we are sampling from $F(x)$ is rejected at the 100$\alpha$% level. If a particular alternative holds, and gives $\pi_i$ as the probability of falling into the $i$th class, then the asymptotic non-null distribution of $X^2$ is $\chi^2_{k-1}(n\phi^2)$, non-central chi-squared with $k-1$ degrees of freedom and parameter of non-centrality $n\phi^2$, where

$$\phi^2 = \sum_{i=1}^{k} \frac{(\pi_i - p_i)^2}{p_i}$$

This distribution may be used to calculate powers, but see Best, Rayner and Turnbull (1983), for references and a simple 3 moment $\chi^2$ approximation to the power function.

For general review material see Cochran (1952, 1954), Lancaster (1969), Horn (1977), and Hutchinson (1979). Here we concentrate on more recent developments.

If the distribution specified by the null hypothesis is continuous, then the statistician has the problem of constructing the cells or classes. As for example, Kempthorne (1967) has pointed out, different conclusions may be reached if different constructions are used. One of the first constructions recommended that the expected cell frequencies, $np_i$, should all be at least 5; see for example Fisher (1925, p.84). If this is so the limiting null distribution of $X^2$, namely $\chi^2_{k-1}$, is a reasonable approximation for relatively small $n$. Recommendations based on the criterion of accurately approximating the null distribution of $X^2$ occur regularly: see the references in Holtzman and Good (1986), and especially Roscoe and Byars (1971), Larntz (1978), Koehler and Larntz (1980), and Lawal (1980). The Roscoe and Byars (1971) recommendations are:

(i) With equiprobable cells, the average expected cell frequency should be at least 1 (that is $n \geq k$) with a 5% significance level, and should be at least 2 (that is, $n \geq 2k$) with a 1% significance level.

(ii) When cells are not approximately equiprobable, the expected cell frequencies in (i) should be
(iii) For \( k = 2 \) the \( X^2_p \) test should be replaced by the test based on the exact binomial distribution.

For the equiprobable Pearson \( X^2 \) test, with \( p_i = k^{-1} \), Katti (1973) and Smith et al. (1979) have tabulated the exact null distribution of this test for \( n \leq 50 \) and various \( k \). At the 'usual' significance levels the nominal and exact sizes are usually acceptably close, but using these tables the exact size can be given. Holtzman and Good (1986) recommend the adjusted \( \chi^2 \) approximation in the equiprobable case: namely if \( y_0 \) is an observable value of \( X^2 \) then

\[
P(X^2_p \geq y_0) \approx P(Y \geq y_0 - k/n \mid Y \text{ is distributed as } \chi^2_{k-1})
\]

However, tests are usually preferred for a variety of reasons, only one of which is the null distribution is known, at least approximately. Tests must have "good" power or efficiency. The desirable power properties of the equiprobable Pearson test include that it is strictly unbiased and it is type D (a local optimality condition) as a multinomial test; see Cohen and Sackrowitz (1975). However it is not type D when used as a test of goodness of fit; see Rayner and Best (1982). The equiprobable Pearson test is appropriate in a distance sense according to Spruill (1976) and, as shown by Cox and Hinkley (1974), if the entropy is to be minimised.

If an \( X^2 \) test is to be used, until recently Lancaster (1980) was virtually alone in not recommending the equiprobable test. He suggested that the interval over which the alternative is defined should be subpartitioned where the slope of the alternative probability density function is largest, and should not be partitioned at all if the slope is less than unity. Although power studies tend to support Lancaster, I have two reservations. If the alternative were known, a more refined analysis, not using \( X^2 \) would probably be used. Also, as reported below, using more classes sometimes reduces the power. So for a given alternative, six classes with boundaries at points of steep slope of the alternative probability density function may be more powerful than six equiprobable classes. But ten classes, no matter how they are chosen, may well be less powerful. Power simulations in Kallenberg (1985) support Lancaster in that appreciable power gains are achieved using non-equiprobable tests for heavy tailed alternatives. The classes should be formed to be tail discriminating: that is, there should be relatively more, and hence lower probability classes in the heavy tails.

For the equiprobable Pearson's \( X^2 \) test, the problem is how to choose \( k \). Mann and Wald (1942) derived the formula

\[
k = 4[2(n - 1)^2/c^2]^{(1/5)}
\]

where \( c \) is the 100\( \alpha \)% point of the standard normal distribution and the test is being applied at the 100\( \alpha \)% level. Several subsequent authors, for example Williams (1950) and Schorr (1974), have recommended much smaller \( k \) than Mann and Wald. See also Harrison (1985). Dahiya and Gurland (1973) observe that the best \( k \) depends on the alternatives. Best and Rayner (1981) have found that for a sequence of fixed level Pearson \( X^2 \) tests with \( k = 2, 3, 4, ... \), the power typically increases to a maximum at \( k_0 \), say, then decreases. This \( k_0 \) is frequently quite small, at about 4 or 5. Using 20 plus classes as the Mann-Wald formula often suggests will then cause a substantial power loss.

The picture has recently been clarified by Kallenberg et al. (1985), Kallenberg (1985), and Oosterhoff (1985). Using local and non-local asymptotic theory on contamination and exponential families, it is shown that for heavy tailed alternatives larger \( k \) produce greater asymptotic power, while for lighter tailed alternatives smaller \( k \) produce greater asymptotic power. However the asymptotic models don't necessarily agree with finite sample calculations. Typically the statistician is presented with a sample of fixed size, and must choose \( k \). This needs to be reconciled with asymptotic theory in which \( k \) is fixed and
n → ∞, or both k and n → ∞ in some functionally dependent way. Nevertheless choosing k large for heavy tailed alternatives does agree well with the simulation studies in the literature. Best and Rayner (1981, 82, 85(a)) have suggested various alternatives as to how to choose k. These are

(i) to perform a sequence of \( X^2 \) tests with an increasing number of classes, or

(ii) use the components of \( X_p^2 \), checking for residual variation.

As an example of a situation where (i) is applicable, consider a preliminary investigation when testing a new random number generator. The sequence of \( X^2 \) tests will determine a \( k^*, \) that is most critical of the data. Effectively a class of alternatives is being singled out, the alternative one hopes to most powerfully detect in the subsequent investigation. An example of option (ii) is given in Best and Rayner (1985(b)). A typical outcome would be that particular components are significantly large. Then in subsequent investigations these can be focussed upon, to more powerfully detect particular alternatives. Using components with as many classes as possible eliminates loss of information due to categorisation. On the other hand, option (i) is easier and more familiar for most users.

The Pearson \( X^2 \) test is a member of the power divergence family of statistics introduced in Cressie and Read (1985). They define

\[
2nI^\lambda(X/n : p) = \left[ 2/\{\lambda(\lambda + 1)\} \right] \sum_{i=1}^{k} X_i \{X_i/(np_i)\}^\lambda - 1
\]

For \( \lambda = 1, 0, -0.5 \) this produces \( X^2 \), the likelihood-ratio test statistic, and the Freeman-Tukey statistics respectively. The performance of this class is further examined by Read (1984 a,b). These studies give perspective to the properties of Pearson’s test. From the point of view of Bahadur efficiency the likelihood-ratio test is the preferred member of the family; by Pitman efficiency the Pearson test is best. Comparisons between these two protagonists abound in the literature: see for example West and Kempthome (1971), Kallenberg et. al (1985), Kallenberg (1985), and Oosterhoff (1985). Rather than delve into these riches, note that Moore (1986) recommends the Pearson test of those in the power divergence family.

5. CHI-SQUARED TESTS WHEN PARAMETERS NEED TO BE ESTIMATED

Suppose now that the null distribution depends on a vector \( \beta \in \mathbb{R}^\ell \) of nuisance parameters. Then in \( X_p^2, \) \( p \) must be replaced by an estimate of \( \hat{p}_i = p_i(\hat{\beta}) \). If the \( \hat{p}_i \) are the maximum likelihood estimators based on the grouped observations the new statistic is \( X_{PF}^2 \), the Pearson-Fisher statistic. \( X_{PF}^2 \) has a null \( X_{k-\ell-1}^2 \) distribution, and a \( X_{k-\ell-1}(n\hat{\phi}^2) \) alternative distribution, where \( \hat{\phi}^2 \) is simply \( \phi^2 \) with both \( \pi_i \) and \( p \) replaced by cell probabilities using the estimate \( \hat{\beta} \) of the nuisance parameter \( \beta \). Although \( X_{PF}^2 \) may be less powerful than its competitors (see Rao and Robson (1974) but note the discussion following in this section), it would seem to provide a more robust test. \( X_{PF}^2 \) depends on the null hypothesis only in the placement of the cell boundaries and through the estimates, but not in its functional form.

If maximum likelihood estimators based on the ungrouped observations are used, we obtain the Chernoff-Lehmann statistic \( X_{CL}^2 \). The null distribution is \( X_{k-\ell-1}^2 \) plus a linear combination of \( \chi_1^2 \) variables, and depends on the unknown nuisance parameters. Different tables of null percentage points are required for each family of distributions for which we test. Using random cell boundaries in \( X_{CL}^2 \) in the manner described in Roy (1956) and Watson (1959) yields the \( X_R^2 \) statistic. In \( X_p^2 \) the dependence of the null distribution on the nuisance parameters is removed, but not the need for tabulation of critical
points for each family tested. Dahiya and Gurland (1972) treat the normal case.

Finally \( X_{RR}^2 \) is the Rao-Robson statistic, the form of which depends on the null probability density function tested for, but whose null distribution is \( X_{k-1}^2 \) no matter how many parameters are estimated.

Moore (1977) gives a "recipe" for producing goodness of fit statistics with chi-squared distributions. These may have fixed or random cells. \( X_{PF}^2 \) and \( X_{RR}^2 \) are both produced by this method; their asymptotic non-null distributions are both non-central chi-squared. \( X_{RR}^2 \) may be viewed as \( X_{CL}^2 \) plus a correction factor to recover the lost degrees of freedom.

It seems to be usually assumed, as in Kopecky and Pierce (1979), that \( X_{PF}^2 \) is less powerful than \( X_{CL}^2 \) and \( X_{RR}^2 \), which are less powerful than \( X_{RR}^2 \). The evidence is a series of particular cases, and the point of Moore and Spruill (1975, p.615), that \( X_{PF}^2 \) is neither always better nor always worse than \( X_{CL}^2 \) and \( X_{RR}^2 \), overlooked. It appears that for any particular alternative, various statistics have different degrees of freedom and different parameters of non-centrality, and the most powerful test is determined by how each factor compensates for the other.

Another route to these \( X^2 \) statistics is via quadratic score statistics. Kopecky and Pierce (1979, section 3) show that some of the Pearson-type statistics mentioned above are particular cases of quadratic score statistics. Thomas and Pierce (1979) point out that these statistics are generalised Neyman statistics with a "correction term" to adjust for estimation. If the nuisance parameters that enter the problem are location-scale parameters, then the suggested statistics do not involve the nuisance parameters. However for each family tested, a different form of the test statistic must be calculated.

### 6. NEYMAN'S SMOOTH TESTS

So far we have discussed only \( X^2 \) tests. If the null hypotheses is simple (there are no nuisance parameters) and the data ungrouped, Neyman's (1937) smooth test is a competitor to Pearson's \( X^2 \) test. Applying the probability integral transformation \( Y = F(X) \) specifies that \( Y \) is uniformly distributed on \((0,1)\). An alternative of order \( k \) to \( H_0 \) has probability density function

\[
g(y; \theta) = \exp \left\{ \sum_{i=1}^{k} \theta_i \pi_i(y) - K(\theta) \right\}, \quad 0 < y < 1; \quad k = 1, 2, 3, \ldots
\]

where \( \theta^T = (\theta_1, \ldots, \theta_k) \), \( K(\theta) \) is a normalising constant, and the \( \pi_i(x) \) are orthonormal polynomials related to the Legendre polynomials. The first five such polynomials are:

\[
\pi_0(x) = 1, \quad \pi_1(x) = \sqrt{3}(2x - 1), \quad \pi_2(x) = \sqrt{5}(6x^2 - 6x + 1)
\]

\[
\pi_3(x) = \sqrt{7}(20x^3 - 30x^2 + 12x - 1), \quad \pi_4(x) = 3(70x^4 - 140x^3 + 90x^2 - 20x + 1)
\]

\( \pi_i(x) \) is of degree \( i \) and the \( \pi_i(x) \) are orthonormal on \((0,1)\).

To test the null hypothesis \( H : \theta_1 = \ldots = \theta_k = 0 \) the Neyman statistic

\[
\text{Neyman statistic}
\]

---

---
\[ \psi_k^2 = \sum_{i=1}^{k} U_i^2 \quad \text{where} \quad U_i = \sum_{j=1}^{n} \pi_j(y_j) / \sqrt{n} \]

is used.

Neyman's smooth test based on \( \psi_k^2 \) was constructed to be locally most powerful unbiased size \( \alpha \) for testing for uniformity against the order \( k \) alternatives. Its power function was also constrained to be symmetric, depending on \( \theta \) only through \( \lambda^2 = \theta_1^2 + \ldots + \theta_k^2 \). Neyman (1937) notes that his solution is only approximate; only asymptotically is the test of size \( \alpha \), unbiased, and most powerful.

Barton (1953, 1955, 1956) works with probability density function's asymptotically equivalent to \( g(y; \theta) \), namely

\[ h(y; \theta) = 1 + \sum_{i=1}^{k} \theta_i \pi_i(y), \quad 0 < y < 1; \quad k = 1, 2, 3, \ldots \]

His 1956 paper deals with probability density functions involving nuisance parameters, but the statistic derived has an inconvenient distribution. As Kopecky and Pierce (1979) point out, the quadratic score statistic has a more convenient distribution.

An interesting but little known result is that the Pearson \( X^2 \) test (not necessarily equiprobable), is a categorised form of the Neyman-Barton tests. Suppose a multinomial with \( g \) classes is specified by the null hypothesis. Barton (1955) considered order \( k \) alternatives of the form \( h(y; \theta) \), but with the polynomials \( \pi_i(y) \) replaced by an orthonormal system on the multinomial distribution. He then defines a statistic \( B(g, k) \) that as \( g \to \infty \) tends to \( \psi_k^2 \), to be distributed as \( \chi^2_k \), and tends to optimality. Moreover \( B(k + 1, k) \) is shown to be \( X^2 \) based on \( k + 1 \) classes. The importance of this result is that the Pearson \( X^2 \) test can be expected to have good power properties against order \( k \) alternatives, especially for a moderate to large number of classes, when it will be very similar to the optimal \( \psi_k^2 \). Subsequently Kendall and Stuart (1973, p.44) showed the \( B(g, k) \) may be obtained by partitioning \( X^2 \).

Watson (1959) extended a result of Barton (1956), and Hamdan (1961, 1962, 1963), considered various simple smooth tests. He used the Hermite-Chebyshev polynomials to construct a test for the standard normal distribution, and an orthonormal set on the multinomial and the Walsh functions to construct tests for the uniform distribution.

These tests aroused little interest. For hand computations the calculations required are heavy, and the tests could not deal practically with composite null hypotheses, a major interest in applications. So it was not until 1979 that the next significant step forward appeared. By then computers had solved the computation problem, and practical composite tests were proposed.

Thomas and Pierce (1979) noted that it may be counterproductive to work with orthogonal polynomials, and instead defined an order \( k \) probability density function by

\[ \exp \left\{ \sum_{i=1}^{k} \theta_i F_i(x; \beta) - K(\theta) \right\} f(x; \beta) \]

The quadratic score statistic based on this model is \( W_k^* \), given in detail in Thomas and Pierce (1979, p.443), and the test based on \( W_k \) can be interpreted as a generalised smooth test.

Tests based on what might be called the Pierce approach include Bargal and Thomas' (1983) test for the (censored) Weibull, and Bargal's (1986) test for the (censored) gamma. Unfortunately a consequence of using powers is that tables of percentage points are required for the implementation of these tests.
Rayner and Best (1986) have combined the Neyman and Pierce approaches to produce smooth tests for location-scale families using orthonormal functions. The order \( k \) alternative has the form
\[
g^*(y; \xi) = \exp\left\{ \sum_{i=1}^{k} \theta_i h_i(y; \beta) - K(\xi) \right\} f(y; \beta) ; \quad k = 1, 2, 3, \ldots
\]
The appropriate test statistic is
\[
S_k^2 = \sum_{i=1}^{k} \hat{V}_i^2, \quad \text{where} \quad \hat{V}_i = \sum_{j=1}^{n} h_i \{ (X_j - \hat{\mu})/\hat{\sigma} \} / \sqrt{n}
\]
in which \( h_i(z) \) is orthonormal on the standardised probability density function. \( \hat{\mu} \) and \( \hat{\sigma} \) are maximum likelihood estimates of the location parameter \( \mu \) and the scale parameter \( \sigma \) respectively, and the summation is over \( k \) non-zero summands. For example, in testing for the normal distribution with unspecified mean \( \mu \) and variance \( \sigma^2 \), the first two \( V_i \)'s are zero; the next two assess skewness and kurtosis. The test statistic and the components \( V_i \)'s all have asymptotic \( \chi^2 \) distributions, so the test is easy to implement. Moreover the components may be individually informative. Examples are given in Best and Rayner (1985b). This approach can readily be extended beyond location-scale families, and to categorised data and multivariate distributions. This work is in hand.

To complete the picture, I should note that the smooth model appropriate in the discrete case is
\[
\pi_j = C(\xi) \exp\left\{ \sum_{i=1}^{m} \theta_i h_{ij} \right\} p_j, \quad j = 1, \ldots, k
\]
Here \( m \) is the order of the alternative, and the \( \chi^2 \) test results if in the score statistic we take \( m = k - 1 \) and choose the \( h_{ij} \) appropriately. The appeal in this formulation is that if the null hypothesis is rejected, with significant components indicating particular \( \theta_i \) positive, then \( \{\pi_j\} \) above specifies an alternative model. This formulation is investigated in Rayner (1986).

REFERENCES


Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen by random sampling, *Philos.Mag.*, 50, 157-175.


Neyman-type smooth tests for location-scale families

BY J. C. W. RAYNER

Department of Mathematics and Statistics, University of Otago, Dunedin,
New Zealand

AND D. J. BEST

CSIRO, Division of Mathematics and Statistics, North Ryde, N.S.W. 2113,
Australia

SUMMARY

The smooth tests of Neyman (1937) are modified to permit composite alternatives and a choice of orthonormal functions in their construction. These orthonormal functions may be chosen to improve detection of particular alternatives. The tests are easy to apply and have useful orthogonal components. A test for exponentiality is examined, and shown to perform well against its competitors.

Some key words: Complete orthonormal basis; Power; Simple and composite hypothesis; Smooth goodness-of-fit test; Test for exponentiality.
Goodness-of-fit for grouped data using components of Pearson’s $X^2$

D.J. BEST
CSIRO, Division of Mathematics and Statistics, P.O. Box 52, North Ryde, NSW 2113, Australia

J.C.W. RAYNER
Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand

Received 24 April 1986
Revised 22 September 1986

Abstract: Suppose we wish to test whether observations come from some specified distribution when, perhaps, only grouped data exists. Components of $X^2$ are shown, in a small power study, to perform well.

Keywords: Categorized data, Discrete data, Tests of fit, Power comparison, Chi-square tests, Cressie–Read statistics.
CONSTRUCTING A USABLE OVERLAPPING CELLS $X^2$ GOODNESS OF FIT TEST

J.C.W. RAYNER
Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand

Received March 1986
Revised May 1987

Abstract: A smooth alternative is proposed for a multinomial model and is used to construct, via score statistics, a class of $X^2$ tests. Following Hall (1985), a subclass of overlapping cells tests is extracted from this class.

Keywords: components, Pearson's $X^2$ test, score, size, statistic.

A TEST FOR BIVARIATE NORMALITY

D.J. BEST
CSIRO Division of Mathematics and Statistics, P.O. Box 52, North Ryde, NSW, 2113, Australia

J.C.W. RAYNER
Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand

Received June 1987
Revised September 1987

Abstract: A test of multivariate normality given by Koziol (1986, 1987) is examined in some detail for the bivariate case. The small sample null distribution is considered and power comparisons given. Examples are given to illustrate the use of components of the test statistic.

Keywords: goodness of fit, power comparisons, multivariate skewness and kurtosis.

Article removed for copyright reasons, please see print copy for the article:
SMOOTH TESTS OF GOODNESS OF FIT FOR
REGULAR DISTRIBUTIONS

J.C.W. Rayner and D.J. Best

Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand.

CSIRO Division of Mathematics and Statistics, North Ryde, NSW, Australia.

Key Words and Phrases: components; multivariate normal; orthonormal functions; Poisson; score statistic.

ABSTRACT

Smooth tests of goodness of fit based on orthonormal functions for location-scale families were introduced in Rayner and Best (1986). This paper extends this class of tests from location-scale families to 'regular' families. The extension preserves the desirable properties of the class, such as weak optimality, accessible components and convenient distribution theory.

The Goodness of Fit Publications

of

J.C.W. Rayner

Volume 2
Foreword

This thesis has two parts. The first is an overview, "The Development of the Smooth Tests of Goodness of Fit", of the submitted work. This overview is relatively brief, being approximately 5000 words long. The second part consists of the submitted publications themselves. These are listed below by year of publication, and alphabetically within year.

The bulk of the thesis has necessitated it being bound in two volumes. The first volume consists of the overview and publications [1] to [18]. The second volume consists of publications [19] to [25].

The Goodness of Fit Publications of J.C.W. Rayner


Smooth Tests of Goodness of Fit

J. C. W. RAYNER
Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand

D. J. BEST
CSIRO, IAPP Biometrics Unit, Food Research Laboratory, North Ryde, Australia

New York Oxford
OXFORD UNIVERSITY PRESS
1989
J. C. W. R.: To the Rayners, Fred, Carol, Glen and Eric. And my friend and colleague the first fleeter.

D. J. B.: To my parents and to Helen, Warwick and Rohan.
The importance of probabilistic or statistical modeling in the modern world cannot be overrated. With the advent of high-speed computers, complex models for important processes can now be constructed and implemented. These models and the associated statistical analyses are of great assistance in making decisions in diverse fields, from marketing, medicine, and management, to politics, weapons systems, and food science. Goodness of fit is concerned with assessing the validity of models involving statistical distributions, an essential and sometimes forgotten aspect of the modeling exercise. One can only speculate on how many wrong decisions are made due to the use of an incorrect model.

Karl Pearson pioneered goodness of fit in 1900, when his paper introducing the $X^2$ test appeared. Since then, perhaps reflecting the needs and importance of the subject, a great many new tests have been constructed. The smooth tests are a class of goodness of fit tests that are informative, easy to apply, and generally applicable. Typically they can be derived as score tests, and hence are, in a sense, optimal for large sample sizes. For moderate and small samples they are very competitive in the cases we have examined. Pearson’s $X^2$ test is in fact a smooth test. In the formulation we prefer, components with simple graphic interpretations are readily available. We suggest that the properties of the smooth tests are such that a new goodness of fit test must be in some way superior to a corresponding smooth test if it is to be worthy of consideration.

This book is complementary to that by D’Agostino and Stephens (Goodness of Fit Techniques, 1986) in that they do not cover the smooth tests in any detail, while we do not cover in detail topics such as tests based on the empirical distribution function, and tests based on regression and correlation. There is some overlap in the coverage of $X^2$ tests. The tests that they discuss and our smooth tests are in competition with each other. We give some comparisons, and, not surprisingly, recommend use of the smooth tests. Usually, the smooth tests are more informative than their
competitors. The D'Agostino and Stephens book covers a broad range of topics, generally omitting mathematical details and including many tables and examples so that it reads as a handbook of methods. Since our book concentrates on smooth methods, we have been able to present derivations and mathematical details that might have been omitted in a more comprehensive treatment of goodness of fit in its entirety. We consider this to be highly desirable because the development of the smooth tests of fit is far from complete. Indeed, we hope that researchers will read this book and be motivated to help with its further development.

In spite of this last statement, many economists, scientists and engineers who have taken an introductory mathematical statistics course, to the level of Larsen and Marx (1981), will be able to read this book. The more technical details are clearly signposted and are in sections that may be omitted or skimmed. Undergraduates with a sufficient background in statistics and calculus should be able to absorb almost everything. Practical examples are given to illustrate use of the techniques.

The smooth tests for the uniform distribution were introduced by Neyman (1937), but they were slow to gain acceptance because the computations are heavy by hand. This is no longer a barrier. Many of the techniques we discuss are readily implemented on modern computers, and we give some algorithms to assist in doing this. When used in conjunction with density estimate plots or Q–Q plots, the smooth tests can play an important part in many analyses.

Chapter 1 outlines the goodness of fit problem, gives a brief history of the smooth tests, outlines the monograph, and gives some examples of the sort of problems that arise in practice. A review of Pearson (1900), and an outline of the early development of the tests for simple and composite hypotheses is given in Chapter 2. In Chapter 3, tests that are asymptotically optimal are introduced; these include the score tests that are particularly important later in the book. Using score tests and smooth models, tests of completely specified null hypotheses are derived in Chapters 4 and 5. These chapters cover both uncategorized (discrete or continuous) and categorized the null distributions. The tests are essentially tests for uniformity. Then, in Chapters 6 and 7, we consider tests for composite null hypotheses, again treating both the categorized and uncategorized cases. Chapters 4 to 7 emphasize the components our tests yield. In Chapter 6 we look at tests for the univariate and later for the multivariate normal, the Poisson, geometric and exponential distributions. These are extracted from a class of smooth goodness of fit tests. In Chapter 7, we discuss $X^2$ statistics for composite hypotheses. We conclude with a review and an examination of some of the other uses to which our techniques may be put.

Our interest in the subject of goodness of fit came about partly from questions from colleagues relating to the "Gaussian" assumption in routine statistical analyses, and partly from work J. C. W. R. had begun as a student of H. O. Lancaster. Our approach is based on the use of orthonormal functions, emphasized in Lancaster (The Chi-Squared

Dunedin, New Zealand
North Ryde, Australia

J. C. W. R.
D. J. B.
We would like to thank many friends and colleagues for their generous contributions, ranging from encouragement and moral support, to proofreading, typing, and technical advice. In particular, we would like to thank, in alphabetical order, Mary-Jane Campbell, Peter Diggle, Geoff Eagleson, Lynette Grant, Derek Holton, Fred Lam, Bryan Manly, Lyn McAlevey, Tony Pettitt, Allyson Seyb, and Terry Speed.

In addition a grant made by the Otago University Research Committee is gratefully acknowledged.
1 Introduction, 3

1.1 The Problem Defined, 3
1.2 A Brief History of Smooth Tests, 6
1.3 Monograph Outline, 10
1.4 Examples, 12

2 Pearson's $X^2$ Test, 20

2.1 Introduction, 20
2.2 Foundations, 20
2.3 The Pearson's $X^2$ Test—an Update, 22
   2.3.1 Notation, definition of the test, and class construction, 23
   2.3.2 Power related properties, 25
2.4 $X^2$ Tests of Composite Hypotheses, 27
2.5 Examples, 28

3 Asymptotically Optimal Tests, 33

3.1 Introduction, 33
3.2 The Likelihood Ratio, Wald, and Score Tests for a Simple Null
    Hypothesis, 34
3.3 The Likelihood Ratio, Wald, and Score Tests for Composite Null
    Hypotheses, 37
3.4 Properties of the Asymptotically Optimal Tests, 41

4 Neyman Smooth Tests for Simple Null Hypotheses, 45

4.1 Neyman's $\Psi^2$ Test, 45
4.2 Neyman Smooth Tests for Uncategorized Simple Null Hypotheses, 48
4.3 Effective Order and Power Comparisons, 51
4.4 Examples, 52
5 Neyman Smooth Tests for Categorized Simple Null Hypotheses, 57
5.1 Smooth Tests for Completely Specified Multinomials, 57
5.2 $X^2$ Effective Order, 61
5.3 Components of $X^2_p$, 63
5.4 Examples, 66
5.5 A More Comprehensive Class of Tests, 71
5.6 Overlapping Cells Tests, 73

6 Neyman Smooth Tests for Uncategorized Composite Null Hypotheses, 77
6.1 Neyman Smooth Tests for Uncategorized Composite Null Hypotheses, 77
6.2 Smooth Tests for the Univariate Normal Distribution, 81
6.3 Smooth Tests for the Exponential Distribution, 87
6.4 Smooth Tests for the Poisson Distribution, 91
6.5 Smooth Tests for the Geometric Distribution, 96
6.6 Smooth Tests for the Multivariate Normal Distribution, 98
6.7 Components of the Rao–Robson $X^2$ Test, 109

7 Neyman Smooth Tests for Categorized Composite Null Hypotheses, 111
7.1 Neyman Smooth Tests for Composite Multinomials, 111
7.2 Components of the Pearson–Fisher Statistic, 115
7.3 Composite Overlapping Cells and Cell Focusing $X^2$ Tests, 116
7.4 A Comparison Between the Pearson–Fisher and Rao–Robson $X^2$ Tests, 121

8 Conclusion, 124
8.1 Introduction, 124
8.2 Further Applications of Smooth Tests of Fit, 124
8.2.1 Density estimation, 124
8.2.2 Outlier detection, 129
8.2.3 Normality testing of several samples, 131
8.2.4 Residuals, 134
8.3 Concluding Example, 135
8.4 Closing Remarks, 137

Appendix 1 Orthogonal Polynomials, 140
Appendix 2 Computer Program for Implementing the Smooth Tests of Fit for the Uniform, Normal, and Exponential Distributions, 142
Appendix 3 Explicit Formulas for the Components $V_1$, and $V_2$ of $X^2$, 147

References, 148
Author Index, 157
Papers Index, 160
Smooth Tests of Goodness of Fit

The following chapters of this book removed for copyright reasons, please see print
copy for the all these chapters.
Goodness of Fit for the Geometric Distribution

D. J. Best
C.S.I.R.O. IAPP Biometrics Unit, North Ryde, Australia
and
J. C. W. Rayner
Department of Mathematics & Statistics, University of Otago, New Zealand

Summary
A Neyman-type smooth test of goodness of fit is derived for the geometric distribution. Some small-sample critical points are given, with two examples.

Key words: Components; Meixner orthonormal polynomials; Sizes; Smooth model.

Article removed for copyright reasons, please see print copy for the article:
Smooth Tests of Goodness of Fit: An Overview

J. C. W. Rayner¹ and D. J. Best²

¹Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand. ²CSIRO, IAPP Biometrics Unit, PO Box 52, North Ryde, NSW 2113, Australia

Summary

A comprehensive approach to goodness of fit testing is possible using the smooth tests described in detail in Rayner & Best (1989). Here we give an overview of this area and demonstrate the power and flexibility of the smooth tests. Our emphasis is on the use of orthonormal functions, tests for composite hypotheses, and tests of categorised data. We have developed tests for families of distributions, such as the univariate and multivariate normal, exponential and Poisson. The tests are essentially omnibus tests but the components provide useful and powerful directional tests. The history of the smooth tests of goodness of fit is reviewed from Neyman (1937), through to Lancaster, to Thomas, Kopecky and Pierce, and to Rayner and Best. The formulation of categorised smooth models leads to $\chi^2$ tests and their components. A generalisation of the smooth categorised model, when allied with Hall's (1985) idea of overlapping, leads to focused tests, and to an alternative to pooling. Examples are taken from D'Agostino & Stephens (1986), who have several different contributors and therefore approaches, none of which is recommended above the others. Our resolution is simple: don't use those other methods—use a smooth test!

Key words: Categorised data; Components; Composite hypotheses; Omnibus and directional tests; Orthonormal functions; Score tests.
SMOOTH GOODNESS OF FIT TESTS FOR CATEGORISED COMPOSITE NULL HYPOTHESES

J.C.W. RAYNER
Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand

L.G. McALEVEY
Department of Quantitative and Computer Studies, University of Otago, P.O. Box 56, Dunedin, New Zealand

Received April 1988
Revised May 1989

Abstract: In this paper we construct a class of smooth goodness of fit tests of a composite multinomial null hypothesis. This class includes the Pearson–Fisher $X^2$ test and its components, composite extensions of Rayner's (1987) overlapping tests and cell focusing tests. Simulations are given for a Poisson null.

Keywords: Components, goodness of fit, Pearson–Fisher $X^2$ test, score test, size, smooth alternatives.

Article removed for copyright reasons, please see print copy for the article:
Disease Clustering in Time

D. J. Best
CSIRO, IAPP Biometrics Unit,
P.O. Box 52, North Ryde 2113, Australia

and

J. C. W. Rayner
Department of Mathematics and Statistics, University of Otago,
P.O. Box 56, Dunedin, New Zealand

SUMMARY
Tango (1984, *Biometrics* 40, 15–26) proposed a clustering index for testing for clusters of disease in time. A test based on this clustering index was shown to compare favourably with other statistical tests. In this article we show that Pearson's $X^2$ and its components perform well in testing for clusters of disease in time. The $r$th of these components identifies a departure from uniformity in moments up to the $r$th and so helps describe the alternative, if any, to uniformity.

---

THE COMPARISON OF USAGE AND AVAILABILITY MEASUREMENTS USING COMPONENTS OF PEARSON'S GOODNESS OF FIT STATISTIC

J.C.W. RAYNER

Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand.

In the same way that contrasts are used in the analysis of variance, components may be used with Pearson's goodness of fit test. The use of these components to make the Pearson test more informative is demonstrated. A fundamental data analysis problem is also discussed, again with the assistance of components.

1. Introduction

Neu et al. (1974) studied moose distributions in a study area centred on the Little Sioux burn. The study area was divided into four regions, as indicated in Table 1. Pearson's test was used to test the hypothesis that:

“moose utilize each habitat category in exact proportion to its occurrence within the study area.”

Table 1 Little Sioux Burn data

<table>
<thead>
<tr>
<th>#</th>
<th>Region</th>
<th>Acreage</th>
<th>Moose observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Burn interior</td>
<td>11283</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>Burn periphery</td>
<td>3345</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>Unburnt periphery</td>
<td>3442</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>Unburnt exterior</td>
<td>15130</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>33200</td>
<td>117</td>
</tr>
</tbody>
</table>

Equivalently, the hypothesis is that the expected number of moose in each region is proportional to the acreage of that region. Neu et al. (1974) say that:
"The chi-square does not determine preference or avoidance of individual categories, so that the data must be inspected to determine which observations contribute most to the calculated chi-square value, and whether a specific type is preferred or avoided."

Rather than subjectively inspecting the contributions of the cells to the value of the test statistic, an objective assessment can be made. The test statistic is partitioned into (asymptotically orthogonal) components or contrasts. The technique is well known in the analysis of variance. Although partitioning chi-squared has a substantial literature dating back to Irvin and Lancaster, it is an underutilized technique. The approach in Rayner and Best (1989) demonstrates that the components are score statistics and thus are weakly optimal. Although here the cell expectations in Pearson's test are assumed to be known, a parallel (and optimal) analysis is possible if the expectations must be estimated. See, for example Rayner and Best (1990, Example 5.2) and Rayner and McAlevey (1990, pp425-426).

Motivated by the Neu data, we will also look at a perplexing data analysis problem. It may well be that in one category the analyst knows a priori that the null hypothesis does not hold. Should the data for that category be ignored? The other categories may perform as the null hypothesis suggests, but the deviant cell could cause the null hypothesis to be rejected. The use of components enables us to clarify this problem.

2. Components of Pearson’s $\chi^2$ Statistic

Suppose we have $m$ cells or categories, and $N_1, ..., N_m$ count the number of observations in the $m$ cells. Under the null hypothesis the probability of an observation in the $j$th cell is specified to be $p_j$, $j = 1, ..., m$. Pearson’s chi-squared test uses the statistic

$$\chi^2 = \sum_{j=1}^m (N_j - np_j)^2 / (np_j),$$

where $n = N_1 + ... + N_m$. If $\chi^2$ is greater than the 100\% point of the $\chi^2_{m-1}$ distribution, the null hypothesis that the cell probabilities are $p_1, ..., p_m$ is rejected at the 100\% level. Perhaps the most commonly encountered null hypothesis of this type is that the cells are equiprobable: $p_j = 1/m$, $j = 1, ..., m$. For the Neu data the $p_j$ are given by the relative areas of the different regions.

Components of $\chi^2$ were defined in Rayner and Best (1989). Put

$$V_r = \sum_{j=1}^m h_{nj} N_j / \sqrt{n}, \quad r = 1, ..., m-1,$$

where the $h_{nj}$ satisfy

$$\sum_j h_{nj} h_{sj} p_j = 0 \text{ for } r \neq s,$$

$$= 1 \text{ for } r = s, \text{ and}$$

$$\sum_j h_{nj} p_j = 0 \text{ for } r = 1, ..., m-1. \quad (1)$$

51
There are m−1 components, one for each degree of freedom. These components are analogous to the contrasts used for treatment means in the analysis of variance. If the null hypothesis is true, the components are asymptotically independent and asymptotically have the standard normal distribution. Provided the conditions (1) are satisfied, \( X_p^2 = V_1^2 + \ldots + V_{m-1}^2 \).

Lancaster (1969, §4.3) suggested, in another context, that the \( h_r \) be based on a Helmert matrix. This leads to components

\[
V_r^* = \left\{ \frac{N_{r+1}}{n_{r+1}} - \frac{(N_1 + \ldots + N_r)}{n_1 + \ldots + n_r} \right\} \sqrt{\frac{n(n_1 + \ldots + n_r)}{(n_1 + \ldots + n_r) + n_{r+1}}}, \quad r = 1, \ldots, m-1;
\]

where \( V_r^* \) compares the \( r+1 \)th cell with the preceding \( r \) cells.

Of course, many sets of components of \( X_p^2 \) may be useful in different contexts. This comes down to a choice of \( \{h_r\} \). However it may be easier to interpret the data if instead of the components being linear combinations of the \( \{N_j\} \), they are defined in terms of some other random variables. For example, consider components of the form

\[
W_r = \sum_{j=1}^{m} a_j X_j / \sqrt{n}, \quad r = 1, \ldots, m-1,
\]

where \( X_j = N_j / (n_j) \) is the relative excess of the \( j \)th observed cell count over the \( j \)th expected cell count, \( j = 1, \ldots, m \). For data analytic purposes, the relative excesses may suggest sets of components. The relative excesses are easier to interpret if they are standardized so that they sum to one. So define \( \beta_j = X_j / \sum X_j, \quad j = 1, \ldots, m \). Similar values of \( \beta \) suggest similar cell proportions. Interestingly, the \( \beta \) are the maximum likelihood estimates of the probability that the next moose will be in habitat type \( j \), when the selection process has a choice of an equal number of individuals in each of the \( m \) classes. See Manly (1974). The \( W_r \) defined in (3) must satisfy conditions parallel to (1) if \( X_p^2 \) is to be partitioned.

For testing purposes the comparisons to be made must be decided before sighting the data. More than one set of components may be defined, but components in different sets are not independent. This complicates inferences.

3. **Components in the Neu Data**

Routine calculation for the Neu data leads to \( X_p^2 = 43.98 \), which is significant at all normally used levels of significance. The Helmertian components (2) compare regions 1 and 2, then region 3 with the average of regions 1 and 2, and finally region 4 with the average of the first three regions. The values of these components are \( V_1^* = 3.73 \), \( V_2^* = 4.89 \) and \( V_3^* = -2.47 \). These are all significant at the 1% level. If "different" means "the moose usages per unit acreage are different", then the burn interior is different from the burn periphery; the unburnt periphery is different from the burn, and the unburnt exterior is different from the regions interior to it.

But are these the comparisons that are of interest? Suppose that before sighting the data we had decided to compare the periphery and non-periphery areas both within
and outside the burn, and also to compare the burn and unburnt areas. The appropriate components are

\[
\begin{align*}
V_1^{**} &= \left[ \frac{N_1}{np_1} - \frac{N_2}{np_2} \right] \sqrt{\frac{np_1p_2}{(p_1+p_2)}}, \\
V_2^{**} &= \left[ \frac{N_3}{np_3} - \frac{N_4}{np_4} \right] \sqrt{\frac{np_3p_4}{(p_3+p_4)}}, \quad \text{and} \\
V_3^{**} &= \left[ \frac{N_1+N_2}{np_1+np_2} - \frac{N_3+N_4}{np_3+np_4} \right] \sqrt{\frac{np_1+np_2}{(p_1+p_2)(p_3+p_4)}}.
\end{align*}
\]

We obtain \(V_1^{**} = -3.73, V_2^{**} = 5.42\) and \(V_3^{**} = -0.85\). The first two are significant at all normally used levels of significance, while the third is not. Apparently the burn and unburnt areas are used similarly by the moose, but both within and outside the burn the fit shown on the periphery is different from that on the non-periphery.

The standardized relative excesses are \(\beta_1 = 0.11, \beta_2 = 0.33, \beta_3 = 0.43\) and \(\beta_4 = 0.13\). This suggests the periphery regions are more heavily utilized than the non-periphery regions. Had this been suggested by a priori considerations, it could be tested by using the components

\[
\begin{align*}
V_1^{**} &= \left[ \frac{N_1}{np_1} - \frac{N_4}{np_4} \right] \sqrt{\frac{np_1p_4}{(p_1+p_4)}}, \\
V_2^{**} &= \left[ \frac{N_2}{np_2} - \frac{N_3}{np_3} \right] \sqrt{\frac{np_2p_3}{(p_2+p_3)}}, \quad \text{and} \\
V_3^{**} &= \left[ \frac{N_1+N_2}{np_1+np_2} - \frac{N_3+N_4}{np_3+np_4} \right] \sqrt{\frac{np_1+np_4}{(p_1+p_4)(p_3+p_4)}}.
\end{align*}
\]

Calculation gives \(V_1^{**} = -0.58, V_2^{**} = -1.48\) and \(V_3^{**} = -6.44\). We conclude that the periphery and non-periphery are utilized differently; and that within each region, the burn and unburnt regions are used similarly.

4. **Discarding Information**

Consider the situation in §2. Denote the chi-squared test statistic by \(\chi^2_p(m,n)\) and the components (2) by \(V_r^*(m,n)\). In both cases, the arguments \(m\) and \(n\) are included to emphasize the dependence on the number of categories \(m\) and the number of observations \(n\). Suppose we know that cells \(k+1, \ldots, m\) do not follow the hypothesised probabilities. If the information in these cells is discarded, we apply the chi-squared test with \(N = n_1 + \ldots + n_k\) observations and cell probabilities \(P_r = p_r/P\), where \(P = p_1 + \ldots + p_k\), and \(r = 1, \ldots, k\). If we just have a strong a priori belief that cells \(k+1, \ldots, m\) do not follow the hypothesised probabilities, which test do we use; that based on \(\chi^2_p(m,n)\), or that based on \(\chi^2_p(k,N)\)? There is clearly a danger that the later cells may reverse a nonsignificant result based on the earlier cells. On the other hand it seems counter to a statistician's natural instincts to ignore some of the data.

It is routine to show that

\[
V_r^{*(k,N)} = V_r^{*(m,n)} \sqrt{\frac{nP}{N}}, \quad r = 1, \ldots, k-1.
\]

If the cell probabilities for the first \(k\) cells are \(p_1, \ldots, p_k\), then the factor \(\sqrt{\frac{nP}{N}}\) will be close to one (it will approach one in probability) and the \(V_r^{*(k,N)}\) will be approximately independent and standard normal in large samples, as will the \(V_r^{*(m,n)}\). Use of either
set of components is acceptable. Since the numerically larger will be more critical of the data, the size of \( \sqrt{nP/N} \) could be applied to determine which components to use.

If the cell probabilities for the first \( k \) cells are not \( p_1, \ldots, p_k \), the distributions of both the \( V_r^*(k,N) \) and the \( V_r^*(m,n) \) will not be approximately independent and standard normal in large samples. Since both sets of components should reject the null hypothesis, it seems reasonable to retain the criterion of the preceding paragraph.

There are two attitudes we could sensibly adopt. Since the numerically larger components will be more critical of the data if \( nP/N > 1 \), use the \( V_r^*(k,N) \) in preference to the \( V_r^*(m,n) \), \( r = 1, \ldots, k \). Otherwise use the \( V_r^*(m,n) \). On the other hand, some users will be more comfortable with the reverse rule, which adopts a conservative policy.

The factor \( \sqrt{nP/N} \) could, via (4), be determined as the quotient of the different components. So that we can determine when \( \sqrt{nP/N} \) is statistically different from 1, we should then ask what the distribution of \( \sqrt{nP/N} \) is? A less formal approach seems desirable, and will be motivated by twice adjusting the Neu data so that the null hypothesis is true with respect to only the first three regions. In the first case, given in Table 2(i), the fourth region, the unburnt exterior, is rarely used.

### Table 2 Adjusted Little Sioux Burn data

(i) Rarely used fourth region

<table>
<thead>
<tr>
<th>Region</th>
<th>Acreage</th>
<th># Moose observed</th>
<th>( \beta_i )</th>
<th>( \beta_i^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burn interior</td>
<td>11285</td>
<td>28</td>
<td>0.353</td>
<td>0.359</td>
</tr>
<tr>
<td>Burn periphery</td>
<td>3345</td>
<td>8</td>
<td>0.340</td>
<td>0.346</td>
</tr>
<tr>
<td>Unburnt periphery</td>
<td>3442</td>
<td>7</td>
<td>0.289</td>
<td>0.294</td>
</tr>
<tr>
<td>Unburnt exterior</td>
<td>15130</td>
<td>2</td>
<td>0.019</td>
<td></td>
</tr>
</tbody>
</table>

33200 45

(ii) Highly used fourth region

<table>
<thead>
<tr>
<th>Region</th>
<th>Acreage</th>
<th># Moose observed</th>
<th>( \beta_i )</th>
<th>( \beta_i^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burn interior</td>
<td>11283</td>
<td>28</td>
<td>0.044</td>
<td>0.359</td>
</tr>
<tr>
<td>Burn periphery</td>
<td>3345</td>
<td>8</td>
<td>0.042</td>
<td>0.346</td>
</tr>
<tr>
<td>Unburnt periphery</td>
<td>3442</td>
<td>7</td>
<td>0.036</td>
<td>0.294</td>
</tr>
<tr>
<td>Unburnt exterior</td>
<td>200</td>
<td>10</td>
<td>0.879</td>
<td></td>
</tr>
</tbody>
</table>

18270 53
The components $V_r(4,45)$ are $-0.124$, $-0.613$ and $-5.540$ for $r = 1$, $2$, $3$ respectively, while the components $V_r(3,43)$ are $-0.094$ and $-0.462$ for $r = 1$ and $2$ respectively. These agree with (4). We obtain $X_p^2$ as 31.078 using all four regions and 0.223 using just the first three. The relative excesses omitting the fourth region are 0.353, 0.340 and 0.290. These are almost identical to the $\beta_1^*$, the relative excesses for the first three regions relative to each other, namely 0.359, 0.346 and 0.294.

We now consider the second modification of the Neu data, with the fourth region highly used. See Table 2(ii). Using all four regions $X_p^2$ is 154.814, but just 0.223 using the first three. The components $V_r(3,43)$ are $-0.093$ and $-0.462$ while the $V_r(4,45)$ are $-0.085$, $-0.419$ and $12.435$ for $r = 1$, $2$, $3$ respectively. The relative excesses omitting the fourth region are 0.359, 0.346 and 0.294. These are identical to three decimals to the $\beta_1^*$, the relative excesses for the first three regions relative to each other.

The message coming from these calculations is that it is the $\beta_i$ that will alert us to regions behaving deviantly, if we were not already aware! Then, if the comparisons and thus the components have been chosen appropriately, either the $V_r^*(k,N)$ or the $V_r^*(m,n)$ may be used. This depends on whether we chose to be conservative or as critical of the data as is possible. This choice will not normally be vital in the analysis. On the other hand, the choice of comparisons and therefore components is. To underscore the message, with the unadjusted Neu data, none of the relative excesses are excessively large.

In summary, a closer scrutiny of the data is possible, if, in addition to $X_p^2$, these suggested components are used. The relative excesses are a valuable additional exploratory tool.

Acknowledgements

The author wishes to thank Lyman McDonald, formerly of the University of Wyoming, for suggesting some of the problems considered here, and for reading and suggesting improvements to the manuscript.

References


S-Sample Smooth Goodness of Fit Tests

J.C.W RAYNER and G.D. RAYNER
Department of Mathematics and Statistics
University of Otago
PO Box 56, Dunedin, New Zealand

SUMMARY
In a development parallel to that for the one sample problem, S-sample smooth goodness of fit tests are introduced. These tests are omnibus tests and have directional components, all based on score tests which thus are weakly optimal for specifiable alternatives. Testing may proceed on a hierarchical basis. We test if the first two samples are consistent with each other, and then if the s-th, s = 3, ..., S, sample is consistent with its predecessors. The test statistics are given in terms of the one sample components for each sample. Then, if all samples are consistent with having come from the same population, we can conveniently perform a one-sample test for that distribution. If the samples are not consistent, an LSD-type analysis can be performed on the one-sample components to identify where the differences occur. Our procedure requires the specification of a target distribution. The subjectivity of the choice of this distribution is balanced by the wealth of information the procedure provides.

Some key words: Anderson-Darling test; components; directional test; omnibus test; orthonormal polynomial; power; size.

1. Introduction
In the S-sample goodness of fit problem we wish to test if S random samples may be considered to come from the same population. Typically that population is not initially specified, although ultimately it may be. There are several tests in the literature that are limited in some way. Tiku (1980), Foutz and Birch (1982), Brunner and Neumann (1986) and Lemmer (1987) all discuss two-sample tests. These and other authors often use rank tests, although the most successful approach so far has probably been to generalise the one sample goodness of fit tests based on the empirical distribution function. See for example Kiefer (1958) and Eplett (1982). Conover (1980) gave a users' treatment. Pettitt's (1976) two-sample Anderson-Darling test was extended by Scholz and Stephens (1987) to the S-sample case. This test is taken to be the standard against which we compare the tests developed here.
It is surprising that there haven't been more attempts to generalise the one-sample approaches, such as the tests based on regression and correlation. Such one sample tests have been discussed in D'Agostino and Stephens (1986). The following development extends to the S-sample case the advantages of the one sample smooth goodness of fit tests, discussed, for example, in Rayner and Best (1989). These advantages include the following:

(i) a wide class of tests is available;
(ii) these tests are score tests and hence are weakly optimal;
(iii) both omnibus and directional tests are available;
(iv) the test statistics have convenient null and alternative distributions and are easily interpreted.

Moreover the approach yields as an intermediate step the statistics from which certain one sample smooth tests can be conducted, and LSD-type comparisons made.

The approach adopted here is that a two-stage procedure may be most informative. Initially we test if the data are consistent with coming from the same population: the S-sample problem. If so, subsequent testing for a specified population, a one-sample goodness of fit problem, may be appropriate. A typical situation is a two-sample location problem. A two-sample goodness of fit test assesses if both samples can be regarded as coming from the same population. If so, we ask if the combined samples are consistent with the normal distribution, a one-sample goodness of fit problem. A positive answer leads to the pooled t-test or Welch's test; a negative answer suggests use of the Wilcoxon test. However if both samples cannot be regarded as coming from the same population, this may be more informative than the initial location test required. A Wilcoxon test may still be applied, but the location question may be too simplistic. Perhaps density estimates are called for. The question "why has the model (distribution) changed?" needs to be addressed.

As with the one-sample smooth tests, the use of components permits a close scrutiny of the data. So for example, we could optimally test whether the samples agree in their location, dispersion, skewness and kurtosis measures. As a corollary we may then test whether these samples, individually or collectively, are consistent with a specified distribution such as the exponential or the normal. An omnibus assessment bases the decision on moments up to specified order \( r \); the directional components gives information on each moment. If the samples are not consistent with each other, an LSD-type analysis will determine the nature of the differences between the populations. Again this analysis gives information on the differences between the samples for each moment up to order \( r \). That we can give much more than a single P-value in assessing the S-sample goodness of fit problem, is surely a considerable strength of this approach.
Figure 1  LSD comparisons of four samples using Table 5(a) data.

Order One Comparison

\[
\begin{array}{cccc}
P_4 & P_3 & P_2 & P_1 \\
-2.53 & -1.51 & -1.08 & 0.23
\end{array}
\]

Order Two Comparison

\[
\begin{array}{cccc}
P_2 & P_1 & P_3 & P_4 \\
-2.50 & -1.94 & -1.62 & -0.25
\end{array}
\]

Order Three Comparison

\[
\begin{array}{cccccc}
P_1 & P_3 & P_2 & P_4 \\
0.80 & 1.60 & 2.06 & 2.16
\end{array}
\]

Order Four Comparison

\[
\begin{array}{cccc}
P_4 & P_3 & P_2 & P_1 \\
-1.89 & 0.48 & 1.27 & 1.88
\end{array}
\]

In the two-sample situation, we wish to test if two random samples \(X_{s1}, \ldots, X_{sn_s}, \ s = 1 \text{ and } 2\), may be considered to come from the same population. Although there is no obvious a priori specification for a probability density function \(f(x)\), we seek to model the probability density functions for each population by functions of the form (2.1), and then test if the corresponding parameters are the same, in which case we can conclude that the populations are consistent with each other. If the current testing problem is preliminary in some sense to assessing if the two populations have the same specified probability density function, then that specified probability density function will be used in our model. Otherwise we seek a target probability density function that models our populations by order \(k\) probability density functions with \(k\) not large: hopefully at most four. Our expectation from the one-sample case is that this limitation on \(k\) will lead to more powerful procedures. With \(k\) large enough, the data can be modelled exactly. In practice we find that few densities project substantially into more that four dimensions: see Rayner Best and Dodds (1985). However, if the target is poorly
chosen, then more than order four differences will need to be assessed. Reduced power will result, and this is the cost of poor targeting.

We model the $s$th probability density function by

$$C_s(\xi_{1s}, \ldots, \xi_{ss}) \exp\left[\sum_{i=1}^{k} \xi_{si} h_i(x)\right] f(x), s = 1, \ldots, S.$$ 

Although the $\{h_i(x)\}$ need not be orthonormal functions in general, here we will make that assumption. Were the $f(x)$ assumed to be uniform, the $\{h_i(x)\}$ could usually be taken to be the Legendre polynomials, or $\{\sqrt{2}\cos(ntj)\}$, or some other functions that will enable representation of the data using relatively few terms. If the populations are not uniform, then some $\xi_{si}$ will be non-zero.

In one sample goodness of fit testing, we wish to test if all the parameters in the order k alternative are zero. If the problem was specifying a parsimonious model, we would seek an $f(x)$ so that as few $\xi_{si}$ as possible were needed to describe the data. For the two-sample goodness of fit problem, we want to show that $\xi_1 = (\xi_{11}), \xi_2 = (\xi_{21}).$ In many practical situations, if the populations are consistent with each other, it will also be of interest to know if they are consistent with the target probability density function. Having found that $\xi_1 = \xi_2$, we would then ask if $\xi_1 = \xi_2 = 0$, corresponding to a one sample goodness of fit testing problem.

3 Derivation of the Two-sample No Nuisance Parameter Test Statistic

We now derive the score test of $H: \xi_1 = \xi_2$ against $K: \xi_1 \neq \xi_2$ given two random samples, $X_{s1}, \ldots, X_{s_n}, s = 1$ and 2. Abbreviating meaningfully, the likelihood is

$$C_1^{n1}(\xi_1) \exp\left[\sum_{i=1}^{k} \xi_{1i} \sum_{j=1}^{n1} h_i(x_{1j})\right] \prod_{j=1}^{n1} f(x_{1j}) C_2^{n2}(\xi_2) \exp\left[\sum_{i=1}^{k} \xi_{2i} \sum_{j=1}^{n2} h_i(x_{2j})\right] \prod_{j=1}^{n2} f(x_{2j}).$$

Write $G_{sr} = \sum_{r=1}^{k} h_i(x_{sr})$, for $s = 1, 2$ and $r = 1, \ldots, k \leq n - 1$, $f_{X_1} = \prod_{j=1}^{n1} f(x_{1j})$ and $f_{X_2} = \prod_{j=1}^{n2} f(x_{2j})$. The logarithm of the likelihood is then

$$\ell = n_1 \ell n C_1 + n_1 \xi_{1i} G_{1i} + n_1 \ell n f_{X_1} + n_2 \ell n C_2 + n_2 \xi_{2i} G_{2i} + n_2 \ell n f_{X_2} + \ell.$$ 

Now reparametrize by putting $\theta = \xi_1 - \xi_2 = (\theta_i)$ and $\beta = \xi_1 + \xi_2 = (\beta_i)$. Then $\xi_1 = (\theta + \beta)/2$ and $\xi_2 = (\beta - \theta)/2$. We wish to test $H: \theta = 0$ against $H: \theta \neq 0$ in the presence of the vector of nuisance parameters $\beta$. 
Differentiation yields

\[
\frac{\partial l}{\partial \theta_r} = \frac{n_1}{2} \frac{\partial \ln C_1}{\partial \xi_{1r}} + \frac{G_{1r}}{2} - \frac{n_2}{2} \frac{\partial \ln C_2}{\partial \xi_{2r}} + \frac{G_{2r}}{2},
\]

\[
\frac{\partial l}{\partial \beta_r} = \frac{n_1}{2} \frac{\partial \ln C_1}{\partial \xi_{1r}} + \frac{G_{1r}}{2} + \frac{n_2}{2} \frac{\partial \ln C_2}{\partial \xi_{2r}},
\]

\[
\frac{\partial^2 l}{\partial \theta_r \partial \theta_s} = \frac{n_1}{4} \frac{\partial^2 \ln C_1}{\partial \xi_{1r} \partial \xi_{1s}} + \frac{n_2}{4} \frac{\partial^2 \ln C_2}{\partial \xi_{2r} \partial \xi_{2s}} = (I_{\theta \theta})_{rs},
\]

\[
\frac{\partial^2 l}{\partial \theta_r \partial \beta_s} = \frac{n_1}{4} \frac{\partial^2 \ln C_1}{\partial \xi_{1r} \partial \xi_{1s}} + \frac{n_2}{4} \frac{\partial^2 \ln C_2}{\partial \xi_{2r} \partial \xi_{2s}} = (I_{\theta \beta})_{rs},
\]

\[
\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} = \frac{n_1}{4} \frac{\partial^2 \ln C_1}{\partial \xi_{1r} \partial \xi_{1s}} + \frac{n_2}{4} \frac{\partial^2 \ln C_2}{\partial \xi_{2r} \partial \xi_{2s}} = (I_{\beta \beta})_{rs}.
\]

From Lehmann (1959, p. 58), \( \frac{\partial \ln C_i}{\partial \xi_{ir}} = -E_{\xi_1}[h_r] \) and \( \frac{\partial^2 \ln C_i}{\partial \xi_{ir} \partial \xi_{is}} = -\text{cov}_{\xi_1}(h_r, h_s) \), for \( i = 1, 2 \). It follows that the elements of the efficient score are given by

\[
2U_r = G_{1r} - G_{2r} - [n_1E_{\xi_1}[h_r] - n_2E_{\xi_2}[h_r]].
\]

The expectations in \( U_r \) are unknown, but in the usable form of the score test are estimated by maximum likelihood under the null hypothesis. These equations are \( \frac{\partial l}{\partial \beta_r} = 0, \ r = 1, \ldots, q, \) and yield, if we write \( E_{\beta} \) for the expectation under the null hypothesis,

\[
G_{1r} + G_{2r} = (n_1 + n_2)E_{\beta}[h_r].
\]

Substituting in the efficient score and simplifying gives, under the null hypothesis,

\[
U_r = (n_2G_{1r} - n_1G_{2r})/(n_1+n_2).
\]

Write \( \Sigma = (\text{cov}_{\xi}(h_r, h_s)) \). The partitioned information matrix is given by

\[
\begin{pmatrix}
I_{\theta \theta} & I_{\theta \beta} \\
I_{\beta \theta} & I_{\beta \beta}
\end{pmatrix} = \frac{1}{4} \begin{pmatrix}
(n_1+n_2)\Sigma & (n_1-n_2)\Sigma \\
(n_1-n_2)\Sigma & (n_1+n_2)\Sigma
\end{pmatrix}.
\]
The inverse of the asymptotic covariance matrix of \( U_{\theta} \) is \( I_{\theta} - I_{\theta} \beta I_{\beta}^{-1} I_{\beta} \) = \( n_1 n_2 \Sigma / (n_1 + n_2) \). Denote \( \Sigma \) when evaluated under the null hypothesis by \( \Sigma_{\beta} \). The score statistic is then

\[
(n_1 + n_2)(U_{\tau})^T \Sigma_{\beta}^{-1}(U_{\tau}) / (n_1 n_2).
\]

To make this more accessible, we replace the \( \xi_i \) in the original model by \( \xi_i / \sqrt{n_1 + n_2} \), \( i = 1 \) and \( 2 \). The \( s \)th probability density function is thus assumed to be

\[
C_s(\xi_{s1}, \ldots, \xi_{sk}) \exp\left(\sum_{i=1}^{k} \xi_{si} h_i(x) / \sqrt{n_1 + n_2}\right) f(x), \quad s = 1, 2. \tag{3.1}
\]

Such distributions are said to be contiguos, and tend to the target distributions with probability density functions \( f(x) \) as the sample sizes tend to infinity. Each probability density function is now asymptotically equivalent to that used in Theorem 1 of Rayner and Best (1986). From the proof of that theorem, under the alternative hypothesis there, it follows that asymptotically \( E_{\beta} [G_{sr}] = n_s \theta_{sr} / \sqrt{n_1 + n_2} \), \( s = 1, 2 \), and \( \Sigma = I_k \), the \( k \times k \) identity matrix. Under this revised model the score statistic is

\[
S_k = V_1^2 + \ldots + V_k^2,
\]

in which

\[
V_k = \{n_2 \sum_j h_r(X_j) - n_1 \sum_j h_r(Y_j)\} / \sqrt{n_1 n_2 (n_1 + n_2)}.
\]

These \( V_k \) are asymptotically independent and asymptotically have the standard normal distribution. Moreover they may be derived as score statistics in their own right, and hence are weakly optimal for detecting specifiable alternatives.

If we define \( V_{1r} = \sum_j h_r(X_j) / n_1 \) and \( V_{2r} = \sum_j h_r(Y_j) / n_2 \), the usual components from the one sample goodness of fit problem (see Rayner and Best, 1989, Chapter 4), then

\[
V_r = \sqrt{n_2 V_{1r}} - \sqrt{n_1 V_{2r}} / \sqrt{n_1 + n_2}.
\]

So we can calculate the one sample components and combine them as indicated to assess whether the two populations can be considered to be the same. As a corollary, the one sample components could then be used for subsequent testing for the so-called target distribution.

4 Small Sample Size Study

In order to assess the distributional conclusions of the preceding section, a small Monte Carlo study was undertaken. With no nuisance parameters included the
target distribution was taken to be the uniform, and the \(h_r(x)\) were taken to be the Legendre polynomials. We calculated the proportion of exceedances for the asymptotic 99%, 95%, 90%, 80%, 10%, 5% and 1% points for each of the first four components. These components all asymptotically follow the standard normal distribution. If the common null distribution of the two samples is uniform, the simulated exceedances are within sampling error of the nominal exceedances for sample sizes as low as \(n_1 = n_2 = 5\).

We next considered the family of truncated exponential-type distributions with probability density functions

\[
f_X(x; \theta) = \lambda \exp(\lambda x / \sqrt{n}) / [\sqrt{n}[\exp(\lambda / \sqrt{n}) - 1]], \quad x \in (0, 1)
\]

\[
= 0 \text{ otherwise},
\]

(4.1)

in which \(\lambda > 0\).

With \(\lambda = 5\) the nominal and simulated exceedances were close, even for \(n_1 = n_2 = 10\). Table 1 gives results for \(\lambda = 10\). In all cases the number of simulations is 1,000.

The reason for the good agreement for small \(\lambda\) relative to the sample sizes, is that then (4.1) is well approximated by probability density functions of the form (3.1) with \(k\) small, or, equivalently, most of the \(\xi_{si}\) small. It should be remembered that the asymptotic distribution of the \(V_i\) depends on the Central Limit Theorem and the \(\xi_{si}\). For unimodal \(h_r(x)\) the Central Limit Theorem will converge rapidly; for higher order polynomials (larger \(r\)) convergence will be slower. Also the moments of the \(V_r\) will more closely approximate their asymptotic values if the \(n_i\) are large or if the \(\xi_{si}\) are small: that is, if the target distribution is well-chosen. Similar comments apply to the power calculations.

If \(n_1\) and \(n_2\) are unequal, we conjecture that the exceedances are better than for \(n_1 = n_2 = \) the minimum of the unequal sample sizes, and perhaps similar to \(n_1 = n_2 = \) the mean of the unequal sample sizes.

It is interesting to note that for non-contiguous nulls, the exceedances can be poor and do not improve with increasing sample size. See Table 2, in which the common null distribution has cumulative distribution function \(F_X(x) = x^{0.25}\).
Table 1 Proportions exceeding nominal percentage points for a contiguous truncated exponential distribution with $\lambda = 10$ and for components based on the first four Legendre polynomials. Based on 1,000 simulations for equal sample sizes $m = n = 5, 10, 20, 50$.

<table>
<thead>
<tr>
<th>Nominal exceedances</th>
<th>99%</th>
<th>95%</th>
<th>90%</th>
<th>80%</th>
<th>20%</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1 = n_2 = 5$</td>
<td>0.999</td>
<td>0.991</td>
<td>0.970</td>
<td>0.894</td>
<td>0.121</td>
<td>0.036</td>
<td>0.010</td>
<td>0.000</td>
</tr>
<tr>
<td>$n_1 = n_2 = 10$</td>
<td>0.999</td>
<td>0.979</td>
<td>0.943</td>
<td>0.851</td>
<td>0.167</td>
<td>0.060</td>
<td>0.018</td>
<td>0.003</td>
</tr>
<tr>
<td>$n_1 = n_2 = 20$</td>
<td>0.997</td>
<td>0.970</td>
<td>0.930</td>
<td>0.830</td>
<td>0.168</td>
<td>0.077</td>
<td>0.034</td>
<td>0.006</td>
</tr>
<tr>
<td>$n_1 = n_2 = 50$</td>
<td>0.993</td>
<td>0.959</td>
<td>0.908</td>
<td>0.805</td>
<td>0.147</td>
<td>0.075</td>
<td>0.037</td>
<td>0.010</td>
</tr>
<tr>
<td>$V_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1 = n_2 = 5$</td>
<td>0.991</td>
<td>0.942</td>
<td>0.885</td>
<td>0.779</td>
<td>0.229</td>
<td>0.118</td>
<td>0.061</td>
<td>0.014</td>
</tr>
<tr>
<td>$n_1 = n_2 = 10$</td>
<td>0.993</td>
<td>0.952</td>
<td>0.902</td>
<td>0.811</td>
<td>0.216</td>
<td>0.112</td>
<td>0.059</td>
<td>0.015</td>
</tr>
<tr>
<td>$n_1 = n_2 = 20$</td>
<td>0.990</td>
<td>0.953</td>
<td>0.898</td>
<td>0.789</td>
<td>0.214</td>
<td>0.102</td>
<td>0.051</td>
<td>0.010</td>
</tr>
<tr>
<td>$n_1 = n_2 = 50$</td>
<td>0.986</td>
<td>0.943</td>
<td>0.881</td>
<td>0.782</td>
<td>0.207</td>
<td>0.103</td>
<td>0.051</td>
<td>0.019</td>
</tr>
<tr>
<td>$V_3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1 = n_2 = 5$</td>
<td>0.976</td>
<td>0.920</td>
<td>0.856</td>
<td>0.783</td>
<td>0.227</td>
<td>0.128</td>
<td>0.067</td>
<td>0.019</td>
</tr>
<tr>
<td>$n_1 = n_2 = 10$</td>
<td>0.982</td>
<td>0.936</td>
<td>0.879</td>
<td>0.774</td>
<td>0.219</td>
<td>0.111</td>
<td>0.061</td>
<td>0.016</td>
</tr>
<tr>
<td>$n_1 = n_2 = 20$</td>
<td>0.990</td>
<td>0.946</td>
<td>0.884</td>
<td>0.783</td>
<td>0.192</td>
<td>0.109</td>
<td>0.063</td>
<td>0.014</td>
</tr>
<tr>
<td>$n_1 = n_2 = 50$</td>
<td>0.983</td>
<td>0.947</td>
<td>0.891</td>
<td>0.800</td>
<td>0.198</td>
<td>0.099</td>
<td>0.052</td>
<td>0.009</td>
</tr>
<tr>
<td>$V_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1 = n_2 = 5$</td>
<td>0.979</td>
<td>0.908</td>
<td>0.863</td>
<td>0.778</td>
<td>0.246</td>
<td>0.133</td>
<td>0.076</td>
<td>0.018</td>
</tr>
<tr>
<td>$n_1 = n_2 = 10$</td>
<td>0.982</td>
<td>0.944</td>
<td>0.890</td>
<td>0.796</td>
<td>0.215</td>
<td>0.106</td>
<td>0.059</td>
<td>0.014</td>
</tr>
<tr>
<td>$n_1 = n_2 = 20$</td>
<td>0.978</td>
<td>0.931</td>
<td>0.879</td>
<td>0.773</td>
<td>0.193</td>
<td>0.104</td>
<td>0.057</td>
<td>0.016</td>
</tr>
<tr>
<td>$n_1 = n_2 = 50$</td>
<td>0.985</td>
<td>0.947</td>
<td>0.902</td>
<td>0.790</td>
<td>0.205</td>
<td>0.113</td>
<td>0.057</td>
<td>0.010</td>
</tr>
</tbody>
</table>
Table 2  Proportions exceeding nominal percentage points for a (non-contiguous) fourth root distribution and for components based on the first four Legendre polynomials. Based on 1,000 simulations for equal sample sizes $m = n = 10, 50, 100, 200$.

<table>
<thead>
<tr>
<th>Nominal exceedances</th>
<th>99%</th>
<th>95%</th>
<th>90%</th>
<th>80%</th>
<th>20%</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1=n_2=10$</td>
<td>0.995</td>
<td>0.959</td>
<td>0.918</td>
<td>0.822</td>
<td>0.197</td>
<td>0.087</td>
<td>0.039</td>
<td>0.004</td>
</tr>
<tr>
<td>$n_1=n_2=50$</td>
<td>0.995</td>
<td>0.963</td>
<td>0.923</td>
<td>0.817'</td>
<td>0.164</td>
<td>0.073</td>
<td>0.033</td>
<td>0.007</td>
</tr>
<tr>
<td>$n_1=n_2=100$</td>
<td>0.992</td>
<td>0.958</td>
<td>0.908</td>
<td>0.818</td>
<td>0.172</td>
<td>0.086</td>
<td>0.044</td>
<td>0.009</td>
</tr>
<tr>
<td>$n_1=n_2=200$</td>
<td>0.992</td>
<td>0.963</td>
<td>0.914</td>
<td>0.815</td>
<td>0.177</td>
<td>0.079</td>
<td>0.036</td>
<td>0.006</td>
</tr>
<tr>
<td>$V_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1=n_2=10$</td>
<td>0.974</td>
<td>0.909</td>
<td>0.853</td>
<td>0.746</td>
<td>0.218</td>
<td>0.120</td>
<td>0.074</td>
<td>0.026</td>
</tr>
<tr>
<td>$n_1=n_2=50$</td>
<td>0.968</td>
<td>0.907</td>
<td>0.849</td>
<td>0.749</td>
<td>0.258</td>
<td>0.164</td>
<td>0.112</td>
<td>0.042</td>
</tr>
<tr>
<td>$n_1=n_2=100$</td>
<td>0.966</td>
<td>0.912</td>
<td>0.849</td>
<td>0.749</td>
<td>0.252</td>
<td>0.155</td>
<td>0.086</td>
<td>0.036</td>
</tr>
<tr>
<td>$n_1=n_2=200$</td>
<td>0.973</td>
<td>0.898</td>
<td>0.843</td>
<td>0.739</td>
<td>0.225</td>
<td>0.136</td>
<td>0.079</td>
<td>0.024</td>
</tr>
<tr>
<td>$V_3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1=n_2=10$</td>
<td>0.948</td>
<td>0.880</td>
<td>0.829</td>
<td>0.744</td>
<td>0.285</td>
<td>0.180</td>
<td>0.126</td>
<td>0.048</td>
</tr>
<tr>
<td>$n_1=n_2=50$</td>
<td>0.945</td>
<td>0.872</td>
<td>0.809</td>
<td>0.712</td>
<td>0.261</td>
<td>0.166</td>
<td>0.108</td>
<td>0.037</td>
</tr>
<tr>
<td>$n_1=n_2=100$</td>
<td>0.949</td>
<td>0.875</td>
<td>0.804</td>
<td>0.714</td>
<td>0.281</td>
<td>0.198</td>
<td>0.130</td>
<td>0.047</td>
</tr>
<tr>
<td>$n_1=n_2=200$</td>
<td>0.946</td>
<td>0.879</td>
<td>0.810</td>
<td>0.711</td>
<td>0.271</td>
<td>0.177</td>
<td>0.103</td>
<td>0.037</td>
</tr>
<tr>
<td>$V_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1=n_2=10$</td>
<td>0.933</td>
<td>0.858</td>
<td>0.803</td>
<td>0.709</td>
<td>0.330</td>
<td>0.226</td>
<td>0.168</td>
<td>0.089</td>
</tr>
<tr>
<td>$n_1=n_2=50$</td>
<td>0.928</td>
<td>0.842</td>
<td>0.779</td>
<td>0.693</td>
<td>0.297</td>
<td>0.206</td>
<td>0.145</td>
<td>0.063</td>
</tr>
<tr>
<td>$n_1=n_2=100$</td>
<td>0.925</td>
<td>0.842</td>
<td>0.784</td>
<td>0.705</td>
<td>0.296</td>
<td>0.217</td>
<td>0.152</td>
<td>0.066</td>
</tr>
<tr>
<td>$n_1=n_2=200$</td>
<td>0.923</td>
<td>0.853</td>
<td>0.805</td>
<td>0.724</td>
<td>0.297</td>
<td>0.203</td>
<td>0.146</td>
<td>0.075</td>
</tr>
</tbody>
</table>
5 Small Sample Power Study

To assess the power of the two-sample smooth tests, we assumed the two samples had probability density functions (4.1) with parameters $\lambda_1$ and $\lambda_2$ respectively. The parameter space is two dimensional with the null hypothesis being the line $\lambda_1 = \lambda_2$.

To achieve test sizes of exactly 5%, 10,000 samples of sizes $m$ and $n$ were generated, the test statistics ordered, and the 9,500th selected. With these critical points, the proportions of exceedances for $\lambda_i = 2$ $(i = 1, 2)$ were calculated using a further 1,000 samples. The results, for $V_1^2$, $V_2^2$, $V_3^2$, $V_4^2$, $S_2$, $S_3$, $S_4$, and $A_{2n}^2$, the Anderson-Darling statistic of Scholz and Stephens (1987), are displayed in Table 3.

It is worth noting that the simulated 5% Anderson-Darling critical point for $n_1 = n_2 = 20$ was 2.4239. The asymptotic standardized percentiles technique of Scholz and Stephens (1987) yielded 2.4256. Using the asymptotic variance they quote from Pettitt (1976) yields a value of 2.4334, which was more convenient to calculate, and sufficient to check the simulated value. Unlike the smooth test statistics, $A_{2n}^2$ is distribution free. So while the smooth statistics were randomised to give exact size only at $\lambda_1 = \lambda_2 = 10$, $A_{2n}^2$ can be expected to have accurate size for all null points.

In view of the comments in the preceding section, we expected that the smaller $\lambda$, the better the power of the lower order smooth tests. The performance of the Anderson-Darling test relative to the smooth tests depends on the nature of its decomposition. For example, Durbin and Knott (1971) expressed the one sample Anderson-Darling statistic as a linear combination of variables having independent $\chi_1^2$ distributions, with weights $[j(j + 1)]^{-1}$. So that test is effectively of low order (see Rayner Best and Dodds, 1985), weighting the first component heavily, and the second component significantly, and subsequent components lightly. This assessment is consistent with the calculations in Table 3. For high values of $\lambda_1$ and $\lambda_2$ the sizes for the $V_1$ test are unfortunately low, because we randomised only for $\lambda_1 = \lambda_2 = 10$. The low size confers an advantage on the $A_{2n}^2$ test. Without that disadvantage, the $V_1$ test usually outperforms the $A_{2n}^2$ test.

There appears to be a non-uniform ordering of the tests: $V_1^2$, $A_{2n}^2$, $S_2$, $V_2^2$, $S_3$, $S_4$, $V_3^2$, $V_4^2$ in decreasing effectiveness. See Figure 2.

We repeated the simulations above using instead of the truncated exponential-type distribution, a trigonometric-type distribution with probability density function
Figure 2. Comparative powers for $A_2$, $V_1$, $V_2$, $V_3$, $V_4$ for target distribution (4.1). Vertical axis is power $\times$ 1000, horizontal axes are $\lambda_1/2, \lambda_2/2$. 
\[ f_X(x; \omega) = \frac{2\pi \omega}{1 + \sin(2\pi \omega)} / \left[ 1 + 2\pi \omega - \cos(2\pi \omega) \right], x \in (0, 1) \]
\[ = 0 \text{ otherwise,} \]
in which \( \omega > 0 \).

The results are in Table 4. Our aim was to assess alternatives that the polynomial orthonormal functions would not represent well. On the other hand, the lower frequency alternatives with small \( \omega \) would perhaps be well represented by the Legendre polynomials.

For the smooth tests we found critical values based on 10,000 simulations for \( \omega_1 = \omega_2 = 3 \). The powers based on 1,000 simulations at points with \( \omega_1 = \omega_2 \) varied more than we would have preferred, indicating some sampling variation and that the critical values are not as robust as the \( A_2n^2 \) critical values. Of course the alternatives are quite different from the target distribution. We again found that the \( A_2n^2 \) test performed similarly to the the \( V_1 \) test, and sometimes better when one of the subsequent components had substantial power. When two or more components were substantial, the omnibus tests performed very well. When both \( \omega_1 \) and \( \omega_2 \) were at least 3, none of the tests were very powerful.

Our conclusions with regard to the \( A_2n^2 \) test are that it appears to be predominantly what we would call a first order test, being sensitive especially to location shifts. Although its distribution theory is most adequate for applications, it is reasonable to ask if the Anderson-Darling test provides information beyond that supplied by the Wilcoxon test.

The smooth tests provide a wealth of information. If there is some preliminary knowledge about the alternatives it is best to protect against, this information can be built into the choice of target distribution and the orthonormal polynomials. If the target distribution is well chosen, the asymptotic null distributions of the smooth tests is probably adequate for moderate sample sizes.
Table 3: Powers of the tests based on $V_1^2$, $V_2^2$, $V_3^2$, $V_4^2$, $S_2$, $S_3$, $S_4$, and $A_{2n}^2$ for exponential-type alternatives and size exactly 5% if $\lambda_1 = \lambda_2 = 10$. Based on 1,000 simulations for equal sample sizes $m = n = 20$. In each cell the first line gives the power for the test based on $A_{2n}^2$, the second line gives the powers based on the components $V_1^2$, $V_2^2$, $V_3^2$, $V_4^2$, and the third line gives the powers based on the omnibus statistics $S_2$, $S_3$, $S_4$.

<table>
<thead>
<tr>
<th>$\lambda_1/\lambda_2$</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>18</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>50</td>
<td>73</td>
<td>118</td>
<td>215</td>
<td>314</td>
<td>458</td>
<td>589</td>
<td>685</td>
<td>775</td>
<td>863</td>
</tr>
<tr>
<td></td>
<td>78, 52, 47, 47</td>
<td>99, 59, 37, 41</td>
<td>170, 61, 36, 35</td>
<td>287, 61, 55, 44</td>
<td>378, 83, 48, 53</td>
<td>515, 98, 46, 52</td>
<td>657, 145, 71, 54</td>
<td>748, 208, 72, 52</td>
<td>802, 257, 79, 77</td>
<td>853, 343, 103,</td>
</tr>
<tr>
<td></td>
<td>64, 55, 42</td>
<td>81, 69, 54</td>
<td>125, 100, 77</td>
<td>204, 157, 132</td>
<td>302, 252, 220</td>
<td>441, 338, 284</td>
<td>563, 487, 425</td>
<td>706, 609, 547</td>
<td>776, 721, 664</td>
<td>856, 790, 732</td>
</tr>
<tr>
<td>4</td>
<td>69</td>
<td>50</td>
<td>61</td>
<td>135</td>
<td>195</td>
<td>294</td>
<td>416</td>
<td>548</td>
<td>664</td>
<td>746</td>
</tr>
<tr>
<td></td>
<td>102, 55, 46, 28</td>
<td>84, 37, 36, 43</td>
<td>106, 46, 42, 40</td>
<td>168, 54, 48, 31</td>
<td>341, 109, 51, 45</td>
<td>474, 142, 48, 63</td>
<td>630, 182, 63, 67</td>
<td>680, 237, 81, 58</td>
<td>780, 350, 105,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>86, 69, 51</td>
<td>61, 62, 49</td>
<td>78, 75, 60</td>
<td>121, 98, 82</td>
<td>206, 167, 151</td>
<td>289, 238, 196</td>
<td>416, 352, 303</td>
<td>579, 491, 454</td>
<td>641, 561, 518</td>
<td>785, 722, 661</td>
</tr>
<tr>
<td>6</td>
<td>102</td>
<td>70</td>
<td>51</td>
<td>66</td>
<td>120</td>
<td>169</td>
<td>276</td>
<td>374</td>
<td>497</td>
<td>597</td>
</tr>
<tr>
<td></td>
<td>115, 93, 73</td>
<td>77, 69, 59</td>
<td>53, 48, 39</td>
<td>83, 59, 47</td>
<td>105, 101, 87</td>
<td>178, 154, 130</td>
<td>270, 220, 190</td>
<td>381, 309, 270</td>
<td>493, 425, 370</td>
<td>616, 546, 497</td>
</tr>
<tr>
<td>8</td>
<td>228</td>
<td>120</td>
<td>63</td>
<td>52</td>
<td>62</td>
<td>100</td>
<td>160</td>
<td>240</td>
<td>356</td>
<td>460</td>
</tr>
<tr>
<td></td>
<td>292, 58, 46, 30</td>
<td>154, 39, 38, 46</td>
<td>93, 47, 52, 47</td>
<td>76, 54, 52, 45</td>
<td>76, 55, 52, 49</td>
<td>116, 74, 51, 57</td>
<td>168, 81, 54, 46</td>
<td>259, 157, 72, 69</td>
<td>347, 197, 81, 57</td>
<td>393, 238, 81, 5</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>114</td>
<td>69</td>
<td>46</td>
<td>65</td>
<td>93</td>
<td>139</td>
<td>207</td>
<td>341</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td></td>
<td>323</td>
<td>195</td>
<td>114</td>
<td>69</td>
<td>46</td>
<td>65</td>
<td>93</td>
<td>139</td>
<td>207</td>
<td></td>
</tr>
<tr>
<td></td>
<td>374, 72, 39, 47</td>
<td>251, 84, 62, 58</td>
<td>136, 59, 53, 44</td>
<td>78, 55, 50, 48</td>
<td>52, 58, 47, 60</td>
<td>63, 57, 43, 51</td>
<td>100, 68, 56, 59</td>
<td>97, 78, 75</td>
<td>149, 101, 75, 69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>291, 209, 183</td>
<td>203, 171, 143</td>
<td>98, 98, 92</td>
<td>71, 75, 65</td>
<td>63, 67, 66</td>
<td>76, 65, 60</td>
<td>97, 78, 75</td>
<td>141, 140, 118</td>
<td>217, 195, 165</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>471</td>
<td>166</td>
<td>110</td>
<td>72</td>
<td>41</td>
<td>64</td>
<td>97</td>
<td>137</td>
<td></td>
</tr>
<tr>
<td></td>
<td>528, 107, 56, 38</td>
<td>343, 87, 52, 43</td>
<td>183, 90, 60, 44</td>
<td>128, 70, 53, 47</td>
<td>54, 59, 54, 59</td>
<td>43, 53, 47, 46</td>
<td>57, 61, 50, 61</td>
<td>88, 88, 60, 59</td>
<td>102, 118, 74, 61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>452, 367, 298</td>
<td>271, 216, 175</td>
<td>158, 125, 105</td>
<td>104, 87, 77</td>
<td>57, 65, 68</td>
<td>43, 48, 41</td>
<td>54, 65, 59</td>
<td>104, 88, 100</td>
<td>134, 126, 118</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>578</td>
<td>264</td>
<td>160</td>
<td>102</td>
<td>63</td>
<td>56</td>
<td>74</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td></td>
<td>665, 135, 48, 57</td>
<td>464, 136, 48, 47</td>
<td>320, 126, 60, 61</td>
<td>164, 103, 59, 47</td>
<td>95, 70, 62, 57</td>
<td>46, 55, 79, 66</td>
<td>41, 67, 65, 64</td>
<td>39, 67, 71, 65</td>
<td>66, 90, 64, 72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>562, 468, 414</td>
<td>416, 331, 289</td>
<td>283, 233, 213</td>
<td>165, 141, 110</td>
<td>88, 85, 84</td>
<td>57, 72, 74</td>
<td>58, 64, 71</td>
<td>48, 66, 61</td>
<td>83, 91, 86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>684</td>
<td>386</td>
<td>241</td>
<td>166</td>
<td>93</td>
<td>72</td>
<td>40</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>747, 189, 61, 52</td>
<td>620, 191, 54, 66</td>
<td>432, 161, 68, 63</td>
<td>267, 136, 47, 55</td>
<td>129, 102, 52, 62</td>
<td>73, 87, 62, 47</td>
<td>37, 61, 54, 70</td>
<td>31, 76, 64, 62</td>
<td>32, 71, 62, 81</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>787</td>
<td>500</td>
<td>342</td>
<td>230</td>
<td>135</td>
<td>97</td>
<td>57</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>828, 284, 66, 64</td>
<td>693, 242, 77, 63</td>
<td>505, 222, 67, 57</td>
<td>331, 169, 74, 66</td>
<td>191, 129, 58, 64</td>
<td>122, 116, 67, 67</td>
<td>37, 91, 71, 67</td>
<td>34, 54, 69, 68</td>
<td>18, 63, 76, 80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>797, 727, 646</td>
<td>646, 578, 527</td>
<td>492, 421, 366</td>
<td>323, 298, 251</td>
<td>204, 172, 162</td>
<td>151, 142, 128</td>
<td>81, 79, 74</td>
<td>47, 59, 70</td>
<td>48, 61, 71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>857</td>
<td>625</td>
<td>420</td>
<td>320</td>
<td>190</td>
<td>131</td>
<td>81</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>879, 320, 100, 70</td>
<td>778, 322, 90, 57</td>
<td>606, 286, 105, 55</td>
<td>434, 230, 101, 66</td>
<td>301, 200, 93, 74</td>
<td>169, 134, 67, 72</td>
<td>74, 104, 80, 48</td>
<td>47, 104, 71, 63</td>
<td>12, 64, 73, 70</td>
<td></td>
</tr>
</tbody>
</table>

The table contains numerical data arranged in a grid format.
Table 4 *Powers of the tests based on $V_1^2$, $V_2^2$, $V_3^2$, $V_4^2$, $S_2$, $S_3$, $S_4$, and $A_{2n^2}$ for trigonometric alternatives and size exactly 5% for $\omega = \omega_2 = 3$. Based on 1,000 simulations for equal sample sizes $m = n = 20$. In each cell the first line gives the power for the test based on $A_{2n^2}$, the second line gives the powers based on the components $V_1^2$, $V_2^2$, $V_3^2$, $V_4^2$, and the third line gives the powers based on the omnibus statistics $S_2$, $S_3$, $S_4$.*

<table>
<thead>
<tr>
<th>$\omega_1/\omega_2$</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
<th>3.5</th>
<th>4.0</th>
<th>4.5</th>
<th>5.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>50</td>
<td>310</td>
<td>303</td>
<td>451</td>
<td>312</td>
<td>452</td>
<td>403</td>
<td>460</td>
<td>358</td>
</tr>
<tr>
<td>1.5</td>
<td>309</td>
<td>55</td>
<td>249</td>
<td>85</td>
<td>138</td>
<td>64</td>
<td>131</td>
<td>62</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>464, 180, 316, 272</td>
<td>85, 26, 84, 33</td>
<td>167, 160, 137, 256</td>
<td>98, 55, 41, 396</td>
<td>126, 170, 124, 271</td>
<td>78, 69, 59, 425</td>
<td>105, 175, 101, 233</td>
<td>87, 95, 64, 456</td>
<td>104, 182, 97, 7</td>
</tr>
<tr>
<td>2.0</td>
<td>262</td>
<td>274</td>
<td>49</td>
<td>121</td>
<td>81</td>
<td>136</td>
<td>94</td>
<td>151</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>120, 50, 650, 47</td>
<td>163, 165, 148, 263</td>
<td>44, 53, 50, 47</td>
<td>163, 127, 142, 81</td>
<td>61, 42, 48, 49</td>
<td>156, 98, 157, 72</td>
<td>73, 50, 59, 55</td>
<td>160, 86, 154, 80</td>
<td>92, 54, 59, 48</td>
</tr>
<tr>
<td></td>
<td>103, 514, 475</td>
<td>218, 249, 368</td>
<td>53, 48, 52</td>
<td>188, 225, 215</td>
<td>60, 50, 48</td>
<td>165, 217, 218</td>
<td>58, 60, 62</td>
<td>152, 199, 191</td>
<td>91, 85, 79</td>
</tr>
</tbody>
</table>
6 Sample Goodness of Fit

Suppose we have $S$ independent random samples, the $s$th of which is $X_{s1}, ..., X_{sn}$. We wish to test if these samples are consistent with coming from the same population. The $s$th sample is assumed to come from a distribution with probability density function

$$C_s(\xi_{s1}, ..., \xi_{sk}) \exp \left\{ \sum_{i=1}^{k} \xi_{si} h_i(x) / \sqrt{(n_1 + ... + n_j)} \right\} f(x), s = 1, ..., S.$$  

In the two-sample derivation we transformed from $(\xi_{s1}, ..., \xi_{sk})$ to $(\theta, \beta)$. There is considerable choice in how we transform the parameters, in both the two and the $S$-sample cases. Here we use a transformation based on the Helmert matrices discussed, for example, in Lancaster (1965). We reparametrize by putting

$$(\theta_1^T, ..., \theta_{S-1}^T, \beta)^T = (H_S \otimes I_S) (\xi_1^T, ..., \xi_S^T)^T,$$

in which $\otimes$ is the Kronecker product, $I_n$ is the $n \times n$ identity matrix, and $H_n$ is the $n \times n$ Helmert matrix with last row $(1, ..., 1)/\sqrt{n}$ and $r+1$th row $(1, ..., 1)/\sqrt{r(r+1)}$, $r = 1, ..., n-1$. Writing $G_s = (\sum_j h_j(X_{sj}; \gamma_s), s = 1, ..., S$, this leads to a score vector proportional to

$$(G_1 - G_2 - E\beta [G_1 - G_2])^T, (G_1 + G_2 - 2G_3 - E\beta [G_1 + G_2 - 2G_3])^T, ..., (G_1 + ... + G_{S-1} - (S-1)G_S - E\beta [G_1 + ... + G_{S-1} - (S-1)G_S])^T)^T.$$  

At this point we now propose a hierarchical procedure. The difficulty is that the maximum likelihood equations are equivalent to

$$G_1 + ... + G_S = E\beta [G_1 + ... + G_S],$$

and substituting this into the score vector does not yield appealing components. On the other hand, under the null hypothesis that the first two samples are consistent with having come from the same population, we may use $G_1 + G_2 = E\beta [G_1 + G_2]$ for the maximum likelihood equations, and substitute this into $(G_1 - G_2 - E\beta [G_1 - G_2])$. We are lead to the vector of components

$$(n_2G_1 - n_1G_2)/\sqrt{n_1n_2(n_1+n_2)}.$$  

If the null hypothesis is accepted, we may proceed to the next step of the hierarchical testing. Under the null hypothesis that the first three samples are consistent with coming from the same population, and we use
$G_1 + G_2 + G_3 = E[\beta_1 G_1 + \beta_2 G_2 + \beta_3 G_3]$ for the maximum likelihood equations in $G_1 + G_2 - 2G_3 = E[\beta_1 G_1 + \beta_2 G_2 - 2\beta_3 G_3]$, leading to the vector of components 

$$[n_3(G_1 + G_2) - (n_1 + n_2)G_3]/\sqrt{(n_1 + n_2)n_3(n_1 + n_2 + n_3)}.$$ 

At the $s$th step ($s = 2, \ldots, S - 1$), we assess the consistency of the $s + 1$th sample with the preceding $s$ samples by using 

$$[n_{s+1}(G_1 + \ldots + G_s) - (n_1 + \ldots + n_s)G_{s+1}]/\sqrt{(n_1 + \ldots + n_s)n_{s+1}(n_1 + \ldots + n_{s+1})}.$$ 

The covariance matrix of this vector is asymptotically the $k \times k$ unit matrix. Write $V_s = (\sum_j h_j(X_j)/\sqrt{n_g})$. In terms of the sample components, at the $s$th step we assess each element of 

$$n_{s+1}(V_1\sqrt{n_1} + \ldots + V_s\sqrt{n_s}) - (n_1 + \ldots + n_s)V_{s+1}\sqrt{n_{s+1}}/\sqrt{(n_1 + \ldots + n_s)n_{s+1}(n_1 + \ldots + n_{s+1})}$$ 

relative to the standard normal distribution. The sum of the squares of these $k$ elements may be compared with the $\chi_k^2$ distribution to give an omnibus assessment of the null hypothesis.

Of course other contrasts may be chosen depending on the situation, and non-hierarchic testing may be preferred by some users.

7 Discussion and Example

As always with omnibus tests and their directional component tests, we have the option of informal data analytic assessment, and formal tests of significance. Our preference is a combination of both: using the omnibus test as the basis for a formal test of significance, and using the components informally to assess the nature of any differences should there be some.

To test for a common distribution, we assess the first two samples by calculating the components $(V_1\sqrt{n_1} - V_2\sqrt{n_2})/\sqrt{n_1 + n_2}$, and their sum of squares for an omnibus comparison. At the $s$th step, $s = 2, \ldots, S - 1$, we use 

$$n_{s+1}(V_1\sqrt{n_1} + \ldots + V_{s+1}\sqrt{n_{s+1}}) - (n_1 + \ldots + n_s)V_{s+1}\sqrt{n_{s+1}}/\sqrt{(n_1 + \ldots + n_s)n_{s+1}(n_1 + \ldots + n_{s+1})}$$ 

to assess the consistency of the order $r$ term of the $s + 1$th sample with the order $r$ terms of the preceding consistent samples. The sum of the squares of these statistics up to order $k$ gives an omnibus assessment based on a statistic with the $\chi_k^2$ distribution. If at any step the samples are considered to be inconsistent,
testing stops. Then the $V_{sr}$ may be used to assess the nature of the inconsistency. With uniform target and $\{h_r(x)\}$ the Legendre polynomials, the $V_{sr}$ reflect the $r$ th moment of the $s$ th population. This assists in giving an informative explanation of the differences.

If the null hypothesis of a common distribution is accepted, we may use the components to see if the target distribution is that common distribution. The statistic $V_{1r}^2 + \ldots + V_{sr}^2$ assesses order $r$ deviations from the target distribution, and a reasonable omnibus assessment would be given by using the statistic

$$\sum_{s=1}^{S} \sum_{r=1}^{R} V_{sr}^2.$$ 

Even if we decide the data are not consistent with the samples being from a common distribution, we can still assess if the $s$ th sample is consistent with the target distribution by using $\sum_{r=1}^{R} V_{sr}^2$.

**Example**

Lehmann (1975) gave data of four sets of eight measurements each of the smoothness of a certain type of paper, obtained from four laboratories. The Kruskal-Wallis test gave evidence of significant differences with a $P$-value of 0.005. The same data were analysed by Scholz and Stephens (1987) using various versions of their Anderson-Darling test. They reported $P$-values of around 0.002. The data are reproduced in Table 5(a).

<table>
<thead>
<tr>
<th>Laboratory</th>
<th>Smoothness</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>38.7</td>
</tr>
<tr>
<td>B</td>
<td>39.2</td>
</tr>
<tr>
<td>C</td>
<td>34.0</td>
</tr>
<tr>
<td>D</td>
<td>34.0</td>
</tr>
</tbody>
</table>
Table 5(b) Components $V_{sr}$ for the data in Table 4(b).

<table>
<thead>
<tr>
<th>Population (s)</th>
<th>Order (r)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.2327</td>
<td>-1.9439</td>
<td>0.7992</td>
<td>1.8767</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>-1.0859</td>
<td>-2.5016</td>
<td>2.0570</td>
<td>1.2663</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>-1.5105</td>
<td>-1.6286</td>
<td>1.5978</td>
<td>0.4755</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>-2.5311</td>
<td>-0.2467</td>
<td>2.1608</td>
<td>-1.8855</td>
</tr>
</tbody>
</table>

With no preconceptions about the data it seems reasonable to assume a target distribution that is uniform between 30 and 60. Probability plots suggest uniformity is as dubious an assumption as normality. In addition, the value 58.0 looks suspiciously large. We used the Legendre polynomials so that the components could be given a moment interpretation. The $V_{sr}$ are given in Table 5(b). Testing hierarchically, we find samples one, two and three are consistent, but sample four is inconsistent with the other three. The test statistics for the final comparison are $V_r = (V_{1r} + V_{2r} + V_{3r} - 3V_{4r})/\sqrt{12}$, $r = 1, \ldots, 4$. These take the values 1.51, -1.54, -0.59 and 2.68. The sum of the squares of these values is 12.16, which yields a P-value of 0.016 on using the $X_{42}^2$ distribution.

Had we chosen initially to compare the first and fourth samples, the component values are 1.95, -1.20, -0.96 and 2.66, with a sum of the squares of 13.26, and a P-value of 0.01. An LSD-type comparison shows clearly in what respects the samples differ. For a pairwise comparison based on samples of size $n_1$ and $n_2$, we define

$$LSD = z(\alpha / 2C\%)/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

in which $z(\alpha / 2C\%)$ is the point giving equal 50\% tails with the standard normal distribution, the asymptotic distribution of the $V_{sr}$. Here we take $\alpha = 5\%$, giving LSD = 1.32. Most notable are the second and fourth order comparisons: the fourth sample is more variable and less peaked than the other three. See Figure 1.

If we take the sum of the squares of the components in Table 5(b), we see that the target uniform distribution is not a good description of the data. This is reflected in that the components of orders two and three all have the same sign, and that
this is almost true for the components of orders one and four. It could be that further $ξ_{s1}$ are needed to model the probability density functions corresponding to the different samples. We could assess this by calculating further $V_{sr}$. It may well be that there are differences in moments beyond the four considered here, although whether we really want to know this is a moot point. Also, with all samples being of size eight, the asymptotic null distributions are questionable. This could be overcome by calculating Monte Carlo P-values for the $V_i$ and $S_i$. The small sample sizes also make the LSD comparisons in Figure 1 dubious. However, it is not in doubt that the populations may be so ordered with regard to their relative means, variances etc.

8 Nuisance Parameters in Target Distribution

In the analysis of variance situation we would like to be able to assess if the samples come from normal populations with possibly different means and variances. We thus wish to permit the target probability density functions $f(x; θ_s)$, $s = 1, \ldots, S$, and we model the $s$th probability density function by

$$C_s(ξ_{s1}, \ldots, ξ_{sk}) \exp\left(\sum_{i=1}^{k} ξ_{si} h_i(x; θ_s)/\sqrt{n_1 + \ldots + n_s}\right) f(x; θ_s), s = 1, \ldots, S.$$  

The $h_i(x; θ)$ are here taken to be orthonormal on $f(x; θ)$. With this modification the nuisance parameter of section 3 changes from $β$ to $(β^T | θ_1^T | θ_2^T)^T$. The information matrix is much more complicated, having elements such as $\text{cov}(h_{1r}, \partial \ln f_1/\partial θ_{i1})$, $i = 1$ and 2, and $\text{cov}(\partial \ln f_1/\partial θ_{1s}, \partial \ln f_2/\partial θ_{2s})$. However the $Iθ_β$ submatrix can be partitioned into a submatrix $(n_1-n_2)\Sigma$ and another submatrix involving the $\text{cov}(h_{1r}, \partial \ln f_1/\partial θ_{i1})$. Following the one sample development in Rayner and Best (1989, Chapter 6), this submatrix is zero when the target distribution is normal, exponential, Poisson, geometric, multivariate normal, and no doubt other distributions also. This has the effect that, for such appropriate target distributions, the inverse of the asymptotic covariance matrix of $U_θ$ is $n_1n_2\Sigma/(n_1+n_2)$. That is, it is as before, but now involving the additional nuisance parameters evaluated at the maximum likelihood estimators under the target model. So the score statistic is

$$S_k = \tilde{V}_1^2 + \ldots + \tilde{V}_k^2,$$

in which $\tilde{V}_r = (\sqrt{n_2} \tilde{V}_{1r} - \sqrt{n_1} \tilde{V}_{2r})/\sqrt{(n_1+n_2)}$.

The $\tilde{V}_{1r}$ and $\tilde{V}_{2r}$ being the usual one-sample components. The summation should be understood to be over $k$ summands not identically zero. For example if the target distribution is the Poisson ($λ$) with unspecified mean, $\tilde{V}_1 = \tilde{V}_{11} = \tilde{V}_{21} = 0$, and the first nonzero component involves $\sum_j h_2(X_1j; X'1)$ and $\sum_j h_2(X_2j; X'2)$. The
extension to the full S-sample problem for such target distributions is routine: the
statistics are all as before, with the additional nuisance parameters being evaluated
at their maximum likelihood estimators under the target distribution.

References
Biometrical Journal 4, 395-402.

John Wiley.

York: Marcell Dekker.


Eplett, W.J.R. (1982). The distributions of Smirnov type two-sample rank tests for


72, 4-12.


(1), 161-168.

Rayner, J.C.W. and Best, D.J. (1986). Neyman-type smooth tests for location-scale
families. Biometrika, 73, 437-446.

