AN ABSTRACT OF THE THESIS OF

Kenneth James Kopecky for the degree of Doctor of Philosophy

in Statistics presented on August 12, 1977

Title: THE EFFICIENCY OF SMOOTH GOODNESS-OF-FIT TESTS

Abstract approved: Redacted for privacy

Donald A. Pierce

Neyman (1939) obtained a quadratic score statistic for testing simple goodness-of-fit hypotheses in a family of "smooth" alternatives. Javitz (1975) and Thomas and Pierce (1977) described the appropriate quadratic score statistic for composite goodness-of-fit hypotheses against "smooth" alternatives. These statistics, ψ_k^2 and $\psi_k^2(\hat{\lambda}_0)$, are described and matrices needed for their application to some examples are calculated. The possibility of applying the procedures to testing multivariate normality is considered and difficulties are seen to arise.

Noncentrality parameters for power calculations are derived, and a measure of asymptotic relative efficiency of the smooth tests with other tests based on quadratic score statistics is derived. This permits comparison of the smooth tests with χ^2 goodness-of-fit tests of the Pearson type as well as locally best tests for parametric alternatives. For the examples considered the smooth tests for composite hypotheses are seen to perform better than the χ^2 tests and,

while still giving protection against a wide range of alternatives, to be at least moderately efficient with respect to tests for parametric departures.

The Appendix contains a summary of the asymptotic distribution theory of ψ_k^2 and $\psi_k^2(\hat{\lambda}_0)$ and discussion of some optimality properties of the tests.

THE EFFICIENCY OF SMOOTH GOODNESS-OF-FIT TESTS

by

Kenneth James Kopecky

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Completed August 1977

Commencement June 1978

APPROVED:

Redacted for privacy

Professor of Statistics

in charge of major

Redacted for privacy

Chairman of Department of Statistics

Redacted for privacy

Dean of Graduate School

Date thesis is presented August 12, 1977

Typed by Clover Redfern for Kenneth James Kopecky

ACKNOWLEDGMENT

I wish to record my appreciation of Professor Don Pierce, my major professor, for his continual and enthusiastic support of my academic career, and to the faculty of the Department of Statistics for their ready willingness to be of assistance.

This thesis is dedicated by my parents in recognition of their immeasurable loving impact on my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THE EFFICIENCY OF SMOOTH GOODNESS-OF-FIT TESTS

I. INTRODUCTION

I. 1. Goodness-of-Fit Tests

The application of any statistical procedure presupposes some model describing the observed phenomena. A simple and widely utilized type of model is the specification of a probability distribution for some appropriate random variable. Such a model might completely specify a single distribution or might encompass a family of probability laws which contains the correct but unknown distribution. The value of a statistical procedure in a given problem (as measured by its power, efficiency, etc.) depends on how nearly the unknown true probability law matches that the statistician assumed to be true in applying the procedure. Clearly, it is a problem of some importance to be able to evaluate how well a set of data agrees with the hypothesized probability distribution or class of distributions, i. e., to be able to test the goodness of fit of a distribution (or family) to a data set.

The development of goodness-of-fit tests has occurred throughout this century, and the statistician now has available a wide variety of different procedures. For example, Pearson's χ^2, as well as the Kolmogorov-Smirnov, Cramér-von Mises, Anderson-Darling, Watson, and other criteria are all appropriate for an arbitrary simple goodness-of-fit hypothesis. In most applications, however, the hypothesis is composite. Typically, it places the distribution in a class of probability laws indexed by unknown and unspecified parameter values. A common procedure is to replace the unknown "nuisance" parameter values by estimates and calculate Pearson's χ^2 statistic, which is taken to have a χ^2 distribution with degrees of freedom reduced by one for each parameter estimated. However, with commonly used estimators of the nuisance parameter, the statistic is no longer asymptotically distributed as χ^2; see Chernoff and Lehmann (1954) or Kendall and Stuart (1973; Sections 30.11-30.19). Lilliefors (1967, 1969), Durbin (1975), and Margolin and Maurer(1976) have considered the use of Kolmogorov-Smirnov statistics for composite hypotheses, especially in the cases of normality and exponentiality. Pearson and Hartley (1972; Table 54) give percentage points for several statistics in testing normality and exponentiality with unknown parameters.

Because of the central role it plays in many statistical models, the normal distribution has received special attention in the development of goodness-of-fit tests. For example, tests based on the sample skewness and kurtosis and the Shapiro-Wilk test are appropriate for testing normality, as are, of course, all those tests applicable to arbitrary continuous distributions. For discussion of testing for normality, the reader is referred to Bowman (1973), Bowman and Shenton (1973 JASA, 1973 Biometrika, 1975), D'Agostino (1970, 1971, 1973), D'Agostino and Pearson (1973), D'Agostino and Rosman (1974), D'Agostino and Tietjen (1971), Dyer (1974), Prescott (1976), Shapiro, Wilk and Chen (1968), Shenton and Bowman (1975), and Stephens (1974).

Many goodness-of-fit tests, including several mentioned above, share the characteristic of giving moderate protection against many alternatives at the expense of not giving very good protection against any particular alternative, which might be acceptable if one has no idea of the type of alternative likely to be encountered. Very often, however, some class of alternatives can be identified. In this case, one might be able to embed the null distribution (or class) in a parametric family in such a way that the goodness-of-fit hypothesis corresponds to a hypothesis about certain parameter values. So, for example, can one consider the hypothesis that a Weibull shape parameter is unity to be a hypothesis of goodness of fit of the exponential distribution. In this manner, one can often construct a test which has very good power against the parametric alternatives, although one must often sacrifice power against alternatives outside the given class.

It is a general rule that, as the class of alternatives increases, the power of tests designed for the class tends to decrease for individual alternatives, and this is exemplified in the discussion above: parametric tests have good power in relatively small classes of parametric alternatives, while the tests which seek to protect against all alternatives tend to have low power overall. The desire to find a compromise, a procedure which has reasonably good power over a relatively wide class of alternatives, leads to the smooth goodness-of-fit tests of Neyman and Barton. These tests are obtained by embedding the null density or class of densities in an exponential family containing the null density or class, which then corresponds to a hypothesis on the exponential family parameters, and also containing alternatives which are only slight or "smooth" departures from the null case. In this paper, the smooth goodness-of-fit tests are described in detail, and they are compared, chiefly in terms of efficiency, with procedures having wide-spread, moderate power, and with parametric procedures having highly focused power.

I. 2. Preview

Smooth goodness-of-fit tests are neither well-known nor widely used. For this reason the derivation of the tests for simple and composite hypotheses is presented in Chapter II. The details of the asymptotic distribution theory and optimality properties are postponed to the Appendix. The application of smooth goodness-of-fit tests to the multivariate normal case is also considered.

The comparisons of smooth tests with certain other parametric tests, in terms of power and efficiency, reduce to comparisons of noncentral χ^2 distributions with, in general, differing noncentrality parameters and degrees of freedom. Such comparisons not only provide a means for evaluating the smooth tests, but also aid in the selection of the order or degrees of freedom of the test. The derivations of the noncentrality parameters and measures of efficiency are contained in Chapter III, along with calculations for several examples.

Chapter IV consists of discussion of the results of Chapter III and some concluding remarks.

II. SMOOTH GOODNESS-OF-FIT TESTS

II. 1. Quadratic Score Statistics

Neyman's smooth goodness-of-fit test is obtained by embedding the distribution specified by the null hypothesis in a parametric class indexed by a vector parameter θ and then testing the hypothesis $H_0: \theta = 0$, which is equivalent to the original goodness-of-fit hypothesis. The test of θ is based on the quadratic score statistic, which is particularly convenient for the smooth test. In this section, the quadratic score statistics are discussed in general. Details of the asymptotic theory and optimality properties are left to the Appendix. In the subsequent sections of this chapter these results are applied to smooth goodness-of-fit tests and several important examples are considered.

Let $X_1, \ldots, X_n$ denote a random sample from a distribution with density function $g(x|\theta)$, where $\theta \in \mathbb{R}^k$. It is desired to test $H_0 : \theta = 0$. Suppressing the dependence on the observed $x_1, \ldots, x_n$, let $\ell.(\theta)$ denote the log likelihood of θ given $x_1, \ldots, x_n$, i.e.,

$$\ell.(\theta) = \sum_{i=1}^{n} \ln g(x_i | \theta) .$$

The subscripted dot will always denote summation over the n observations. For $j = 1, \ldots, k$ we define the efficient scores for the ith observation

$$U_{\theta_j}^{(i)} = U_{\theta_j}^{(i)}(0) = \partial \ln g(x_i \,|\, \theta)/\partial\theta_j \big|_{\theta=0}$$

and

$$U_{\cdot\theta_j} = U_{\cdot\theta_j}(0) = \sum_{i=1}^{n} U_{\theta_j}^{(i)} .$$

Finally, denote the vector of efficient scores by

$$U_{\cdot\theta} = U_{\cdot\theta}(0) = [U_{\cdot\theta_1}, \ldots, U_{\cdot\theta_k}]' .$$

Under suitable regularity conditions on the density $g(x\,|\,\theta)$

$$E_0 U_{\theta_j} = \int U_{\theta_j} g(x\,|\,\theta=0)dx = 0$$

$$i_{\theta_j\theta_j} = i_{\theta_j\theta_j}(0) = \mathrm{Var}_0\, U_{\theta_j} < \infty$$

for $j = 1, \ldots, k$, from which it follows that

$$i_{\theta_j\theta_\ell} = i_{\theta_j\theta_\ell}(0) = \mathrm{Cov}_0(U_{\theta_j}, U_{\theta_\ell}) < \infty.$$

Let the information matrix be denoted

$$i_{\theta\theta} = \begin{bmatrix} i_{\theta_1\theta_1} & \cdots & i_{\theta_1\theta_k} \\ & \ddots & \vdots \\ & & i_{\theta_k\theta_k} \end{bmatrix},$$

then the covariance matrix of $U._\theta$ is given by

$$i._{\theta\theta} = ni_{\theta\theta} .$$

In most cases of practical interest, $i_{\theta\theta}$ will be nonsingular; indeed this will be true for the goodness-of-fit statistics to be considered below. If, however, $i_{\theta\theta}$ is found to be singular, one can often reparameterize to a parameter, perhaps of lower dimension, for which the information matrix is nonsingular. Since $U_\theta^{(1)}, \ldots, U_\theta^{(n)}$ are independently and identically distributed random vectors with finite second moments, it follows by a multivariate central limit theorem, for example, Anderson (1958; Theorem 4.2.3), that $n^{-1/2}U._\theta$ is asymptotically distributed as normal with mean vector 0 and covariance matrix $i_{\theta\theta}$. Thus the quadratic score statistic

$$W_u = U'._\theta i.^{-1}_{\theta\theta} U._\theta$$

is asymptotically distributed as χ_k^2.

Rao (1948) suggested the use of W_u for testing the simple hypothesis $H_0 : \theta = 0$. In fact, W_u is asymptotically equivalent to Wald's statistic, $\hat{\theta}' i._{\theta\theta}(\hat{\theta})\hat{\theta}$, and to the likelihood ratio statistic,

$2[\ell.(\hat{\theta})-\ell.(0)]$, under standard regularity conditions; refer, for example, to Cox and Hinkley (1974; 9.3(ii)). Here $\hat{\theta}$ denotes the maximum likelihood estimator of θ. Because of this equivalence, the quadratic score statistic inherits the optimality properties verified by Wald (1943) for the other two statistics; refer to the Appendix. In addition W_u is often easier to calculate than either alternative statistic since, unlike the other two, W_u does not require the evaluation of $\hat{\theta}$. In the goodness-of-fit tests to be considered below, the parameter θ is of little direct interest, and the simplicity of W_u far outweighs any desire to estimate θ.

The utility of W_u for testing $H_0: \theta = 0$ arises from the fact that for θ^* near $\theta = 0$,

$$E_{\theta^*}U_{\cdot\theta} \doteq i_{\cdot\theta\theta}\theta^*$$

$$\mathrm{Var}_{\theta^*}U_{\theta} \doteq i_{\cdot\theta\theta},$$

so that for large n

$$n^{-1/2}U_{\cdot\theta} \mathrel{\dot\sim} N_k(n^{1/2}i_{\theta\theta}\theta^*, i_{\theta\theta}),$$

where $Z \mathrel{\dot\sim} \mathcal{L}$ means that Z is approximately distributed as $\mathcal{L}$, in which case,

$$W_u = U_{\cdot\theta}' i_{\cdot\theta\theta}^{-1} U_{\cdot\theta} \mathrel{\dot\sim} \chi^2_{k,\,\delta(\theta^*)}$$

where the noncentrality parameter is given by

$$\delta(\theta^*) = (ni_{\theta\theta}\theta^*)'(ni_{\theta\theta})^{-1}(ni_{\theta\theta}\theta^*) \tag{1.1}$$

$$= n\theta^{*\prime} i_{\theta\theta}\theta^* .$$

Of course, simple hypotheses rarely arise in practical situations, so a generalization of the results above to composite hypotheses is imperative. Rao (1948), Bartlett (1953), and Neyman (1959) each suggested the suitable modification of W_u which will be described below. Suppose now that $X_1, \ldots, X_n$ constitute a random sample from a distribution with density $g(x|\theta, \lambda)$, where $\theta \in \mathbb{R}^k$ is the parameter of interest, as before, while $\lambda \in \mathbb{R}^s$ is a nuisance parameter. It is desired to test $H_0: \theta = 0$, leaving λ unspecified. The likelihood is now a function of both θ and λ for the observed $x_1, \ldots, x_n$, and the efficient scores are also: we now consider $U_{\cdot\theta}(0, \lambda)$ and $U_{\cdot\lambda}(0, \lambda)$, where the λ-scores are defined in the same manner as the θ-scores. In addition, we deal with a partitioned covariance matrix:

$$\mathrm{Cov}_{0,\lambda}\begin{bmatrix} U_{\cdot\theta}(0,\lambda) \\ U_{\cdot\lambda}(0,\lambda) \end{bmatrix} = \begin{bmatrix} i_{\cdot\theta\theta}(0,\lambda) & i_{\cdot\theta\lambda}(0,\lambda) \\ & i_{\cdot\lambda\lambda}(0,\lambda) \end{bmatrix}.$$

We can define adjusted θ-scores by

$$U_{\cdot\theta|\lambda}(0,\lambda) = U_{\cdot\theta}(0,\lambda) - i_{\cdot\theta\lambda}(0,\lambda) i_{\cdot\lambda\lambda}^{-1}(0,\lambda) U_{\cdot\lambda}(0,\lambda).$$

Notice that $U_{\cdot\theta|\lambda}(0,\lambda)$ represents the residuals from the regression of $U_{\cdot\theta}(0,\lambda)$ on $U_{\cdot\lambda}(0,\lambda)$, i.e., that part of $U_{\cdot\theta}(0,\lambda)$ which is "free of" or orthogonal to $U_{\cdot\lambda}(0,\lambda)$. Similarly the adjusted information matrix is given by

$$\begin{aligned} i_{\cdot\theta\theta|\lambda}(0,\lambda) &= \mathrm{Cov}_{0,\lambda} U_{\cdot\theta|\lambda}(0,\lambda) \\ &= i_{\cdot\theta\theta}(0,\lambda) - i_{\cdot\theta\lambda}(0,\lambda) i_{\cdot\lambda\lambda}^{-1}(0,\lambda) i_{\cdot\lambda\theta}(0,\lambda). \end{aligned}$$

Following results in the Appendix, we consider a statistic

$$W_u(\lambda) = U'_{\cdot\theta|\lambda}(0,\lambda) i_{\cdot\theta\theta|\lambda}^{-1}(0,\lambda) U_{\cdot\theta|\lambda}(0,\lambda).$$

The dependence on the unknown λ is removed, and a practical statistic thereby obtained, by replacing λ with an estimator $\hat{\lambda}$ which is guaranteed to be sufficiently close to λ in a probabilistic sense. Under regularity conditions on $g(x|\theta,\lambda)$, the maximum likelihood estimator of λ under $H_0 : \theta = 0$, say $\hat{\lambda}_0$, may be used, in which case

$$U_{\cdot\theta|\lambda}(0,\hat{\lambda}_0) = U_{\cdot\theta}(0,\hat{\lambda}_0)$$

and the test of θ may be based on

$$W_u = W_u(\hat{\lambda}_0) = U'_{\cdot\theta}(0,\hat{\lambda}_0) i^{-1}_{\cdot\theta\theta|\lambda}(0,\hat{\lambda}_0) U_{\cdot\theta}(0,\hat{\lambda}_0) .$$

Moran (1970) demonstrated the asymptotic equivalence of $W_u(\hat{\lambda}_0)$ to Wald's statistic, $\hat{\theta}' i_{\cdot\theta\theta|\lambda}(\hat{\theta},\hat{\lambda})\hat{\theta}$, and to the likelihood ratio statistic, $2[\ell_{\cdot}(\hat{\theta},\hat{\lambda}) - \ell_{\cdot}(0,\hat{\lambda}_0)]$, under regularity conditions. Here $(\hat{\theta},\hat{\lambda})$ is the unrestricted joint maximum likelihood estimator. Wald (1943) demonstrated the asymptotic equivalence of the latter two statistics, as well as several optimality properties which are inherited by the quadratic score statistic $W_u(\hat{\lambda}_0)$.

As for the test of a simple hypothesis, $W_u(\hat{\lambda}_0)$ is distributed as χ^2_k under the null hypothesis. And for θ^* near 0, if $U_{\cdot\theta}(0,\lambda)$ and $U_{\cdot\lambda}(0,\lambda)$ have finite variances, then for large n

$$n^{-1/2} U_{\cdot\theta|\lambda}(0,\hat{\lambda}_0) \doteq N_k(n^{1/2} i_{\theta\theta|\lambda}(0,\lambda)\theta^*, i_{\theta\theta|\lambda}(0,\lambda)),$$

so that

$$W_u(\hat{\lambda}_0) \doteq \chi^2_{k,\,\delta(\theta^*|\lambda)}$$

where

$$\delta(\theta^*|\lambda) = n\theta^{*\prime} i_{\theta\theta|\lambda}(0,\lambda)\theta^* \tag{1.2}$$

For a more detailed discussion, refer to the Appendix. See also Neyman (1959) or Bhat and Nagnur (1965).

II. 2. Smooth Goodness-of-Fit Test: Simple Hypothesis

Neyman (1937) proposed what he referred to as the smooth goodness-of-fit test for a simple hypothesis. The procedure attempts to draw a compromise between the overall low power of omnibus tests and the highly focused power of parametric tests by establishing a reasonably rich family of parametric alternatives. The alternative densities are considered to be "smooth" departures from the null density in the sense that they remain close to the null density and cross it no more than a few times.

The alternatives are defined as follows. Suppose that under the null hypothesis, the distribution of a random variable X has density function $f(x)$. Let $F(x) = \int_{-\infty}^{x} f(u)du$ denote the cumulative distribution function of X. Then it is well known that if f is the true density, then the random variable $Y = F(X)$ is uniformly distributed on the interval $[0, 1]$. Neyman considered a class of alternatives for which the densities of Y, indexed by a vector parameter $\theta = (\theta_1, \ldots, \theta_k)' \in \mathbb{R}^k$, are given by

$$h(y|\theta) = \exp\left\{\sum_{i=1}^{k} \theta_i y^i - K(\theta)\right\}, \quad 0 \leq y \leq 1.$$

Here the normalizing constant is

$$K(\theta) = \ln\left[\int_0^1 \exp\left\{\sum_{i=1}^{k} \theta_i y^i\right\} dy\right].$$

The corresponding densities for X are given by

$$(2.1) \qquad g(x|\theta) = f(x)\exp\left\{\sum_{i=1}^{k} \theta_i F^i(x) - K(\theta)\right\}$$

and the hypothesis that the true density is f corresponds to $H_0: \theta = 0$. In fact, Neyman's model was parameterized in terms of modified Legendre polynomials in $y = F(x)$ in order to capitalize on the orthogonality of the efficient scores and obtain a statistic which is simply a sum of squares. However, the simplification obtained is of little value when calculations may be readily computerized, so the current parameterization will be employed, since it leads to clearer derivations.

Since $Y^* = \overline{F}(X) = 1 - F(X)$ is uniformly distributed on $(0, 1)$ as well as Y when F is the correct cumulative distribution function, the alternatives (2.1) could be written with $\overline{F}$ in place of F. This corresponds to a nonsingular linear transformation on the parameter space and leaves the goodness-of-fit tests to be described below unchanged. Some simplification of the test statistic may thereby be obtained; for example, the negative exponential distribution has $\overline{F}(x) = e^{-x}$.

Barton (1953) proposed a system of alternatives which are first order approximations of (2.1):

$$g^*(x \mid \theta) = f(x) \left\{ 1 + \sum_{i=1}^{k} \theta_i \pi_i[F(x)] \right\} \tag{2.2}$$

where $\pi_i(y)$ denotes the ith order modified Legendre polynomial in y. While this system provides some simplifications over Neyman's, it imposes constraints on the θ's, and we shall retain Neyman's parameterization.

From (2.1) it is easy to see that the alternatives are indeed smooth, in the sense described by Neyman, so long as θ is reasonably near 0 and k is not too large. In addition, even by taking k no greater than three or four one expects to obtain a rich family of alternatives. Neyman (1937; Fig. 4) illustrates two particular alternatives to normality. Figures 1 through 4 of this paper show how the Neyman alternatives with $k = 2$ depart from normality in four different "directions": when $\theta_2 = 0$, $\theta_1 = 0$, $\theta_1 = \theta_2$, and $\theta_1 = -\theta_2$. Figures 5 through 8 show the same departures from the negative exponential distribution. In each figure, the null density is represented by a dashed curve, while the alternatives, corresponding to θ-values of ± 0.2, ± 0.4, and ± 0.6, are represented by solid curves.

As a step in deriving the quadratic score statistic to test the goodness-of-fit hypothesis $H_0 : \theta = 0$, note that from Lehmann

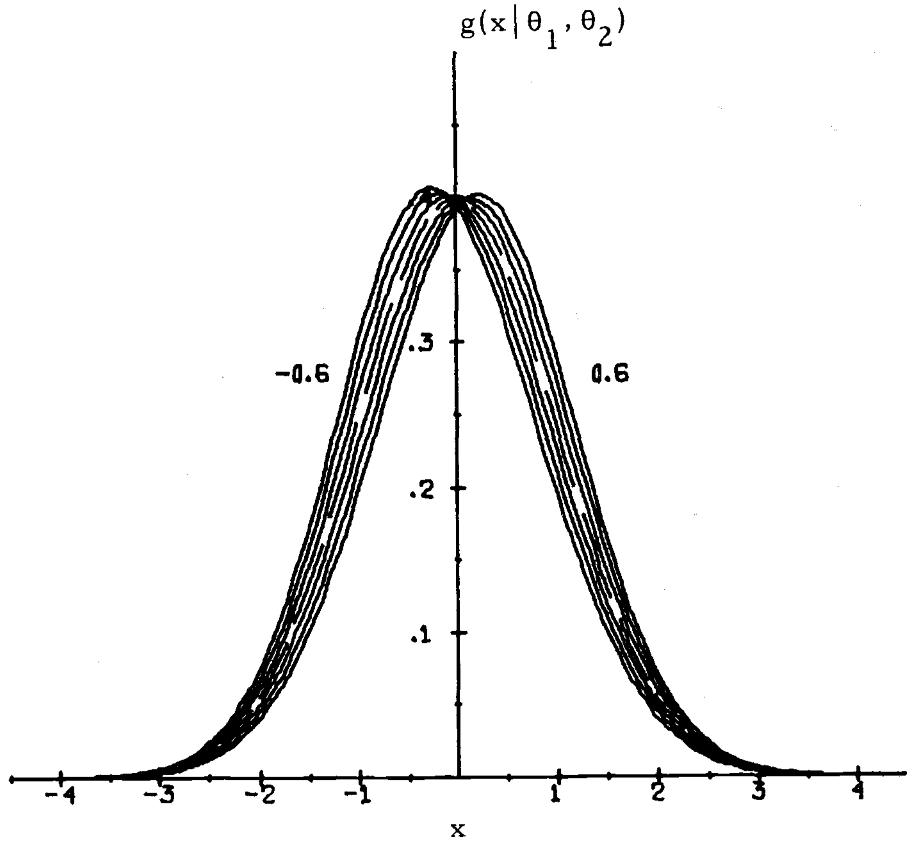

Figure 1. Neyman alternatives to normality:
$\theta_1 = -0.6(0.2)0.6$, $\theta_2 = 0$.

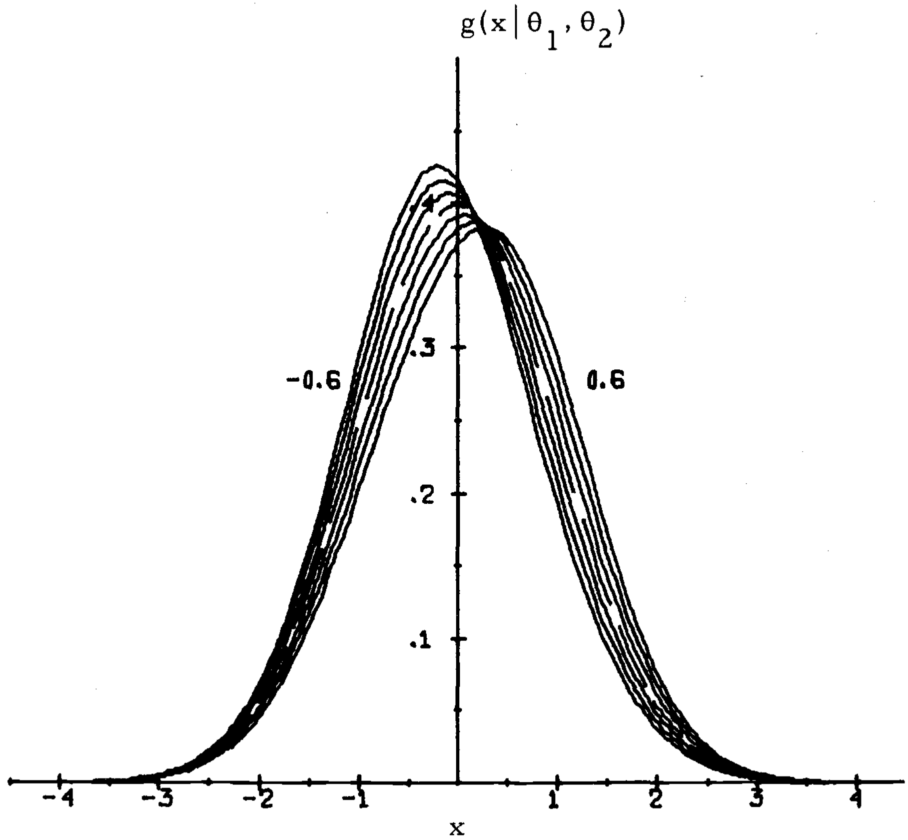

Figure 2. Neyman alternatives to normality:
$\theta_1 = 0,\ \theta_2 = -0.6(0.2)0.6.$

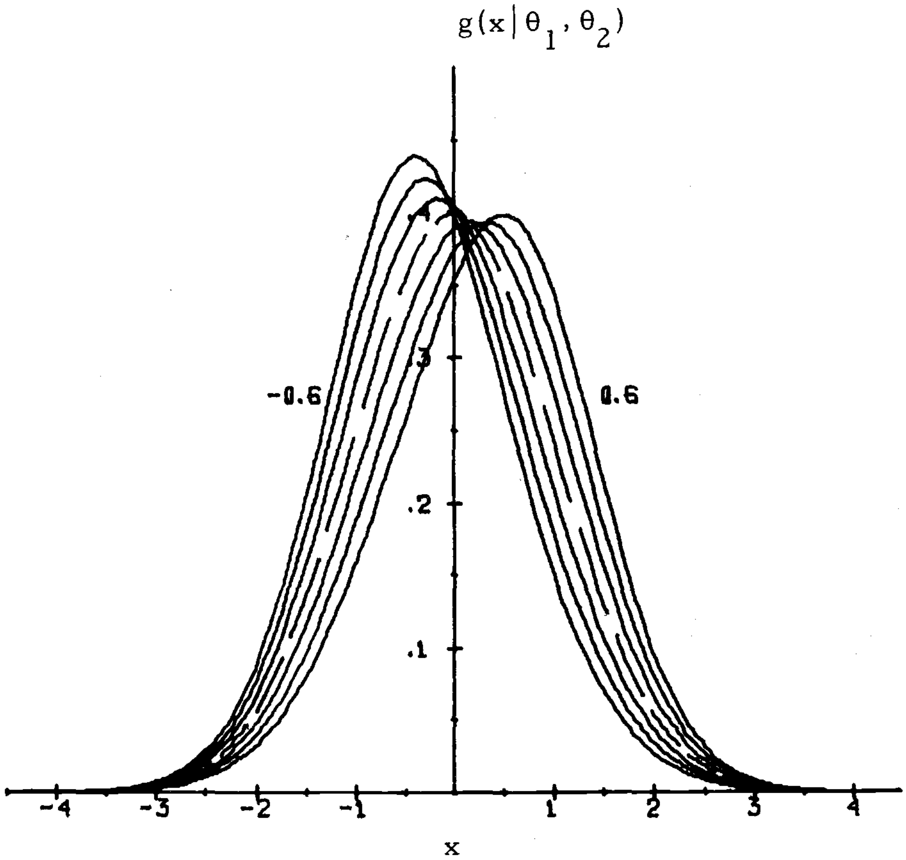

Figure 3. Neyman alternatives to normality: $\theta_1 = \theta_2 = -0.6(0.2)0.6$.

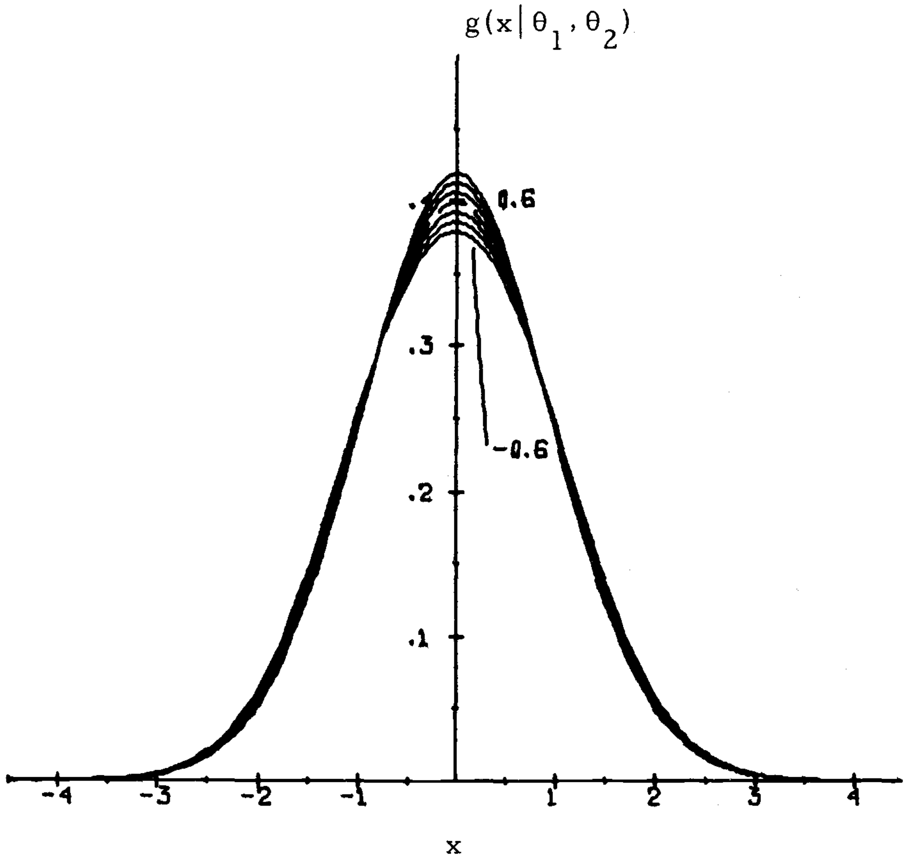

Figure 4. Neyman alternatives to normality: $\theta_1 = -\theta_2 = -0.6(0.2)0.6$.

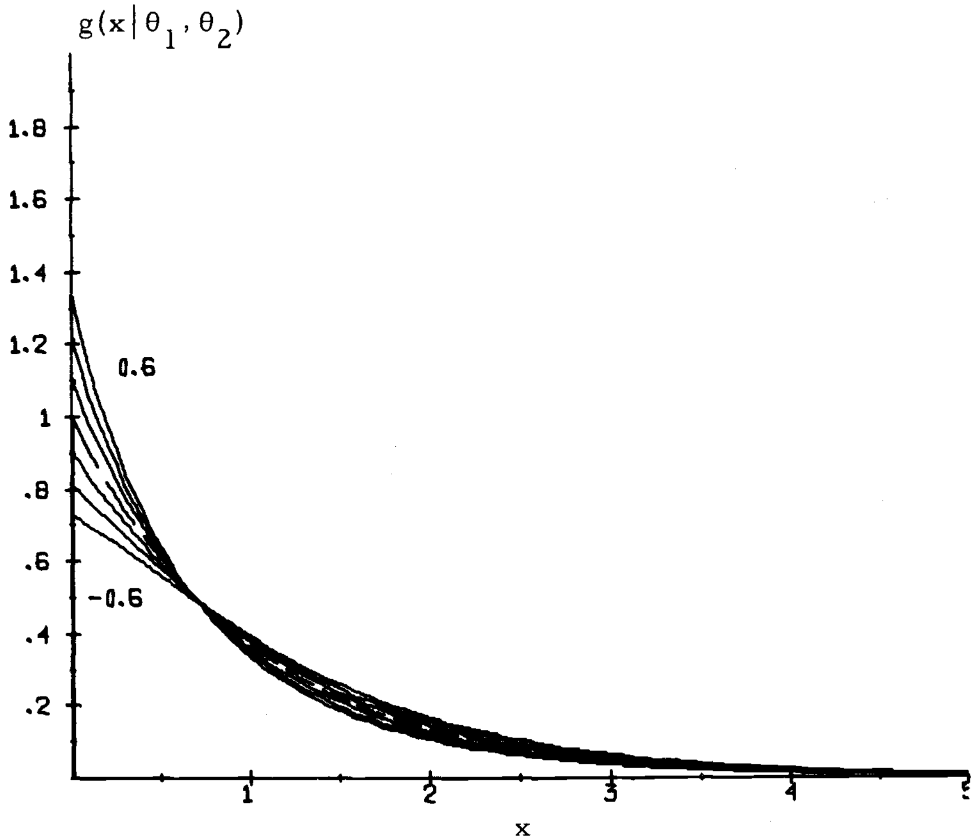

Figure 5. Neyman alternatives to exponentiality: $\theta_1 = -0.6(0.2)0.6$, $\theta_2 = 0$.

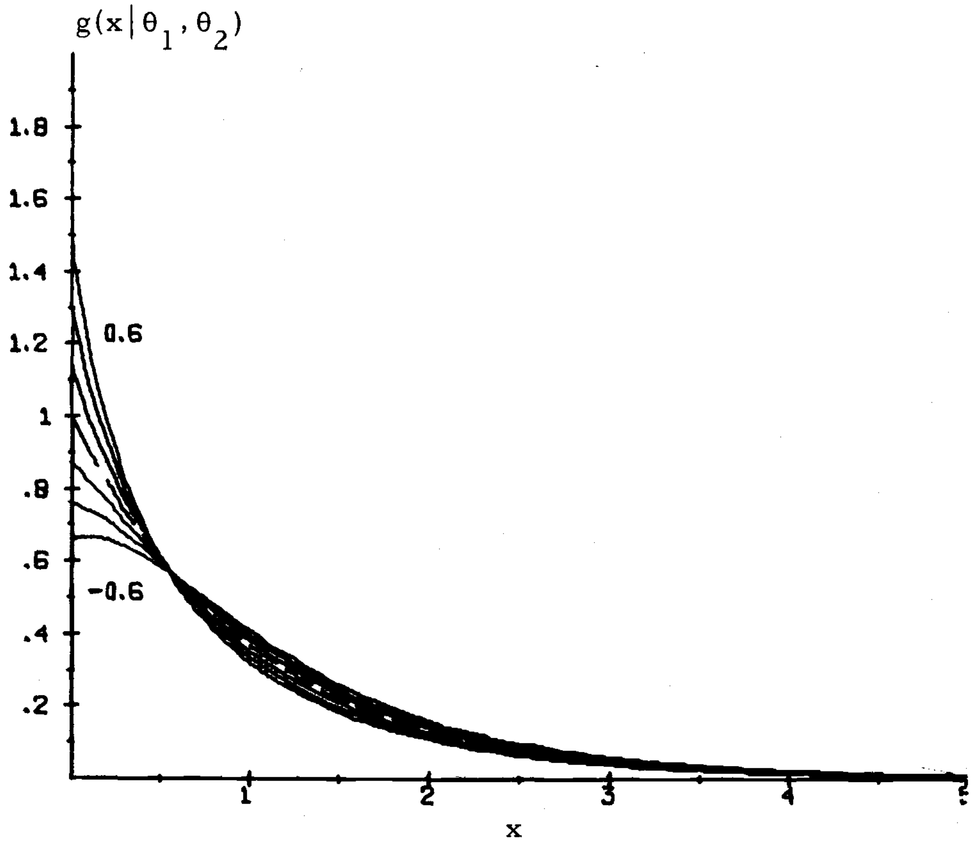

Figure 6. Neyman alternatives to exponentiality: $\theta_1 = 0$, $\theta_2 = -0.6(0.2)0.6$.

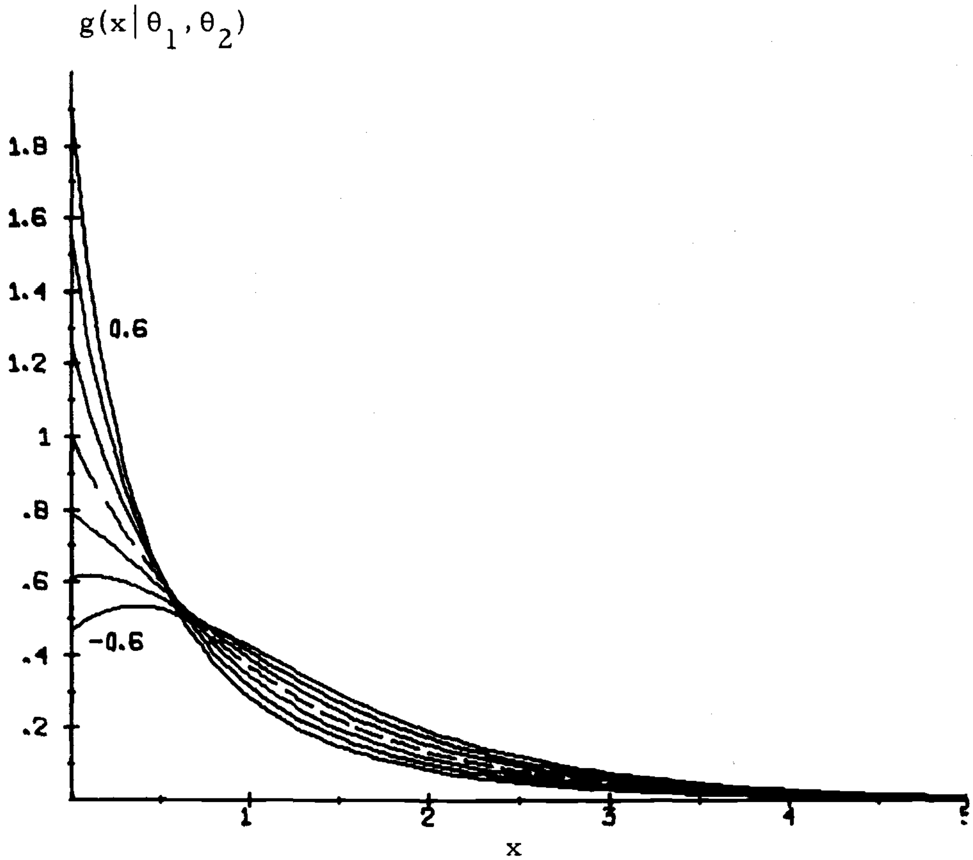

Figure 7. Neyman alternatives to exponentiality:
$\theta_1 = \theta_2 = -0.6(0.2)0.6.$

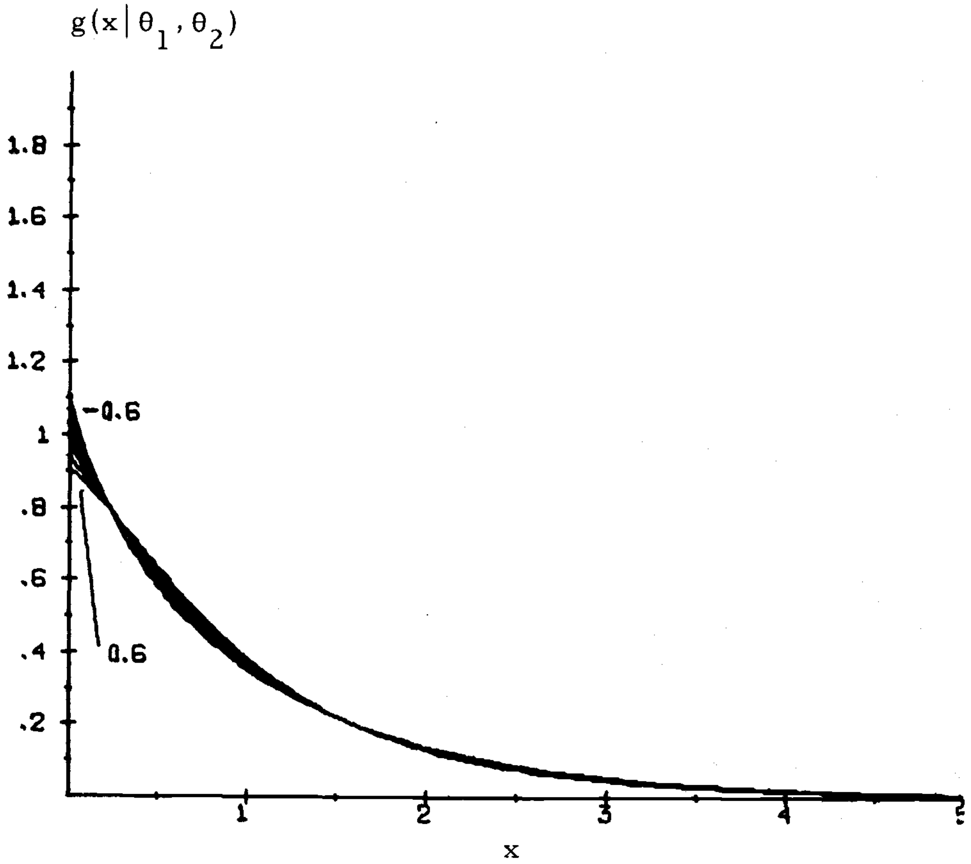

Figure 8. Neyman alternatives to exponentiality: $\theta_1 = -\theta_2 = -0.6(0.2)0.6$.

(1959; Theorem 9, p. 52)

$$\frac{\partial}{\partial\theta_j}\int_0^1 \exp\left\{\sum_{i=1}^k \theta_i y^i\right\} dy = \int_0^1 y^j \exp\left\{\sum_{i=1}^k \theta_i y^i\right\} dy$$

so

$$\frac{\partial}{\partial\theta_j} K(\theta)\Big|_{\theta=0} = \int_0^1 y^j dy \Big/ \exp\{K(0)\} = (j+1)^{-1}.$$

Since the log likelihood for θ is

$$\ell(\theta) = \ln f(x) + \sum_{i=1}^k \theta_i F^i(x) - K(\theta),$$

it follows that the efficient score for θ_j is

$$U_{\theta_j} = U_{\theta_j}(0) = F^j(x) - (j+1)^{-1}.$$

The information matrix is easily calculated since

$$\begin{aligned}\mathrm{Cov}_0(U_{\theta_j}, U_{\theta_\ell}) &= E_0[F^j(X)F^\ell(X)] - (j+1)^{-1}(\ell+1)^{-1} \\ &= (j+\ell+1)^{-1} - (j+1)^{-1}(\ell+1)^{-1} \\ &= j\ell(j+\ell+1)^{-1}(j+1)^{-1}(\ell+1)^{-1}.\end{aligned}$$

For k up to four, the information matrix is just the $k \times k$ upper left submatrix of

$$(2.3) \qquad i_{\theta\theta} = \begin{bmatrix} \frac{1}{12} & \frac{1}{12} & \frac{3}{40} & \frac{1}{15} \\ & \frac{4}{45} & \frac{1}{12} & \frac{8}{105} \\ & & \frac{9}{112} & \frac{3}{40} \\ & & & \frac{16}{225} \end{bmatrix}$$

Since the calculations essentially involve only the moments of $Y = F(X)$, the distribution of which is uniform on $[0, 1]$ for any F, the information matrix is the same regardless of the null distribution, and may be denoted simply $i_{\theta\theta}$. For k up to four, $i_{\theta\theta}^{-1}$ is given below:

<u>$k = 1$</u>

$$i_{\theta\theta}^{-1} = [12]$$

<u>$k = 2$</u>

$$i_{\theta\theta}^{-1} = \begin{bmatrix} 192 & -180 \\ & 180 \end{bmatrix}$$

<u>$k = 3$</u>

$$i_{\theta\theta}^{-1} = \begin{bmatrix} 1200 & -2700 & 1680 \\ & 6480 & -4200 \\ & & 2800 \end{bmatrix}$$

k = 4

$$i_{\theta\theta}^{-1} = \begin{bmatrix} 4800 & -18900 & 26880 & -12600 \\ & 79380 & -117600 & 56700 \\ & & 179200 & -88200 \\ & & & 44100 \end{bmatrix}$$

Also

$$U._{\theta} = \left[\sum_{i=1}^{n} F(x_i) - \frac{n}{2}, \ldots, \sum_{i=1}^{n} F^k(x_i) - \frac{n}{k+1} \right],$$

so the quadratic score statistic

$$\psi_k^2 = W_u = U!_{\theta} i._{\theta\theta}^{-1} U._{\theta}$$

is easily calculated. The symbol ψ_k^2, originated by Neyman, serves to emphasize k, the order or degrees of freedom of the test, and will be used in this paper.

The selection of k immediately presents itself as a crucial problem in the application of the smooth goodness-of-fit test. In accordance with the general rule mentioned above, it is necessary to seek a compromise between the rich class of alternatives provided by a large value of k and the better power characteristics provided by a small value of k. Neyman (1937) and Pearson (1938) pointed out that for $k^* \leq k_1 < k_2$, the k_2-order test will be less sensitive than the k_1-order test against alternatives which are modeled by the polynomial of order k^* in (2.1). Moreover, the test based on ψ_k^2 will have no power against departures of $\theta_{k'}$ from 0 when $k' > k$. Neyman based these comments on an examination of the asymptotic power function, which he showed by a direct argument to be of the noncentral χ^2 form with noncentrality parameter given by (1.1). Of course, to insure a test of the desired size, k must be selected independently of the data, and in the absence of special knowledge about a particular problem, Neyman recommended that k generally need not be greater than four. However, the results of the following chapter will suggest that one might do better to use a smaller value of k.

Pearson (1938) noted the asymptotic equivalence of Neyman's smooth goodness-of-fit test to the likelihood ratio test. Neyman (1937) showed that the smooth test had, locally and asymptotically, the property later defined by Wald (1943) to be "uniformly best constant power" on the contours defined by setting the noncentrality parameter equal to a constant; i.e., asymptotically the test is locally uniformly most powerful among all tests of $H_0: \theta = 0$ which have power functions depending on θ only through $\delta(\theta) = n\theta' i_{\theta\theta}\theta$. Because of its asymptotic χ^2 distribution, the test is consistent and therefore asymptotically unbiased. In fact, for $k = 1$, the test is asymptotically uniformly most powerful among all unbiased tests for $H_0: \theta_1 = 0$;

see, for example, Lehmann (1959; Section 4.2).

David (1939) calculated exact moments for ψ_1^2 and ψ_2^2, and Barton (1953), using these, fitted Pearson curves and an Edgeworth expansion to approximate the distributions of ψ_1^2 and ψ_1, respectively, in addition to calculating the exact distribution of ψ_1^2. Based on these he tabulated exact and approximate critical points for ψ_1^2 when n is small; refer to his Table II. Barton (1955) also attempted to generalize the smooth goodness-of-fit test to grouped and discrete data.

II.3. Smooth Goodness-of-Fit Test: Composite Hypothesis

Barton (1956), Javitz (1975), and Thomas and Pierce (1977) considered generalizations of Neyman's smooth test to the more useful case of a composite null hypothesis. Suppose we wish to test whether the distribution of X has density $f(x|\lambda)$, where $\lambda \in \mathbb{R}^s$ is to be left unspecified. For example, we might wish to determine if X is normally distributed, regardless of the values of the mean and variance. As before, we consider the class of alternatives with densities

$$g(x|\theta, \lambda) = f(x|\lambda) \exp\left\{\sum_{i=1}^{k} \theta_i F^i(x|\lambda) - K(\theta)\right\},$$

where $F(x|\lambda)$ is the cumulative distribution function for X under

λ. The normalizing constant $K(\theta)$ is free of λ since, substituting $y = F(x|\lambda)$ for any λ,

$$K(\theta) = \ln\left[\int_{-\infty}^{\infty} f(x|\lambda)\exp\left\{\sum_{i=1}^{k} \theta_i F^i(x|\lambda)\right\} dx\right]$$

$$= \ln\left[\int_{-\infty}^{\infty} \exp\left\{\sum_{i=1}^{k} \theta_i y^i\right\} dy\right].$$

Proceeding as in the previous section we find that ψ_k^2 is a function of λ. Barton's procedure was to substitute an estimate of λ in the statistic, which then was, he showed, distributed as a linear combination of χ^2 random variables. As an alternative to this rather complicated criterion, Javitz and, independently, Thomas and Pierce, proposed using the quadratic score statistic as described in the first section of this chapter. Thus the test is based on

$$\psi_k^2(\hat{\lambda}_0) = W_u(\hat{\lambda}_0) = U'_{\cdot\theta}(0,\hat{\lambda}_0)i^{-1}_{\cdot\theta\theta|\lambda}(0,\hat{\lambda}_0)U_{\cdot\theta}(0,\hat{\lambda}_0) \tag{3.1}$$

where

$$U_{\cdot\theta}(0,\hat{\lambda}_0) = \left[\sum_{i=1}^{n} F(x_i|\hat{\lambda}_0) - \frac{n}{2}, \ldots, \sum_{i=1}^{n} F^k(x_i|\hat{\lambda}_0) - \frac{n}{k+1}\right]'$$

and

$$i_{\cdot\theta\theta|\lambda}(0,\hat{\lambda}_0) = n[i_{\theta\theta}(0,\hat{\lambda}_0) - i_{\theta\lambda}(0,\hat{\lambda}_0)i^{-1}_{\lambda\lambda}(0,\hat{\lambda}_0)i_{\lambda\theta}(0,\hat{\lambda}_0)].$$

Here $\hat{\lambda}_0$ is the maximum likelihood estimator of λ under the null hypothesis $H_0: \theta=0$. Recall from Section 1 of this chapter that if $\tilde{\lambda}$ is a consistent estimator, not necessarily the maximum likelihood estimator, then the adjusted efficient score vector

$$U_{\cdot\theta|\lambda}(0,\tilde{\lambda}) = U_{\cdot\theta}(0,\tilde{\lambda}) - i_{\theta\lambda}(0,\tilde{\lambda})i_{\lambda\lambda}^{-1}(0,\tilde{\lambda})U_{\cdot\lambda}(0,\tilde{\lambda})$$

should be used in (3. 1) in place of $U_{\cdot\theta}(0,\hat{\lambda}_0)$.

It is an identity of linear algebra that, dropping the arguments $(0,\hat{\lambda}_0)$,

$$\begin{aligned} i_{\theta\theta|\lambda}^{-1} &= (i_{\theta\theta}-i_{\theta\lambda}i_{\lambda\lambda}^{-1}i_{\lambda\theta})^{-1} \\ &= i_{\theta\theta}^{-1} + i_{\theta\theta}^{-1}i_{\theta\lambda}(i_{\lambda\lambda}-i_{\lambda\theta}i_{\theta\theta}^{-1}i_{\theta\lambda})^{-1}i_{\lambda\theta}i_{\theta\theta}^{-1}\,; \end{aligned}$$

see, for example, Rao (1973; p. 33, problem 2.7). Thus the quadratic score statistic is just Barton's statistic plus a term to correct for the estimation of λ.

It will be seen in the examples below that $i_{\theta\theta|\lambda}(0,\lambda)$ is independent of λ in each case considered. This greatly simplifies computation of the statistic, since then the matrix of the quadratic form depends only on the null distribution, and not at all on the data. This is easily shown to be true in general for location-scale families, with $f(x|\mu,\sigma) = \sigma^{-1}f[(x-\mu)/\sigma|0,1]$, and for shape-scale families, with $f(x|\tau,\sigma) = \tau x^{\tau-1}\sigma^{-\tau}f[(x/\sigma)^{\tau}|1,1]$. As mentioned in Section 2 of this chapter, since the calculation of moments of the θ-scores involves essentially only the uniformly distributed random variables, $Y = F(X|\lambda)$ in this case, the information matrix $i_{\theta\theta}$ is the same for any F and any λ, and is given by (2. 3). In addition, since

$$\ln g(x|\theta,\lambda) = \ln f(x|\lambda) + \sum_{i=1}^{k} \theta_i F^i(x|\lambda) - K(\theta),$$

the vector of efficient scores for λ, obtained by differentiating and then evaluating at $\theta = 0$, and $i_{\lambda\lambda}$, the covariance matrix, are the same for the Neyman family of densities $g(x|\theta,\lambda)$ as for the parametric family of densities $f(x|\lambda)$. Finally, notice that

$$i_{\theta_i\lambda_j} = -E_{0,\lambda}[\partial^2 \ln g(X|0,\lambda)/\partial\theta_i\partial\lambda_j] = -E[\partial F^i(X|\lambda)/\partial\lambda_j] .$$

From Section 1 of this chapter we know that, asymptotically, $\psi_k^2(\hat{\lambda}_0)$ is distributed as $\chi^2_{k,\delta(\theta|\lambda)}$ where

$$\delta(\theta|\lambda) = n\theta' i_{\theta\theta|\lambda}(0,\lambda)\theta ,$$

for local alternatives. The comments in Section 2 of this chapter regarding the selection of k apply in this case as well. To compromise between the effects of selecting a high or low value of k, one might be guided by considerations of efficiency, as in the following chapter.

II. 4. Calculation of $i^{-1}_{\theta\theta|\lambda}$ for Examples

A. Normal Distribution

In this case the density is well-known:

$$f(x|\mu,\sigma) = \sigma^{-1}(2\pi)^{-1/2}\exp\{-(x-\mu)^2/2\sigma^2\} .$$

We take $\lambda = (\mu,\sigma)'$. Then

$$U_\mu(0,\lambda) = (x-\mu)\sigma^{-2}$$

$$U_\sigma(0,\lambda) = (x-\mu)^2\sigma^{-3} - \sigma^{-1}$$

and

$$i_{\lambda\lambda} = \sigma^{-2}\begin{bmatrix} 1 & 0 \\ & 2 \end{bmatrix} \tag{4. 1}$$

Moreover, from Barton (1953; Table X), it follows that, for k up to four,

$$i_{\theta\lambda} = \sigma^{-1}\begin{bmatrix} \frac{1}{2\pi^{1/2}} & 0 \\ \frac{1}{2\pi^{1/2}} & \frac{\sqrt{3}}{6\pi} \\ \frac{3S_3+18\pi}{40\pi^{3/2}} & \frac{\sqrt{3}}{4\pi} \\ \frac{3S_3+8\pi}{20\pi^{3/2}} & \frac{\sqrt{3}\,(S_4+8\pi)}{28\pi^2} \end{bmatrix} \tag{4. 2}$$

where

(4.3) $$S_3 = 10 \arcsin(\tfrac{1}{3}) - \pi = 0.256776441$$

$$S_4 = 14 \arcsin(\tfrac{1}{4}) - \pi = 0.395930917.$$

So, for k up to four, recalling that $i_{\theta\theta}$ is given by (2.3),

$$i_{\theta\theta|\lambda} = i_{\theta\theta} - i_{\theta\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\theta}$$

$$= \begin{bmatrix} 3.755862 & 3.755862 & 2.404642 & 1.053422 \\ & 5.089701 & 4.405401 & 3.225992 \\ & & 4.632428 & 4.116791 \\ & & & 4.211002 \end{bmatrix} \times 10^{-3}.$$

For any k, $i_{\theta\theta|\lambda}^{-1}$ is just the inverse of the upper left $k \times k$ submatrix of $i_{\theta\theta|\lambda}$ given above. These inverses are listed in Table 1. Of course the noncentrality parameter $\delta(\theta|\lambda) = n\theta' i_{\theta\theta|\lambda}\theta$ is calculated from the submatrix from $i_{\theta\theta|\lambda}$ of the appropriate order.

B. Exponential Distribution

The density is $f(x|\lambda) = \lambda^{-1}\exp\{-x/\lambda\}$, where the nuisance parameter is λ, the scale parameter. We have

$$U_\lambda(0,\lambda) = \lambda^{-2}(x-\lambda)$$

so

Table 1. Matrices $i^{-1}_{\theta\theta|\lambda}$ for testing normality.

k = 1

$$\begin{bmatrix} 266.2504789951 \end{bmatrix}$$

k = 2

$$\begin{bmatrix} 1015.965828673 & -749.7153496774 \\ & -749.7153496774 \end{bmatrix}$$

k = 3

$$\begin{bmatrix} 9072.727318174 & -14806.08191679 & 9370.911044737 \\ & 25273.39540729 & -16349.1200384 \\ & & 10899.41335893 \end{bmatrix}$$

k = 4

$$\begin{bmatrix} 23058.98126537 & -66471.72954001 & 84729.69856163 & -37679.39415895 \\ & 216127.8696941 & -294726.7739746 & 139188.8284477 \\ & & 416937.1470859 & -203018.8690217 \\ & & & 101509.43559 \end{bmatrix}$$

$$i_{\lambda\lambda} = \lambda^{-2}.$$

Recall from Section 2 that it is sometimes convenient to use $\overline{F}(x)$ in place of $F(x)$, so we take

$$U_{\theta_i}(0,\lambda) = \overline{F}^i(x|\lambda) - (i+1)^{-1} = e^{-ix/\lambda} - (i+1)^{-1}.$$

Thus

$$\mathrm{Cov}(U_\lambda, U_{\theta_i} | 0,\lambda) = \int_0^\infty \frac{x}{\lambda^2} e^{-ix/\lambda}(\frac{1}{\lambda})e^{-x/\lambda}dx - \frac{1}{\lambda(i+1)}$$

$$= \frac{1}{\lambda^2(i+1)}\int_0^\infty x(\frac{i+1}{\lambda})e^{-(i+1)x/\lambda}dx - \frac{1}{\lambda(i+1)}$$

$$= \frac{1}{\lambda^2(i+1)} \times \frac{\lambda}{i+1} - \frac{1}{\lambda(i+1)} = -\frac{1}{\lambda}\frac{i}{(i+1)^2}.$$

Thus, for k up to four,

$$i'_{\theta\lambda}(0,\lambda) = \lambda^{-1}[-\frac{1}{4}, -\frac{2}{9}, -\frac{3}{16}, -\frac{4}{25}]$$

and

$$i_{\theta\theta|\lambda} = \begin{bmatrix} \frac{1}{48} & \frac{1}{36} & \frac{9}{320} & \frac{2}{75} \\ & \frac{16}{405} & \frac{1}{24} & \frac{64}{1575} \\ & & \frac{81}{1792} & \frac{9}{200} \\ & & & \frac{256}{5625} \end{bmatrix}.$$

The inverses of $i_{\theta\theta|\lambda}$ for k up to four are given in Table 2.

Table 2. Matrices $i^{-1}_{\theta\theta|\lambda}$ for testing exponentiality.

k = 1

$$\begin{bmatrix} 48 \end{bmatrix}$$

k = 2

$$\begin{bmatrix} 768 & -540 \\ & 405 \end{bmatrix}$$

k = 3

$$\begin{bmatrix} 4800 & -8100 & 4480 \\ & 14580 & -8400 \\ & & 4977.\overline{7} \end{bmatrix}$$

k = 4

$$\begin{bmatrix} 19200 & -56700 & 71680 & -31500 \\ & 178605 & -235200 & 106312.5 \\ & & 193137.\overline{7} & 147000 \\ & & & 68906.25 \end{bmatrix}$$

C. Weibull Distribution

For the density $f(x|\tau,\sigma) = \tau x^{\tau-1}\sigma^{-\tau}\exp\{-(x/\sigma)^{\tau}\}$ we have for $\lambda = (\tau,\sigma)'$

$$U_{\tau}(0,\lambda) = \tau^{-1} + [1-(x/\sigma)^{\tau}]\ \ln(x/\sigma)$$

$$U_{\sigma}(0,\lambda) = \tau\sigma^{-1}[(x/\sigma)^{\tau}-1].$$

From

$$\partial^2 \ell(\tau,\sigma)/\partial\tau^2 = -\tau^{-2} - (x/\sigma)^\tau \ln^2(x/\sigma) = -\tau^{-2}\{1+(x/\sigma)^\tau \ln^2[(x/\sigma)^\tau]\}$$

it follows that, letting $y = (x/\sigma)^\tau$,

$$i_{\tau\tau} = E(-\partial^2 \ell/\partial\tau^2 \mid \tau,\sigma) = \tau^{-2} + \tau^{-2}\int_0^\infty y(\ln y)^2 e^{-y} dy$$

$$= \tau^{-2}[1+\Gamma''(2)] = 1.828681\tau^{-2}.$$

Also, since

$$\partial^2 \ell(\tau,\sigma)/\partial\sigma^2 = \tau\sigma^{-2} - \tau(\tau+1)\sigma^{-2}(x/\sigma)^\tau$$

we have

$$i_{\sigma\sigma} = \tau(\tau+1)\sigma^{-2}E[(X/\sigma)^\tau \mid \tau,\sigma] - \tau\sigma^{-2} = \tau(\tau+1)\sigma^{-2}E(X\mid 1,1) - \tau\sigma^2$$

$$= \tau^2\sigma^{-2}.$$

And from

$$\partial^2 \ell(\tau,\sigma)/\partial\tau\partial\sigma = -\sigma^{-1} + \sigma^{-1}(x/\sigma)^\tau + \tau\sigma^{-1}(x/\sigma)^\tau \ln(x/\sigma)$$

it follows that

$$i_{\tau\sigma} = \sigma^{-1} - \sigma^{-1}E[(X/\sigma)^\tau \mid \tau,\sigma] - \sigma^{-1}E\{(X/\sigma)^\tau \ln[(X/\sigma)^\tau] \mid \tau,\sigma\}$$

$$= \sigma^{-1} - \sigma^{-1}E(X\mid 1,1) - \sigma^{-1}E(X \ln X \mid 1,1)$$

$$= -\sigma^{-1}\int_0^\infty x(\ln x)e^{-x}dx = -\sigma^{-1}(1-\gamma)$$

where $\gamma = 0.577216\ldots$ is Euler's constant. So

$$i_{\tau\sigma} = -.422784\sigma^{-1}$$

and

$$i_{\lambda\lambda} = \begin{bmatrix} 1.828681\tau^{-2} & -0.422784\sigma^{-1} \\ & \tau^2\sigma^{-2} \end{bmatrix} .$$

As for the negative exponential, it is convenient to write the Neyman alternatives in terms of $\overline{F}(x) = \exp\{-(x/\sigma)^{\tau}\}$, so we have

$$U_{\theta_i}(0,\lambda) = \exp\{-i(x/\sigma)^{\tau}\} - (i+1)^{-1} .$$

Therefore, letting $y = (x/\sigma)^{\tau}$,

$$i_{\theta_i,\sigma} = \mathrm{Cov}[U_{\theta_i}(0,\lambda), U_{\sigma}(0,\lambda) \mid 0,\lambda]$$

$$= E[\exp\{-i(X/\sigma)^{\tau}\}\tau\sigma^{-1}(X/\sigma)^{\tau} \mid 0,\lambda] - \tau\sigma^{-1}(i+1)^{-1}$$

$$= \tau\sigma^{-1}\int_0^{\infty} ye^{-(i+1)y}dy - \tau\sigma^{-1}(i+1)^{-1} = -\tau\sigma^{-1}i(i+1)^{-2} .$$

Also, letting $u = (i+1)y$

$$i_{\theta_i,\tau} = \mathrm{Cov}[U_{\theta_i}(0,\lambda), U_{\tau}(0,\lambda) \mid 0,\lambda]$$

$$= E[\exp\{-i(X/\sigma)^{\tau}\}\ln(X/\sigma) \mid 0,\lambda]$$

$$- E[\exp\{-i(X/\sigma)^{\tau}\}(X/\sigma)^{\tau}\ln(X/\sigma) \mid 0,\lambda] + \tau^{-1}(i+1)^{-1} =$$

$$= \tau^{-1}\int_0^\infty e^{-(i+1)y}\ln y dy - \tau^{-1}\int_0^\infty ye^{-(i+1)y}\ln y dy + \tau^{-1}(i+1)^{-1}$$

$$= \tau^{-1}(i+1)^{-1}\left\{\int_0^\infty [\ln u - \ln(i+1)]e^{-u}du \right.$$

$$\left. - \int_0^\infty u(i+1)[\ln u - \ln(i+1)]e^{-u}du + 1\right\}$$

$$= \tau^{-1}(i+1)^{-1}\{-\gamma-\ln(i+1)-(i+1)^{-1}[1-\gamma-\ln(i+1)]+1\}$$

$$= i[1-\gamma-\ln(i+1)]\tau^{-1}(i+1)^{-2} .$$

So

$$i_{\theta\lambda} = \begin{bmatrix} -.067591/\tau & -\tau/4\sigma \\ -.150184/\tau & -2\tau/9\sigma \\ -.180658/\tau & -3\tau/16\sigma \\ -.189865/\tau & -4\tau/25\sigma \end{bmatrix}$$

and

$$i_{\theta\theta|\lambda} = \begin{bmatrix} 2.641897 & 2.145342 & 0.833043 & -0.371844 \\ & 3.390813 & 3.213743 & 2.539434 \\ & & 4.259354 & 4.439160 \\ & & & 5.327499 \end{bmatrix} \times 10^{-3}$$

For k up to four, $i^{-1}_{\theta\theta|\lambda}$ is given in Table 3.

Table 3. Matrices $i^{-1}_{\theta\theta|\lambda}$ for testing the Weibull distribution.

k = 1

$$\begin{bmatrix} 378.515892 \end{bmatrix}$$

k = 2

$$\begin{bmatrix} 778.479599 & -492.538214 \\ & 606.539764 \end{bmatrix}$$

k = 3

$$\begin{bmatrix} 10276.398086 & -16135.879709 & 10164.895656 \\ & 26371.567486 & -16741.871835 \\ & & 10878.710303 \end{bmatrix}$$

k = 4

$$\begin{bmatrix} 19600.422408 & -54730.219668 & 67235.294397 & -28568.029642 \\ & 186122.660513 & -252969.738092 & 118249.825381 \\ & & 360194.660129 & -174858.922126 \\ & & & 87530.050266 \end{bmatrix}$$

II.5. The Possibility of a Smooth Test for Multivariate Normality

One attractive feature of the Neyman smooth test is its potential for applications to problems beyond the standard goodness-of-fit test. Thomas (1977) considers smooth goodness-of-fit tests for censored data. Javitz (1975) considered smooth tests for independence and equality of distributions, as well as a smooth test for bivariate goodness of fit. Almost all widely used multivariate techniques are based on the assumption of multivariate normality, which therefore has been the focus of nearly all the attention given to multivariate goodness-of-fit tests. Several goodness-of-fit procedures have been discussed by Hensler, Mehrota, and Michalek (1977), Kowalski (1970), Malkovich and Afifi (1973), and Mardia (1974); see also the reviews of Andrews, Gnanadesikan, and Warner (1973) or Gnanadesikan (1977).

For simplicity of exposition, we will consider the case of bivariate normality. Further generalization to the p-variate normal distribution will then be obvious. Moreover, we will consider first the composite null hypothesis that (X_1, X_2) has a bivariate normal distribution with the nuisance parameter $\lambda' = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ unknown, where ρ denotes the correlation coefficient of X_1 and X_2. Let $\phi_2(x_1, x_2 | \lambda)$ denote the joint density

$$\phi_2(x_1, x_2 | \lambda) = (2\pi)^{-1}\sigma_1^{-1}\sigma_2^{-1}(1-\rho^2)^{-1/2}$$

$$\times \exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x_1-\mu_1}{\sigma_1})^2 - 2\rho(\frac{x_1-\mu_1}{\sigma_1})(\frac{x_2-\mu_2}{\sigma_2})$$

$$+ (\frac{x_2-\mu_2}{\sigma_2})^2]\}$$

and let the marginal probability transform of X_i be denoted by $Y_i = \Phi[(X_i-\mu_i)/\sigma_i]$ for $i = 1, 2$. We consider the family of smooth alternative densities

$$g(x_1, x_2 | \theta, \lambda) = \phi_2(x_1, x_2 | \lambda) \exp\{P_k(\theta, y_1, y_2) - K(\theta)\} \tag{5.1}$$

where P_k is a polynomial of degree k with coefficients given by θ, and $K(\theta)$ is the usual normalizing constant. For example, taking $k = 2$ and letting $\theta' = (\theta_{10}, \theta_{20}, \theta_{01}, \theta_{02}, \theta_{11})$ we have

$$P_k(\theta, y_1, y_2) = \theta_{10}y_1 + \theta_{20}y_1^2 + \theta_{01}y_2 + \theta_{02}y_2^2 + \theta_{11}y_1y_2. \tag{5.2}$$

Javitz (1975) parameterized the alternatives (5.1) in terms of the modified Legendre polynomials in y_1 and in y_2, but that corresponds simply to a linear transformation on the parameter space so long as k is fixed.

Javitz mistakenly asserts that, under the null hypothesis, which may now be written $H_0 : \theta = 0$, the joint distribution of (Y_1, Y_2) is uniform on the unit interval. This implies that Y_1 and Y_2 are independent, which is clearly true if and only if $\rho = 0$. In general, however, the joint density of (Y_1, Y_2) is

$$h(y_1, y_2|\lambda) = (1-\rho^2)^{-1/2}\exp\{-\frac{\rho}{2(1-\rho^2)}[\rho z_1^2 - 2z_1 z_2 + \rho z_2^2]\} \tag{5.3}$$

where $z_i = \Phi^{-1}(y_i)$. The alternative densities (5.1) are given in terms of the y's by

$$h^*(y_1, y_2|\theta, \lambda) = h(y_1, y_2|\lambda)\exp\{P_k(\theta, y_1, y_2) - K(\theta)\}.$$

So, although the null distribution of the Y's is no longer uniform, as in the univariate case, it is still true that the alternatives are smooth departures: the log-likelihood ratio $\ln\{h^*(y_1, y_2|\theta, \lambda)/h^*(y_1, y_2|0, \lambda)\}$ is simply a polynomial in the y's.

The fact that the probability integral transformation in the univariate case produces a random variable with a uniform distribution on $(0, 1)$ provides a major simplification for the smooth test: as noted before, $i_{\theta\theta}$ depends on neither the null hypothesis nor on λ, and is very easily calculated from the moments of the uniform distribution. From (5.3) we see that this simplification no longer occurs in the multivariate case, at least when $\rho \neq 0$. In fact, if $E[T(Z)|\mu, \sigma^2]$ denotes the expectation of $T(Z)$ when $Z \sim N(\mu, \sigma^2)$, then

$$E\{\Phi^i(\frac{X_1-\mu_1}{\sigma_1})\Phi^j(\frac{X_2-\mu_2}{\sigma_2})|\lambda\} = E\{\Phi^i(Z_1)E[\Phi^j(Z_2)|\rho Z_1, 1-\rho^2]|0,1\}$$

for $i, j \geq 1$, and therefore

$$\begin{aligned} i_{\theta_{ij}\theta_{\ell m}}(\rho) &= E\{\Phi^{i+\ell}(Z_1)E[\Phi^{j+m}(Z_2)|\rho Z_1, 1-\rho^2]|0,1\} \\ &- E\{\Phi^i(Z_1)E[\Phi^j(Z_2)|\rho Z_1, 1-\rho^2]|0,1\} \\ &\times E\{\Phi^\ell(Z_1)E[\Phi^m(Z_2)|\rho Z_1, 1-\rho^2]|0,1\}\,. \end{aligned}$$

For $\rho = 0$, each conditional expectation is just a constant and

(5.4) $$i_{\theta_{ij}\theta_{\ell m}}(0) = (i+\ell+1)^{-1}(j+m+1)^{-1} - (i+1)^{-1}(j+1)^{-1}(\ell+1)^{-1}(m+1)^{-1}.$$

We consider a test of the null hypothesis that (X_1, X_2) has a bivariate normal distribution with $\lambda^* = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ unknown and with $\rho = 0$. We do not consider departures in ρ but rather smooth departures (5.1), with $k = 2$ for clarity. From (5.2) it may be seen that the smooth test statistic $\psi_2^2(\hat{\lambda}^*)$ will have five rather than two degrees of freedom. Let

(5.4) $$U'(\lambda^*) = [U_{\theta_{10}}(\lambda^*), U_{\theta_{20}}(\lambda^*), U_{\theta_{01}}(\lambda^*), U_{\theta_{02}}(\lambda^*), U_{\theta_{11}}(\lambda^*)] =$$

$$= \left[\Phi\left(\frac{X_1-\mu_1}{\sigma_1}\right) - \frac{1}{2},\ \Phi^2\left(\frac{X_1-\mu_1}{\sigma_1}\right) - \frac{1}{3},\ \Phi\left(\frac{X_2-\mu_2}{\sigma_2}\right) - \frac{1}{2},\ \Phi^2\left(\frac{X_2-\mu_2}{\sigma_2}\right) - \frac{1}{3},\right.$$

$$\left. \Phi\left(\frac{X_1-\mu_1}{\sigma_1}\right)\Phi\left(\frac{X_2-\mu_2}{\sigma_2}\right) - \frac{1}{4} \right] .$$

Then from (5. 4),

$$i_{\theta\theta} = \mathrm{Cov}(U_\theta(\lambda^*) \mid \theta = 0, \lambda^*)$$

$$= \begin{bmatrix} \frac{1}{12} & \frac{1}{12} & 0 & 0 & \frac{1}{24} \\ & \frac{4}{45} & 0 & 0 & \frac{1}{24} \\ & & \frac{1}{12} & \frac{1}{12} & \frac{1}{24} \\ & & & \frac{4}{45} & \frac{1}{24} \\ & & & & \frac{7}{144} \end{bmatrix} .$$

Now because of the independence of X_1 and X_2 under the null hypothesis, the null density factors:

$$\phi_2(x_1, x_2 \mid \lambda^*) = \phi(x_1 \mid \mu_1, \sigma_1)\phi(x_2 \mid \mu_2, \sigma_2) .$$

So the efficient scores for the components of λ^* and their covariances may be taken from Section 4A. Moreover, scores relating to X_1 are uncorrelated with those relating to X_2 when H_0 is true, so from (4. 1)

$$i_{\lambda^*\lambda^*}(\lambda^*) = \mathrm{diag}(\sigma_1^{-2}, 2\sigma_1^{-2}, \sigma_2^{-2}, 2\sigma_2^{-2}) .$$

Again, since X_1 and X_2 are independent by assumption, it follows that

$$i_{\mu_1\theta_{0m}} = i_{\sigma_1\theta_{0m}} = i_{\mu_2\theta_{m0}} = i_{\sigma_2\theta_{m0}} = 0$$

for $m = 1, 2$. And since expressions involving only X_1 or X_2 depend only on the marginal distributions, the covariances $i_{\mu_1\theta_{m0}}, i_{\sigma_1\theta_{m0}}, i_{\mu_2\theta_{0m}}, i_{\sigma_2\theta_{0m}}$ may also be taken directly from Section 4A. Finally, again by independence,

$$\begin{aligned} i_{\mu_1\theta_{11}}(\lambda^*) &= E\{U_{\mu_1}(\lambda^*)\Phi(\frac{X_1-\mu_1}{\sigma_1})\Phi(\frac{X_2-\mu_2}{\sigma_2})|\lambda^*\} \\ &= i_{\mu_1\theta_{10}}(\lambda^*)E\{\Phi(\frac{X_2-\mu_2}{\sigma_2})|\lambda^*\} \\ &= i_{\mu_1\theta_{10}}(\lambda^*)/2 \end{aligned}$$

with analogous expressions for the rest of $i_{\lambda^*\theta_{11}}(\lambda^*)$. Thus we have

$$i_{\theta\lambda*}(\lambda*) = \begin{bmatrix} \frac{1}{2\pi^{1/2}\sigma_1} & 0 & 0 & 0 \\ \frac{1}{2\pi^{1/2}\sigma_1} & \frac{\sqrt{3}}{6\pi\sigma_1} & 0 & 0 \\ 0 & 0 & \frac{1}{2\pi^{1/2}\sigma_2} & 0 \\ 0 & 0 & \frac{1}{2\pi^{1/2}\sigma_2} & \frac{\sqrt{3}}{6\pi\sigma_2} \\ \frac{1}{4\pi^{1/2}\sigma_1} & 0 & \frac{1}{4\pi^{1/2}\sigma_2} & 0 \end{bmatrix}$$

and

$$i_{\theta\theta|\lambda*} = \begin{bmatrix} \frac{\pi-3}{12\pi} & \frac{\pi-3}{12\pi} & 0 & 0 & \frac{\pi-3}{24\pi} \\ & \frac{32\pi^2-90\pi-15}{360\pi^2} & 0 & 0 & \frac{\pi-3}{24\pi} \\ & & \frac{\pi-3}{12\pi} & \frac{\pi-3}{12\pi} & \frac{\pi-3}{24\pi} \\ & & & \frac{32\pi^2-90\pi-15}{360\pi^2} & \frac{\pi-3}{24\pi} \\ & & & & \frac{7\pi-18}{144\pi} \end{bmatrix} .$$

Numerically,

$$(\pi-3)/12\pi = 0.003756$$

$$(\pi-3)/24\pi = 0.001878$$

$$(32\pi^2-90\pi-15)/360\pi^2 = 0.005090$$

$$(7\pi-18)/144\pi = 0.008822.$$

So

$$(5.5) \qquad i_{\theta\theta|\lambda*} = \begin{bmatrix} A & -B & 36 & 0 & -72 \\ & B & 0 & 0 & 0 \\ & & A & -B & -72 \\ & & & B & 0 \\ & & & & 144 \end{bmatrix}$$

where

$$A = (454\pi^3 - 1098\pi^2 - 720\pi + 105)/(2\pi^3 - 6\pi^2 - 15\pi + 45)$$

$$= 1051.965825$$

$$B = 360\pi^2/(2\pi^2 - 15) = 749.715349.$$

From (5.4) and (5.5) we obtain the quadratic score statistic

$$(5.6) \qquad \psi_2^2(\hat{\lambda}_0^*) = n^{-1} U'_{\cdot\theta}(\hat{\lambda}_0^*) i^{-1}_{\theta\theta|\lambda} U_{\cdot\theta}(\hat{\lambda}_0^*)$$

where

$$\hat{\lambda}_0^* = (\bar{x}_1, \bar{x}_2, s_1, s_2)$$

is the maximum likelihood estimator of λ^*.

It is extremely important to note that (5.6) is only appropriate for testing that X_1 and X_2 are independently distributed with a bivariate normal distribution relative to a fixed, known coordinate system. To paraphrase Gnanadesikan (1977; Section 5.4.2), the test presupposes a total commitment to the given coordinate system:

interest must be completely confined to the observed coordinates. For example, suppose a new pair of coordinate axes is obtained by a rigid rotation through angle η. Let $V = (v_1, v_2)'$ denote the new coordinates of the point with original coordinates $X = (x_1, x_2)$, i.e.,

$$V = A_\eta X = \begin{bmatrix} \cos\eta & \sin\eta \\ -\sin\eta & \cos\eta \end{bmatrix} X.$$

The use of $\psi_2^2(\hat{\lambda}_0^*)$ defined in (5.6) is based on the assumption that X_1 and X_2 are independent:

$$\mathrm{Cov}(X \mid \lambda^*) = D_X = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$$

where σ_1^2 and σ_2^2 are unknown. But

$$\mathrm{Cov}(V \mid \lambda^*) = A_\eta D_X A_\eta' = \begin{bmatrix} \sigma_1^2\cos^2\eta+\sigma_2^2\sin^2\eta & (\sigma_2^2-\sigma_1^2)\cos\eta\sin\eta \\ & \sigma_1^2\sin^2\eta+\sigma_2^2\cos^2\eta \end{bmatrix}.$$

Thus V_1 and V_2 are uncorrelated if and only if $\sigma_1^2 = \sigma_2^2$ or $\eta = N\pi/2$ for some integer N, and in general $\psi_2^2(\hat{\lambda}_0^*)$ is not appropriate for data V obtained from the original X. This lack of invariance, which is related to the difficulty of assuming ρ is unknown, is a major drawback to the use of the smooth test statistic for multivariate normality, and arises from the use of the marginal probability integral transformations. The problem of testing the composite hypothesis of normality remains invariant under location and scale changes, rotations, and reflections. However, the smooth test statistics based on the marginal Y's are only location and scale invariant. Further work in this area may be possible. For example, correlated data might be orthogonalized by regressing X_2 on X_1 and then testing X_1 and the residuals, or by transforming the data to its principal components.

III. POWER AND EFFICIENCY

III. 1. Introduction

In most practical applications, Neyman's smooth test will be utilized to test a composite hypothesis about a parameter vector of dimension greater than unity. In such a situation one does not expect to find a procedure with any strong optimality properties, such as the property of being uniformly most powerful. Nevertheless, by virtue of being asymptotically equivalent to the likelihood ratio, Neyman's test does have the optimality properties described by Neyman (1959) and Wald (1943); refer to Chapter II and the Appendix.

When the selection of a test procedure is not dictated by any compelling optimality considerations, one is led to comparison of power functions and to measures such as relative efficiency in evaluating the available procedures. In the case of the smooth goodness-of-fit test this will assist the statistician in answering two closely related questions. How well do the smooth tests compare with the other omnibus tests, on the one hand, and with the parametric tests, on the other? In addition, what value of k is best, and will one choice for k serve in general?

The information about a test's performance is contained in its power functions. In Section 2 the power of the smooth goodness-of-fit tests against departures in various "directions" from the null hypothesis is discussed. Sections 3, 4, and 5 contain examples of calculations described in Section 2. Of course, it is awkward to try to compare procedures in terms of entire power functions, so we seek summary measures which are algebraically simpler and conceptually more concise. In Section 6 it is seen that asymptotic relative efficiency provides such a measure which is suitable for comparing smooth tests with other parametric tests based on quadratic score statistics. Section 7 describes such a comparison of the smooth tests with the omnibus Pearson's χ^2 test and a generalization.

III. 2. The Power of Smooth Goodness-of-Fit Tests

Neyman (1937) derived directly the asymptotic noncentral χ^2 distribution of ψ_k^2 under local alternatives, when testing a simple goodness-of-fit hypothesis. In Chapter II this result, as well as the analogous result for a composite hypothesis, was seen to be a special case of general results for quadratic score statistics. Simply stated, the noncentrality parameter is a quadratic form in θ, the vector parameter indexing the smooth departures from the null density, given by (II. 1.1) and (II. 1.2) for simple and composite hypotheses, respectively. Neyman (1937) and Javitz (1975) examine the noncentrality parameter of ψ_k^2 when the smooth alternatives are of various orders other than k. Barton (1953) derived a formula giving the exact power of ψ_k^2 for any n when the alternatives are of the form

(II. 2. 2), however the expression is dismayingly complicated: a polynomial of order n in the θ's.

It is of interest to consider the power of the smooth tests against alternatives other than those for which the tests were designed. Hamdan (1962) calculated the noncentrality parameter when the smooth test of the simple hypothesis of standard normality is applied and the true distribution is $N(\mu, 1)$ for $\mu = 0.05(0.05)0.20$. Javitz (1975) considered the power of the smooth test against composite, "non-smooth" alternatives. He also obtained empirical powers for tests of the normal, beta, gamma, and Poisson distributions by computer simulation. He considered $\psi_k^2(\hat{\lambda}_0)$ for $k = 2, 3, 4$ as well as Pearson's χ^2 and in each case the alternatives were mixtures of densities of the null family with alterations in parameter values. In addition he estimated the power of the tests of normality when the distribution is in fact a standard Cauchy or a Gamma (4, 2). These latter results are reproduced in Table 4. The six entries in each column are based on one set of 100 samples of the indicated size and are therefore correlated. Critical levels were taken from the theoretical asymptotic χ_k^2 distribution for $\psi_k^2(\hat{\lambda}_0)$, and, as in its common usage, the Pearson χ^2 statistics were also taken to have χ^2 distributions. Kang (1977) obtained empirical powers of several tests of normality: $\psi_k^2(\hat{\lambda}_0)$ for $k = 1, \ldots, 4$, as well as the omnibus χ^2 and Kolmogorov-Smirnov tests and the tests for normality

Table 4. Empirical powers of tests for normality (Javitz, 1975).

Alternative:	Gamma (4, 2)			Cauchy	
n:	50	100	150	25	50
Test Statistic					
$\psi_2^2(\hat{\lambda}_0)$	33	65	86	91	100
$\psi_3^2(\hat{\lambda}_0)$	31	55	72	90	99
$\psi_4^2(\hat{\lambda}_0)$	32	53	78	90	99
χ^2 (20%/cell)	21	28	44	86	98
χ^2 (10%/cell)	25	29	47	79	96
χ^2 (5%/cell)	25	26	40	76	91

Note: entries are percentages rounded to the nearest digit; tests are at $\alpha = 0.10$; all powers against Cauchy are 100% for $n \geq 75$.

based on b_1 and b_2, and the Shapiro-Wilk statistic. He considered several alternatives from among those used by Shapiro, Wilk and Chen (1968), including the Weibull with scale parameter 2 (denoted W2), standard Laplace (Lap.), standard logistic (Log.), extreme value with parameters 0 and 0.5 (EV(.5)), and double χ^2 with parameter -0.5 (D(-0.5)). This last density is proportional to $\exp\{-x^4\}$ and therefore has light tails relative to the normal density. He first estimated the critical value of the statistics from 2000 trials for each statistic at each sample size. The powers relative to these empirical critical values were then estimated from 2000 trials for W2 and Lap. and 1000 trials for the rest, at each sample size. The results are shown in Table 5. The smooth tests with $k = 1$ and 2 are seen to perform very well, having maximum power in half of the cases. Notice that the smooth tests dominate the Pearson's χ^2 and Kolmogorov-Smirnov tests. The power of $\psi_2^2(\hat{\lambda}_0)$ is never more than 8% below the maximum power, and the smooth tests with one or two degrees of freedom perform as well as or better than procedures specifically designed for testing normality, at least for the alternatives considered.

In the remainder of this section we consider the power of the smooth tests against parametric departures from the null density. Suppose we have a family of densities $f(x|\beta)$, indexed by a scalar parameter β, which contains the null density $f(x) = f(x|\beta = 0)$.

Table 5. Empirical powers of tests for normality (Kang, 1977).

Alternative:	W2		Lap.		Log.		EV(.5)		D(-.5)	
n:	20	50	20	50	20	50	20	50	20	50
Test Statistic										
$\psi_1^2(\hat{\lambda}_0)$	26*	55	29	31	18	20	45*	83*	6	6
$\psi_2^2(\hat{\lambda}_0)$	22	52	39*	69*	21	32	40	76	14	38
$\psi_3^2(\hat{\lambda}_0)$	18	42	38	66	20	32	36	69	13	27
$\psi_4^2(\hat{\lambda}_0)$	20	40	39	66	21	32	38	69	13	20
χ_2^2	12	--	25	--	13	--	23	--	14	--
χ_7^2	--	23	--	38	--	14	--	35	--	17
D	18	34	31	54	15	20	31	55	15	22
b_1	24	52	34	43	21*	27	43	81	2	2
b_2	15	21	35	61	18	33*	26	42	20*	46*
W	25	58*	35	50	19	22	41	78	15	40

Note: entries are percentages rounded to nearest digit; tests are at $\alpha = 0.10$; D = Kolmogorov-Smirnov, W = Shapiro-Wilk; maximum power in each column denoted by asterisk.

Here β is not a nuisance parameter, and should one be required, we consider densities $f(x|\beta,\lambda)$ with the null densities given by $f(x|\lambda) = f(x|\beta = 0,\lambda)$. First, however, we consider the case of a simple hypothesis. Let U_β denote the efficient score for β:

$$U_\beta = U_\beta(0) = [\partial \ln f(x|\beta)/\partial\beta]_{\beta=0} .$$

The locally best test for $H_0 : \beta = 0$ is based on the asymptotic normal distribution of $n^{1/2}U._\beta$; see, for example, Cox and Hinkley (1974; Section 4.8).

We assume that ψ_k^2 for some k is used to test the goodness-of-fit of the density $f(x) = f(x|0)$ when in fact the departure is in the "β direction". Consider the family of densities which includes both smooth alternatives as well as β-departures from $f(x|0)$:

$$g(x|\theta,\beta) = f(x|\beta)\exp\left\{\sum_{i=1}^{k} \theta_i F^i(x|0) - K(\theta,\beta)\right\} .$$

Here $K(\theta,\beta)$ is simply the normalizing constant, and clearly $K(0,\beta) = 0$ identically in β. The use of the smooth test of $H_0 : \theta = 0$ is of course based on the assumption that $\beta = 0$. Proceeding as before we find that

$$U_\theta(\theta=0,\beta=0) = \begin{bmatrix} F(x|0)-1/2 \\ \vdots \\ F^k(x|0)-(k+1)^{-1} \end{bmatrix} .$$

Also

$$U_\beta(\theta=0, \beta=0) = U_\beta ,$$

that is, the β-score for the larger family is just the same as the β-score for the parametric family of densities $f(x|\beta)$. So we may denote the scores by U_θ and U_β without ambiguity. The covariance matrix of the vector of scores $[U_\theta', U_\beta]'$ may be partitioned in an obvious fashion as

$$\begin{bmatrix} i_{\theta\theta} & i_{\theta\beta} \\ & i_{\beta\beta} \end{bmatrix}. \tag{2.1}$$

Here $i_{\theta\theta}$ is the same as before, given by (II. 2. 3), and $i_{\beta\beta}$ is just the information about β contained in x, relative to the family $f(x|\beta)$. If this quantity is finite, since $i_{\theta_j\theta_j}$ is finite for all j, it follows that $i_{\theta_j\beta} = \mathrm{Cov}[F^i(x|0), U_\beta|\theta=0, \beta=0]$ is also finite for all j.

Under regularity conditions on the family of densities $f(x|\beta)$--see the Appendix--the approximate large-sample distribution of the efficient score vector is given by

$$n^{-1/2}\begin{bmatrix} U_{\cdot\theta} \\ U_{\cdot\beta} \end{bmatrix} \mathrel{\dot\sim} N_{k+1}\left(n^{1/2}\begin{bmatrix} i_{\theta\beta} & i_{\theta\beta} \\ & i_{\beta\beta} \end{bmatrix}\begin{bmatrix} \theta^* \\ \beta^* \end{bmatrix}, \begin{bmatrix} i_{\theta\theta} & i_{\theta\beta} \\ & i_{\beta\beta} \end{bmatrix}\right)$$

for (θ^*, β^*) near $(0, 0)$. So the marginal distribution of the

θ-scores, when there is a "pure" β departure, is approximately given by

$$n^{-1/2}U_{\cdot\theta} \stackrel{\cdot}{\sim} N_k(n^{1/2} i_{\theta\beta}\beta^*; i_{\theta\theta}) .$$

Therefore the noncentrality parameter for the test based on ψ_k^2, under a local alternative of the $f(x|\beta)$ type, is

$$\text{(2. 2)} \qquad \delta(\beta^*) = n(i_{\theta\beta}\beta^*)' i_{\theta\theta}^{-1}(i_{\theta\beta}\beta^*) = n(\beta^*)^2 i_{\beta\theta} i_{\theta\theta}^{-1} i_{\theta\beta}$$

$$= n(\beta^*)^2 i_{\beta\beta} R^2_{\theta,\beta}$$

where $R^2_{\theta,\beta} = i_{\beta\theta} i_{\theta\theta}^{-1} i_{\theta\beta} / i_{\beta\beta}$ is well known to be the squared multiple correlation coefficient of the scores $U_{\theta_1}, \ldots, U_{\theta_k}$ with U_β, under the null hypothesis; see, for example, Anderson (1958, p. 32).

Equation (2. 2) is intuitively quite appealing. If we suppose that n has a fixed, large value, since $i_{\beta\beta}$ is constant, the rate at which the power increases with $(\beta^*)^2$ depends on how closely the linear span of $\{U_{\theta_1}, \ldots, U_{\theta_k}\}$ comes to containing U_β. If the θ-scores are all orthogonal to (uncorrelated with) U_β, then $R^2_{\theta,\beta} = 0$ and the smooth test has, locally, no power against departures in the β direction. At the other extreme, if U_β is in the span of $\{U_{\theta_1}, \ldots, U_{\theta_k}\}$ then $R^2_{\theta,\beta} = 1$. But $n\beta^2 i_{\beta\beta}$ is the noncentrality parameter for the locally best test of $H_0: \beta = 0$ for the family of densities $f(x|\beta)$, under local departures. Of course if $k = 1$, then $R^2_{\theta,\beta} = 1$ if and only if $U_\beta \propto U_\theta$, in which case ψ_1^2 is exactly the quadratic score statistic for testing $H_0 : \beta = 0$, so the two tests must have identical power. But if $k > 1$, then the alternative distributions are

$$\psi_k^2 \mathrel{\dot{\sim}} \chi^2_{k,\, n\beta^2 i_{\beta\beta}}$$

and

$$W_u = U^2_{\cdot\beta}/i_{\cdot\beta\beta} \mathrel{\dot{\sim}} \chi^2_{1,\, n\beta^2 i_{\beta\beta}}$$

which implies that W_u is locally more powerful than ψ_k^2, since the noncentral χ^2 distributions are stochastically increasing in their degrees of freedom for any fixed value of their noncentrality parameter. Heuristically, ψ_k^2 is indeed giving protection in the right direction, but is losing power in other directions when $k > 1$. Correction for unequal degrees of freedom will be discussed at more length in Section 6.

Before calculating some example values of $R^2_{\theta,\beta}$, we consider the more important case of composite hypotheses. The development follows closely that given above and therefore will only be outlined. We begin with a family of densities $f(x|\beta,\lambda)$ where $\lambda \in \mathbb{R}^s$ is a nuisance parameter. The β-score is then

$$U_\beta(0,\lambda) = [\partial \ln f(x|\beta,\lambda)/\partial\beta]_{\beta=0,\lambda} \cdot$$

Analogously, the λ-score is given by

$$U_{\lambda_j}(0,\lambda) = [\partial \ln f(x|\beta,\lambda)/\partial\lambda_j]_{\beta=0,\lambda}$$

and

$$U_\lambda(0,\lambda) = [U_{\lambda_1}(0,\lambda), \ldots, U_{\lambda_s}(0,\lambda)]'.$$

Consider the density functions

$$g(x|\theta,\beta,\lambda) = f(x|\beta,\lambda)\exp\left\{\sum_{i=1}^{k} \theta_i F^i(x|0,\lambda) - K(\theta,\beta,\lambda)\right\}.$$

Clearly $K(0,\beta,\lambda) = 0$ identically in β and λ, and $K(\theta,0,\lambda)$ is constant in λ. As before

$$U_\theta(\theta=0,\beta=0,\lambda) = \begin{bmatrix} F(x|0,\lambda)-1/2 \\ \vdots \\ F^k(x|0,\lambda)-(k+1)^{-1} \end{bmatrix}$$

and

$$U_\beta(\theta=0,\beta=0,\lambda) = U_\beta(0,\lambda) .$$

That is, the β-score for the larger family of densities is just the β-score for the parametric family $f(x|\beta,\lambda)$. Similarly,

$$U_\lambda(\theta=0,\beta=0,\lambda) = U_\lambda(0,\lambda).$$

The overall information matrix may be written in partitioned form as

$$(2.3) \qquad \begin{bmatrix} i_{\theta\theta} & i_{\theta\beta} & i_{\theta\lambda} \\ & i_{\beta\beta} & i_{\beta\lambda} \\ & & i_{\lambda\lambda} \end{bmatrix}$$

where $i_{\beta\beta}$ and $i_{\lambda\lambda}$ are the information matrices for the smaller parametric families of densities $f(x|\beta, \lambda)$ and $f(x|0, \lambda)$, respectively. The possible dependence of some entries of (2.3) on λ has been suppressed in the notation. As usual $i_{\theta\theta} = i_{\theta\theta}(\theta=0, \beta=0, \lambda)$ is constant in β and λ, and is given by (II. 2.3). The remaining parts of (2.3) may depend on λ. However, in the examples to be considered below, the more important adjusted covariance matrices are free of λ. In the case that $i_{\theta\beta}$ and $i_{\beta\beta}$ are independent of λ, the upper left portion of (2.3) is just the same as (2.1).

Now we consider the adjusted θ- and β-scores

$$\begin{aligned} U_{\theta|\lambda}(\lambda) &= U_{\theta|\lambda}(0, 0, \lambda) \\ &= U_\theta(0, 0, \lambda) - i_{\theta\lambda} i_{\lambda\lambda}^{-1} U_\lambda(\lambda) , \end{aligned}$$

and

$$U_{\beta|\lambda}(\lambda) = U_\beta(0, \lambda) - i_{\beta\lambda} i_{\lambda\lambda}^{-1} U_\lambda(\lambda) ,$$

for which the covariance matrix may be written

$$\begin{bmatrix} i_{\theta\theta|\lambda} & i_{\theta\beta|\lambda} \\ & i_{\beta\beta|\lambda} \end{bmatrix}$$

where

$$i_{\theta\theta|\lambda} = i_{\theta\theta} - i_{\theta\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\theta}$$

as before, and

$$i_{\theta\beta|\lambda} = i_{\theta\beta} - i_{\theta\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\beta}$$

$$i_{\beta\beta|\lambda} = i_{\beta\beta} - i_{\beta\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\beta} .$$

Under regularity conditions discussed in the Appendix it follows that for large n

$$n^{-1/2} \begin{bmatrix} U_{\cdot\theta|\lambda}(\lambda) \\ U_{\cdot\beta|\lambda}(\lambda) \end{bmatrix} \doteq N_{k+1} \left(n^{1/2} \begin{bmatrix} i_{\theta\theta|\lambda} & i_{\theta\beta|\lambda} \\ & i_{\beta\beta|\lambda} \end{bmatrix} \begin{bmatrix} \theta^* \\ \beta^* \end{bmatrix}, \begin{bmatrix} i_{\theta\theta|\lambda} & i_{\theta\beta|\lambda} \\ & i_{\beta\beta|\lambda} \end{bmatrix} \right)$$

when (θ^*, β^*) is near $(0, 0)$. Thus the marginal distribution of the adjusted θ-scores under a β departure is approximately given by

$$n^{-1/2} U_{\cdot\theta|\lambda}(\lambda) \doteq N_k(n^{1/2} i_{\theta\beta|\lambda} \beta^*, i_{\theta\theta|\lambda}) ,$$

and the noncentrality parameter for the test based on $\psi_k^2(\hat{\lambda}_0)$ is, for small β^*, large n, and an appropriate $\hat{\lambda}_0$,

$$(2.4) \qquad \delta(\beta^*|\lambda) = n(i_{\theta\beta|\lambda}\beta^*)' i^{-1}_{\theta\theta|\lambda}(i_{\theta\beta|\lambda}\beta^*)$$

$$= n(\beta^*)^2 i_{\beta\theta|\lambda} i^{-1}_{\theta\theta|\lambda} i_{\theta\beta|\lambda}$$

$$= n(\beta^*)^2 i_{\beta\beta|\lambda} R^2_{\theta,\beta|\lambda}$$

where

$$R^2_{\theta,\beta|\lambda} = i_{\beta\theta|\lambda} i^{-1}_{\theta\theta|\lambda} i_{\theta\beta|\lambda} / i_{\beta\beta|\lambda}$$

is, by analogy with $R^2_{\theta,\beta}$, the natural way to define the squared partial multiple correlation coefficient of the scores $U_{\theta_1}(\lambda), \ldots, U_{\theta_k}(\lambda)$ with $U_\beta(\lambda)$, adjusting for $U_\lambda(\lambda)$, when X has density $f(x|0,\lambda)$. In other words, $R^2_{\theta,\beta|\lambda}$ is the multiple correlation of $U_{\theta_1|\lambda}(\lambda), \ldots, U_{\theta_k|\lambda}(\lambda)$ with $U_{\beta|\lambda}(\lambda)$, the parts of the θ- and β-scores which are orthogonal to, or "free of," $U_\lambda(\lambda)$.

Notice the similarity of (2.4) to (2.2), and recall that $n(\beta^*)^2 i_{\beta\beta|\lambda}$ is the local noncentrality parameter for $W_u(\hat{\lambda}_0)$, the generalization of the locally best test of $H_0: \beta = 0$ for the family of densities $f(x|\beta)$ to the test of $H_0: \beta = 0$, λ unknown, for the family of densities $f(x|\beta, \lambda)$. As before, if $R^2_{\theta,\beta|\lambda} = 1$ and $k = 1$, then this test is equivalent to the smooth test based on $\psi^2_1(\hat{\lambda}_0)$. However, for $k > 1$, $\psi^2_k(\hat{\lambda}_0)$ will be less powerful than $W_u(\hat{\lambda}_0)$. This suggests also that if $R^2_{\theta,\beta|\lambda}$ increases slowly with k, then the power of the smooth test will decrease as k increases.

Before proceeding to the calculation of some exemplary multiple correlation coefficients, we comment on the restriction to scalar β. The densities $g(x|\theta, \beta)$ and $g(x|\theta, \beta, \lambda)$ were introduced in order to permit the calculation of powers and efficiencies of tests against certain departures from the null density. They are not intended to be models for the behavior of observable phenomena, so the additional richness of the class of densities provided by a vector β is of no value in that regard. And in thinking of power properties of a test, it is perhaps most useful to think in terms of one "direction" at a time. For the local departures being considered here, a departure along a straight line from the origin in the parameter space suffices for this, and this can be modeled as an alternative in a scalar parameter, after a reparameterization, if necessary.

III.3. Multiple Correlations for Tests of the Exponential Distribution: Weibull Departures

We consider applying a smooth test for exponential goodness of fit when the alternative is in fact a Weibull distribution. The approach of Section 2 applies directly in this case. The Weibull densities may be written

$$f(x|\beta, \lambda) = \lambda^{-1}(x/\lambda)^{\beta-1}\exp\{-(x/\lambda)^{\beta}\}$$

for $x, \beta, \lambda > 0$. Suppose first that the scale parameter is known:

we take it to be $\lambda = 1$ without loss of generality. Thus we are testing the null hypothesis that the density is $f(x) = \exp(-x)$. Now the necessary matrices are obtained by easy modifications of the results of Section II. 4. C with $\lambda = \sigma$, $\beta = \tau$. Thus

$$i_{\beta\beta} = 1.828681$$

$$i_{\theta_j\beta} = j[1-\gamma-\ln(j+1)]/(j+1)^2$$

where $\gamma = 0.577216...$ is Euler's constant. Numerically then, for k up to 4,

$$i_{\theta\beta} = \begin{bmatrix} -0.067591 \\ -0.150184 \\ -0.180658 \\ -0.189865 \end{bmatrix}.$$

Taking $i_{\theta\theta}^{-1}$ from Section II. 2, the squared multiple correlation coefficients $R^2_{\theta_1 \ldots \theta_k, \beta} = i_{\beta\theta} i_{\theta\theta}^{-1} i_{\theta\beta} / i_{\beta\beta}$ are calculated and listed in Table 6a.

For testing the composite hypothesis that X has an exponential distribution with unspecified scale parameter λ, we have

$$i_{\lambda\lambda}(\lambda) = \lambda^{-2}$$

$$i_{\beta\lambda}(\lambda) = -(1-\gamma)/\lambda = -0.422784\lambda^{-1}$$

$$i_{\lambda\theta}(\lambda) = [-1/4\lambda, -2/9\lambda, -3/16\lambda, -4/25\lambda].$$

So

$$i_{\beta\beta|\lambda} = 1.644934$$

$$i_{\theta\beta|\lambda} = \begin{bmatrix} -0.173287 \\ -0.244136 \\ -0.259930 \\ -0.257510 \end{bmatrix}.$$

Table 6. Multiple correlations for Weibull departures from the exponential distribution.

k	$R^2_{\theta_1 \ldots \theta_k, \beta}$
a. Simple hypothesis	
1	0.030
2	0.703
3	0.729
4	0.860
	$R^2_{\theta_1 \ldots \theta_k, \beta\|\lambda}$
b. Composite hypothesis (λ = scale parameter)	
1	0.876
2	0.918
3	0.965
4	0.972

For k up to 4, $i^{-1}_{\theta\theta|\lambda}$ may be found in Table 2, so the squared partial multiple correlation coefficients $R^2_{\theta_1 \ldots \theta_k, \beta|\lambda} = i_{\beta\theta|\lambda} i^{-1}_{\theta\theta|\lambda} i_{\theta\beta|\lambda} / i_{\beta\beta|\lambda}$ can be calculated, and are listed in Table 6b. Notice that in each case, $R^2_{\theta,\beta} < R^2_{\theta,\beta|\lambda}$. For

$k = 1$ the difference is quite extreme, the squared correlations being 0.030 and 0.876, respectively. In Section 6 it will be seen that this implies that ψ_1^2 is much less efficient for testing the simple hypothesis of exponential goodness of fit than $\psi_1^2(\hat{\lambda}_0)$ is for testing the composite hypothesis. The difference in correlations is vividly illustrated in Figures 9 and 10. The first shows the standardized score functions

$$U_{\theta_1} i_{\theta_1\theta_1}^{-1/2} = (y - \frac{1}{2})/0.288675$$

$$U_\beta i_{\beta\beta}^{-1/2} = [1+(1+\ln y)\ln(-\ln y)]/1.352293$$

and the second shows

$$-U_{\theta_1|\lambda} i_{\theta_1\theta_1|\lambda}^{-1/2} = [-y + \frac{1}{4}\ln y + \frac{3}{4}]/0.144338$$

$$U_{\beta|\lambda} i_{\beta\beta|\lambda}^{-1/2} = [2-\gamma+(1+\ln y)\ln(-\ln y)+(1-\gamma)\ln y]/1.282550.$$

Notice that the score functions are given in terms of $y = \overline{F}(x|\lambda) = \exp(-x/\lambda)$. The standardization and the use of $-U_{\theta_1|\lambda}$ have no effect on the correlations, but make the curves more readily comparable.

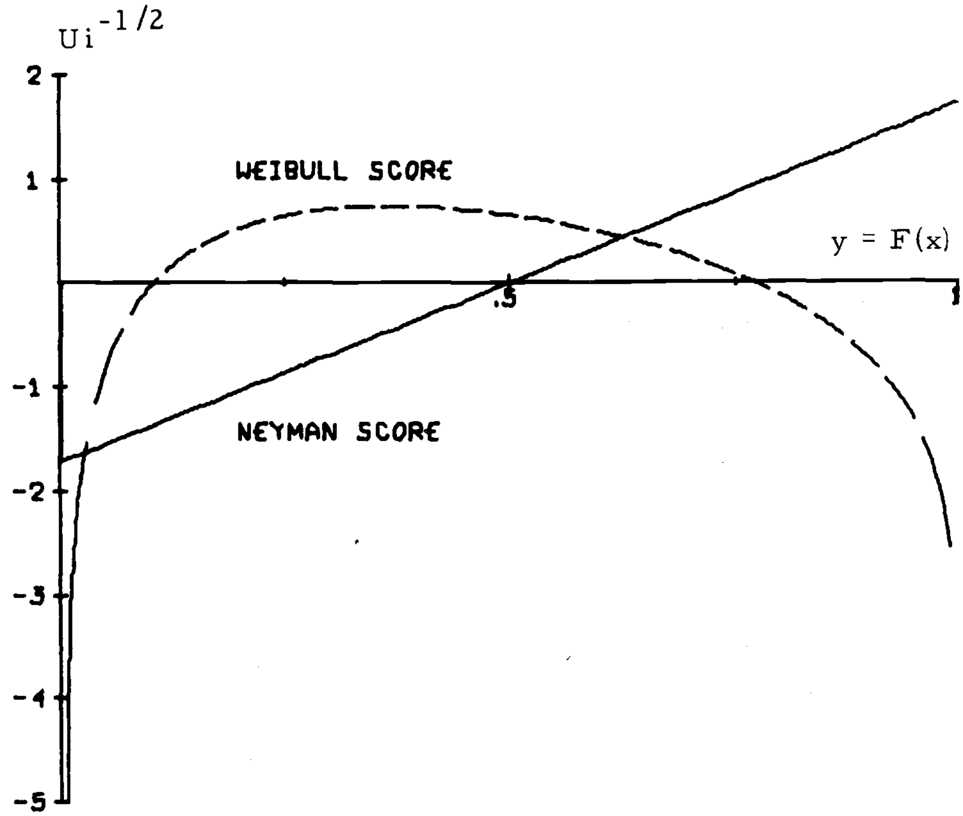

Figure 9. Neyman and Weibull scores for tests of simple exponentiality.

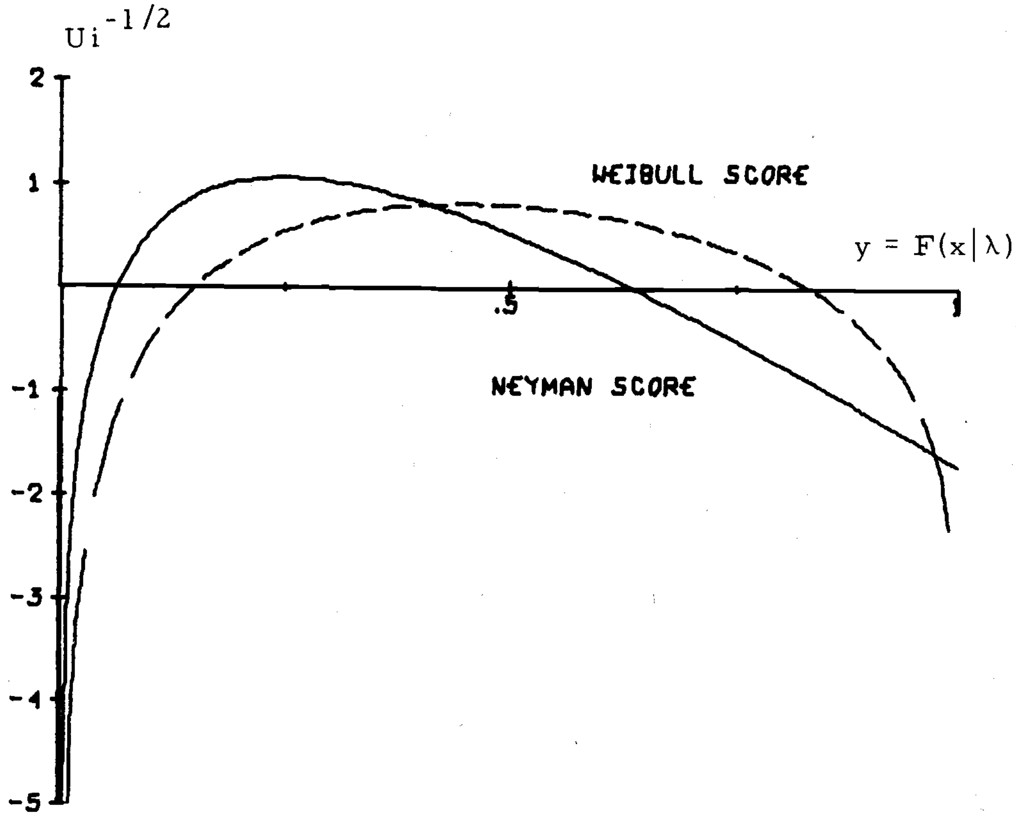

Figure 10. Neyman and Weibull scores for tests of composite exponentiality.

III. 4. Multiple Correlations for Tests of the Exponential Distribution: Gamma Departures

We proceed as in the previous section. The Gamma densities are given by

$$f(x|\beta, \lambda) = \lambda^{-1}[\Gamma(\beta)]^{-1}(x/\lambda)^{\beta-1}\exp(-x/\lambda)$$

for $x, \beta, \lambda > 0$. Assume first we are testing a simple hypothesis of exponential goodness of fit: without loss of generality we take $\lambda = 1$. Now

$$U_\beta(\beta=1) = -\psi(1) + \ln x = \gamma + \ln x$$

where $\psi(\beta) = \partial \ln \Gamma(\beta)/\partial\beta$ and $\psi(1) = -\gamma = -0.577216...$ is the negative of Euler's constant. Since

$$\partial^2 \ln f(x|\beta, 1)/\partial\beta^2 = -\psi'(\beta)$$

it follows that

$$i_{\beta\beta} = -E\{[\partial^2 \ln f(x|\beta, 1)/\partial\beta^2]_{\beta=1}\} = \psi'(1) = 1.644934.$$

Also since $U_{\theta_j} = \exp(-jx) - (j+1)^{-1}$

$$i_{\theta_j\beta} = \int_0^\infty (\ln x)\exp\{-(j+1)x\}dx + \gamma/(j+1)$$

$$= -[\ln(j+1)]/(j+1)$$

or, numerically, for k up to 4,

$$i_{\theta\beta} = \begin{bmatrix} -0.346574 \\ -0.366204 \\ -0.346574 \\ -0.321888 \end{bmatrix} .$$

With $i_{\theta\theta}^{-1}$ from Section II. 2, the squared multiple correlation coefficients $R^2_{\theta_1 \ldots \theta_k, \beta}$ are calculated and listed in Table 7a.

Table 7. Multiple correlations for Gamma departures from the exponential distribution.

k	$R^2_{\theta_1 \ldots \theta_k, \beta}$
a. Simple hypothesis	
1	0.876
2	0.918
3	0.965
4	0.972
	$R^2_{\theta_1 \ldots \theta_k, \beta \mid \lambda}$
b. Composite hypothesis (λ = scale parameter)	
1	0.694
2	0.840
3	0.910
4	0.938

For the test of the composite hypothesis that X has an exponential distribution with unknown scale parameter, it is easily seen that $i_{\beta\beta}$ and $i_{\theta\beta}$ are the same as above. Moreover, since

$$U_\lambda(\lambda) = (x-\lambda)/\lambda^2$$

is distributed under H_0 the same as for Weibull departures in Section 3, it is easy to see that $i_{\lambda\lambda}$ and $i_{\theta\lambda}$ are the same as in Section 3. Also $i_{\beta\lambda} = \lambda^{-1}$. So we may calculate the adjusted covariance matrices

$$i_{\beta\beta|\lambda} = 0.644934$$

and

$$i_{\theta\beta|\lambda} = \begin{bmatrix} -0.096574 \\ -0.143982 \\ -0.159070 \\ -0.161888 \end{bmatrix} .$$

Finally, $i^{-1}_{\theta\theta|\lambda}$ is given in Table 2 for k up to 4, so the squared partial multiple correlation coefficients may be calculated. These are listed in Table 7b. Two features of this table are notable. First, the multiple correlations for a Gamma departure from the simple hypothesis are the same as the partial multiple correlations for the Weibull departure from the composite hypothesis. Second, unlike the previous example, $R^2_{\theta_1 \ldots \theta_k, \beta | \lambda} < R^2_{\theta_1 \ldots \theta_k, \beta}$ for each k.

III. 5. Multiple Correlations for Tests of Normality: Skewness and Kurtosis Departures

In this example we depart somewhat from the situations of the previous sections, however the essential results are not altered. We seek to obtain the noncentrality parameters of the smooth tests for normality when the alternatives differ in skewness or kurtosis. Let $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ be the standard normal density, and consider the functions

$$h(x | \beta_1, \beta_2) = \phi(x) \{ 1 + \beta_1 H_3(x)/6 + (\beta_2 - 3) H_4(x)/24 \} \tag{5. 1}$$

where

$$H_3(x) = x^3 - 3x$$
$$H_4(x) = x^4 - 6x^2 + 3$$

are the third and fourth order Hermite polynomials, respectively. The function h is very nearly a probability density: it integrates to unity for all values of β_1 and β_2, however if $(\beta_1, \beta_2) \neq (0, 3)$, then $h(x | \beta_1, \beta_2) < 0$ for some values of x. But for (β_1, β_2) near $(0, 3)$, the region in which h is negative has very low probability under the null density $\phi(x)$, so we will deal with h as if it were a probability density since we are concerned with local departures.

The alternatives (5. 1) are in the spirit of (II. 2. 2) proposed by Barton (1953) and are simply truncated Gram-Charlier expansions of Type A; see, for example, Kendall and Stuart (1969; Sections 6. 14-6. 16). Taking h as a density, the moments of X are

$$E(X|\beta_1,\beta_2) = 0$$

$$E(X^2|\beta_1,\beta_2) = \mathrm{Var}(X|\beta_1,\beta_2) = 1$$

$$E(X^3|\beta_1,\beta_2) = \mathrm{Skew}(X|\beta_1,\beta_2) = \beta_1$$

$$E(X^4|\beta_1,\beta_2) = \mathrm{Kurt}(X|\beta_1,\beta_2) = \beta_2.$$

There are, of course, alterations in the higher moments as well but these serve to index the departures of interest. The score functions for β_1 and β_2 are easily seen to be

$$U_{\beta_1} = H_3(x)/6 = (x^3-3x)/6$$

$$U_{\beta_2} = H_4(x)/24 = (x^4-6x^2+3)/24.$$

From well-known properties of Hermite polynomials it follows that

(5. 2)
$$i_{\beta_1\beta_1} = 1/6$$

$$i_{\beta_2\beta_2} = 1/24.$$

To compute $i_{\theta\beta_1}$ and $i_{\theta\beta_2}$ we employ Table X in Barton (1953) and after some calculation find that for k up to 4

$$(5.3) \qquad {}^{i}\theta\beta_1 = \begin{bmatrix} -1/24\pi^{1/2} \\ -1/24\pi^{1/2} \\ \{-3(S_3+18\pi)+10\sqrt{2}\}/480\pi^{3/2} \\ \{-3(S_3+8\pi)+10\sqrt{2}\}/240\pi^{3/2} \end{bmatrix}$$

and

$$(5.4) \qquad {}^{i}\theta\beta_2 = \begin{bmatrix} 0 \\ -5\sqrt{3}/432\pi \\ -5\sqrt{3}/288\pi \\ \{-\sqrt{3}(25S_4+200\pi)+21\sqrt{5}\}/10080\pi^2 \end{bmatrix}$$

where

$$S_3 = 10 \arcsin(1/3) - \pi = 0.256776441$$

$$S_4 = 14 \arcsin(1/4) - \pi = 0.395930917.$$

Numerically,

$$(5.5) \qquad {}^{i}\theta\beta_1 = \begin{bmatrix} -.023507899 \\ -.023507899 \\ -.016154184 \\ -.008800469 \end{bmatrix}$$

and

$$(5.6) \qquad i_{\theta\beta_2} = \begin{bmatrix} 0.0 \\ -.006381121 \\ -.009571682 \\ -.010639393 \end{bmatrix}.$$

With these and with $i_{\theta\theta}^{-1}$ from Section II.2 we can calculate $R^2_{\theta\beta_j} = i_{\beta_j\theta} i^{-1}_{\theta\theta} i_{\theta\beta_j} / i_{\beta_j\beta_j}$ for $j = 1, 2$. These multiple correlation coefficients are displayed in Table 8a. Since $R^2_{\theta_1, \beta_2} = 0$, we see that ψ_1^2 has no power against departures from normality which affect the kurtosis but not the symmetry of the distribution. This is not surprising, since $U_{\theta_1} = \Phi(x) - \frac{1}{2}$ is an odd function of x, while the kurtosis score $U_{\beta_2} = H_4(x)/24 = (x^4 - 6x^2 + 3)/24$ is an even function. Notice also that the multiple correlation does not increase if $U_{\theta_3} = \Phi^3(x) - \frac{1}{4}$ is added to U_{θ_1} and U_{θ_2}. Unfortunately this does not bear a simple interpretation: if U_{θ_3} and U_{β_2} are regressed on U_{θ_1} and U_{θ_2}, the two residual functions are

$$\Phi^3(x) - \frac{5}{2}\Phi^2(x) + \frac{3}{5}\Phi(x) + \frac{17}{60}$$

and

$$\frac{1}{24} H_4(x) + \frac{25\sqrt{3}}{36\pi}\left(3\Phi^2(x) - 3\Phi(x) + \frac{1}{2}\right),$$

which can easily be shown to be uncorrelated, but these functions have no apparent significant interpretation. Similar behavior is seen in values of R^2_{θ,β_1}.

The results given above for the standard normal distribution apply to any simple hypothesis of normality, for if μ and σ^2 are known, the observations can be standardized. Now suppose we must test a composite hypothesis of normality with $\lambda = (\mu, \sigma)'$ unknown. Analogous to (5. 1) we consider densities

$$h(x|\beta_1, \beta_2, \lambda) = \sigma^{-1}\phi(z)\{1+\beta_1 H_3(z)/6+(\beta_2-3)H_4(z)/24\} \tag{5.7}$$

where $z = (x-\mu)/\sigma$. Now $i_{\beta_1\beta_1}(\lambda)$, $i_{\beta_2\beta_2}(\lambda)$, $i_{\theta\beta_1}(\lambda)$, and $i_{\theta\beta_2}(\lambda)$ are all independent of λ, and are just the same as given in (5. 2)-(5. 6). Also, $i_{\lambda\lambda}(\lambda)$ and $i_{\theta\lambda}(\lambda)$ are given by (II. 4. 1) and (II. 4. 2), respectively. Finally, since Hermite polynomials of different orders are uncorrelated, we find that $i_{\beta_1\lambda}(\lambda) = i_{\beta_2\lambda}(\lambda) = [0, 0]$. From this last fact, for $j = 1, 2$

$$i_{\theta\beta_j|\lambda} = i_{\theta\beta_j} - i_{\theta\lambda} i^{-1}_{\lambda\lambda}(\lambda) i_{\lambda\beta_j} = i_{\theta\beta_j}$$

$$i_{\beta_j\beta_j|\lambda} = i_{\beta_j\beta_j} - i_{\beta_j\lambda} i^{-1}_{\lambda\lambda}(\lambda) i_{\lambda\beta_j} = i_{\beta_j\beta_j} .$$

Taking $i^{-1}_{\theta\theta|\lambda}$ from Table 1 we can compute

$$R^2_{\theta_1 \ldots \theta_k, \beta_j|\lambda} = i_{\beta_j\theta} i^{-1}_{\theta\theta|\lambda} i_{\theta\beta_j} / i_{\beta_j\beta_j}$$

for $k = 1, \ldots, 4$ and $j = 1, 2$. These values are shown in Table 8b. Notice that $R^2_{\theta_1 \beta_2 | \lambda} = 0$ and that while the correlation coefficients are much higher than for the simple hypothesis, they still exhibit the behavior of increasing only with even or odd k.

Table 8. Multiple correlations for skewness and kurtosis departures from the normal distribution.

k	$R^2_{\theta_1 \ldots \theta_k, \beta_1}$	$R^2_{\theta_1 \ldots \theta_k, \beta_2}$
a. Simple hypothesis		
1	0.040	0.000
2	0.040	0.176
3	0.460	0.176
4	0.460	0.296
	$R^2_{\theta_1 \ldots \theta_k, \beta_1 \mid \lambda}$	$R^2_{\theta_1 \ldots \theta_k, \beta_2 \mid \lambda}$
b. Composite hypothesis ($\lambda = (\mu, \sigma)'$)		
1	0.883	0.000
2	0.883	0.733
3	0.996	0.733
4	0.996	0.863

When the noncentrality parameter is the same, the test with more degrees of freedom will have less power than the test with fewer degrees of freedom. So, for example, in testing for normality, ψ_1^2 has better power than ψ_2^2 against local departures in β_1. If the noncentrality parameter increases with k, it becomes a question of whether that increase is sufficient to overcome the loss in power incurred by the increase of k. The problem of adjusting for unequal degrees of freedom will be explored more carefully in the following section.

III. 5. Asymptotic Relative Efficiency of the Smooth Tests

Pitman's measure of asymptotic relative efficiency is a very useful tool for comparing statistical tests of hypotheses. Briefly, the efficiency is the ratio of the sample sizes required by the two procedures in question in order to insure that they have equal power at each point in a sequence of alternatives converging to the hypothesized law. This is typically done by showing that the power functions of the two tests are asymptotically of the same form, say $G(\cdot)$, but with different arguments, say $G(u_1)$ and $G(u_2)$, where the u's depend on sample sizes, alternative parameter values, etc. The limiting ratio of the sample sizes which insures the equality of the u's in the limit is then taken to be the asymptotic relative efficiency. This is the outline of an argument available in detail in Noether (1955) or Kendall and Stuart (1973; Sections 25.5-6). Very often the test statistics are asymptotically normal, so that $G(\cdot)$ has the same form for both tests, however the same argument applies if the test statistics asymptotically have χ^2 distributions with the same degrees of freedom. In this case, the asymptotic relative efficiency can be shown to be the ratio of the two noncentrality parameters. This was shown by Hannan (1956). Hájek and Šidák (1967) use this as the definition of the asymptotic relative efficiency of two tests based on asymptotically χ^2 statistics with equal degrees of freedom. However the argument does not apply to χ^2 statistics with unequal degrees of freedom, since the power functions have different forms, say G_1 and G_2, and equating the arguments of the two functions no longer insures equal power. This point was apparently overlooked by Kendall and Stuart (1973; Section 25.15). Hamdan (1964) computed the Pitman efficiency of a smooth test based on Walsh functions relative to Neyman's smooth test with the same degrees of freedom for a simple hypothesis of normality when the alternative is a shift in the mean.

Blomqvist (1950) defined the relative efficiency of two tests to be the limiting ratio of sample sizes required to insure that the asymptotic power functions have equal slopes at the value specified by the null hypothesis. Noether (1955) showed that under weak conditions this is equivalent to Pitman's measure of efficiency; see also Kendall and Stuart (1973; Section 25.8). Shirahata (1976) redefined the local asymptotic efficiency of Hájek and Šidák (1967) as the slope of the power function at the origin. In his Theorem 5.1 he then showed the following. Suppose two test statistics of $H_0: \delta = 0$, T_1 and T_2, are asymptotically distributed as $\chi^2_{k_1, c_1\delta^2}$ and $\chi^2_{k_2, c_2\delta^2}$, respectively. Then for tests of size α, the local asymptotic relative efficiency of the first test to the second is

$$(6.1) \qquad e(T_1, T_2; \alpha) = \frac{c_1}{c_2} \times \frac{1-\alpha-\Pr\{\chi^2_{k_1+2} \leq \chi^2_{k_1}(1-\alpha)\}}{1-\alpha-\Pr\{\chi^2_{k_2+2} \leq \chi^2_{k_2}(1-\alpha)\}} .$$

This is just the ratio of slopes of the power functions where, for convenience, we may think of differentiating with respect to δ^2. The local asymptotic relative efficiency is the product of the ratio of non-centrality parameters with a term which adjusts for the unequal degrees of freedom. Note that this measure of efficiency depends on α, the size of the test. For reference, values of $1 - \alpha - \Pr\{\chi^2_{k+2} \leq \chi^2_k(1-\alpha)\}$ are tabulated in Table 9 for various k and α.

Table 9. Values of $1 - \alpha - \Pr\{\chi^2_{k+2} \leq \chi^2_k(1-\alpha)\}$.

	α		
k	.10	.05	.01
1	.339	.229	.075
2	.230	.150	.046
3	.183	.117	.035
4	.155	.098	.029
5	.136	.086	.025
6	.123	.077	.022
7	.112	.070	.020
8	.104	.065	.019
9	.098	.060	.017
10	.092	.057	.016
11	.087	.054	.015
12	.083	.051	.014
13	.079	.049	.014
14	.076	.047	.013

The evaluation of (6.1) is convenient and illuminating for the comparison of smooth tests with the quadratic score statistic test for some parametric departure as described in Section 2. Testing, for example, a simple hypothesis, we have

$$\psi_k^2 \sim \chi^2_{k,\, n\beta^2 i_{\beta\beta} R^2_{\theta\beta}}$$

$$W_u \sim \chi^2_{1,\, n\beta^2 i_{\beta\beta}}$$

so that

$$(6.2) \qquad e(\psi_k^2, W_u; \alpha) = R^2_{\theta\beta} \frac{1-\alpha-\Pr\{\chi^2_{k+2} \leq \chi^2_k(1-\alpha)\}}{1-\alpha-\Pr\{\chi^2_3 \leq \chi^2_1(1-\alpha)\}}.$$

Similarly, for a composite hypothesis,

$$(6.3) \qquad e(\psi_k^2(\hat{\lambda}_0), W_u(\hat{\lambda}_0); \alpha) = R^2_{\theta\beta|\lambda} \frac{1-\alpha-\Pr\{\chi^2_{k+2} \leq \chi^2_k(1-\alpha)\}}{1-\alpha-\Pr\{\chi^2_3 \leq \chi^2_1(1-\alpha)\}}.$$

Formulas (6.2) and (6.3) make explicit the trade-off that takes place as one increases k, the order of the smooth test. As k increases, so does $R^2_{\theta\beta}$ or $R^2_{\theta\beta|\lambda}$, and hence so does the non-centrality parameter. At the same time, however, increasing the degrees of freedom tends to reduce power, and this is reflected in the term $1 - \alpha - \Pr\{\chi^2_{k+2} \leq \chi^2_k(1-\alpha)\}$. One can expect the efficiency to first increase with k, and then decrease. By selecting the k which maximizes (6. 2) or (6. 3), one selects the best smooth test, in terms of efficiency, for the alternative under consideration. Notice that (6. 2) and (6. 3) represent a generalization of the result of van Eeden (1963) who related Pitman efficiency to the correlation of test statistics. The efficiencies (6. 2) and (6. 3) for the examples previously considered are listed in Table 10.

To obtain a measure of efficiency which does not depend on the size of the test, α, Shirahata (1976) considers

$$e(T_1, T_2) = \lim_{\alpha \to 0} e(T_1, T_2; \alpha)$$

and shows that, following the notation of (6. 1),

$$e(T_1, T_2) = \frac{c_1}{c_2} \times \frac{k_2}{k_1} \tag{6.4}$$

an expression reminiscent of Kendall and Stuart's (1973) equation (25. 64). However, $e(T_1, T_2; \alpha)$ seems to be a more suitable measure of efficiency for two reasons. First, the ratio of the terms $1 - \alpha - \Pr\{\chi^2_{k+2} \leq \chi^2_k(1-\alpha)\}$ when the two degrees of freedom, k_1 and k_2, are fixed is fairly stable as α varies between 0. 01 and 0. 10, the range of typical test sizes. Second, the values of the ratio on that range are closer to each other than to the limiting value

Table 10. Local asymptotic relative efficiencies (α = 0.10).

k	$e(\psi_k^2, W_u; .10)$	$e(\psi_k^2(\hat{\lambda}_0), W_u(\hat{\lambda}_0); .10)$
a. Normal distribution: skewness departure		
1	0.040	0.883
2	0.027	0.599
3	0.248	0.538
4	0.210	0.455
b. Normal distribution: kurtosis departure		
1	0.000	0.000
2	0.119	0.497
3	0.095	0.396
4	0.135	0.395
c. Exponential distribution: Weibull departure		
1	0.030	0.876
2	0.477	0.623
3	0.394	0.521
4	0.393	0.444
d. Exponential distribution: gamma departure		
1	0.876	0.694
2	0.623	0.570
3	0.521	0.491
4	0.444	0.429

k_2/k_1. For example, with $k_1 = 1$, $k_2 = 5$, the value of the ratio is 2.49, 2.66, and 3.00 for $\alpha = 0.1$, 0.05, and 0.01, respectively, while the limiting value is $k_2/k_1 = 5.00$.

It is often of interest to compare the behavior of tests in regions of moderate power, and this leads to a third measure of efficiency. Let $\delta(k, \alpha, \pi)$ denote the value of the noncentrality parameter necessary for a size-α test based on a χ^2 statistic with k degrees of freedom to have power π. For the smooth tests and the parametric tests, we have seen that these may be written

$$\delta(k, \alpha, \pi) = n i_{\beta\beta} R^2_{\theta_1 \ldots \theta_k, \beta} \beta^2(\alpha, \pi)$$

$$\delta(1, \alpha, \pi) = n i_{\beta\beta} \beta^2(\alpha, \pi)$$

for ψ_k^2 and W_u, respectively. Notice that α and π enter the expressions only through β. For moderate values of π, say, for example, $\pi = 0.5$, the consistency of χ^2 tests insures that power π may be attained locally if n is sufficiently large. Now for α and π fixed, in order to insure that power π is achieved by both tests at the same value of $\beta^2(\alpha, \pi)$ we must have

$$\delta(k, \alpha, \pi) / n_1 i_{\beta\beta} R^2_{\theta_1 \ldots \theta_k, \beta} = \delta(1, \alpha, \pi) / n_2 i_{\beta\beta}$$

or

$$n_2/n_1 = R^2_{\theta_1 \ldots \theta_k, \beta} \frac{\delta(1, \alpha, \pi)}{\delta(k, \alpha, \pi)}$$

where n_1, n_2 are the sample sizes for ψ_k^2 and W_u, respectively. So it is natural to define a moderate-power relative efficiency of ψ_k^2 to W_u, for α and π given, by the asymptotic ratio of sample sizes:

$$e(\psi_k^2, W_u; \alpha, \pi) = R^2_{\theta_1 \ldots \theta_k, \beta} \frac{\delta(1, \alpha, \pi)}{\delta(k, \alpha, \pi)} \tag{6.5}$$

Values of the noncentrality parameters $\delta(k, \alpha, \pi)$ are tabulated in, for example, Pearson and Hartley (1972; Table 25). More extensive tables, based on work by G. E. Hayman, Z. Govindarajulu, and F. C. Leone, are reproduced in Harter and Owen (1970). From these tables it may be seen that $\delta(k, \alpha, \pi)$ increases with the degrees of freedom, k, when α and π are fixed. The generalization of (6.5) to tests of composite hypotheses is obvious. Investigation of these efficiencies will not be pursued at length since the measure depends on both α and π, however Table 11 lists values of $e(\psi_k^2(\hat{\lambda}_0), W_u(\hat{\lambda}_0); .10, .5)$ for tests of the normal and exponential distributions. Notice that the entries in Table 11 are roughly comparable to, and never less than, the corresponding entries in Table 10.

It is of interest to note that each of the three measures of efficiency (6.2), (6.4), and (6.5), or its obvious generalization to composite hypotheses, is of the form

$$e = R^2(k)g(k)$$

where $g(k)$ is a decreasing function in k which depends on the measure of efficiency being used and represents the "cost" of increasing the degrees of freedom, which must be balanced against the gain in $R^2(k)$.

Table 11. Moderate-power efficiencies of smooth tests for composite hypotheses ($\alpha = 0.1$, $\pi = 0.5$).

k	β_1 Departure	β_2 Departure
a. Normal distribution		
1	0.883	0.000
2	0.671	0.557
3	0.644	0.474
4	0.574	0.497
	Weibull Departure	Gamma Departure
b. Exponential distribution		
1	0.876	0.694
2	0.697	0.638
3	0.624	0.588
4	0.560	0.540

Clearly, one is not restricted to comparing only smooth tests with parametric tests by the means described above. Suppose one considers testing the hypothesis by means of another quadratic score statistic ϕ_m^2 with m degrees of freedom based on the m-dimensional score vector U_ω. The parameter ω might, for example, arise by embedding the null density in a family indexed by ω in a manner analogous to the formation of the Neyman-type smooth alternatives. Following the derivation above we could calculate the efficiency (6.2) for the parametric β departure: with obvious notation,

$$e(\phi_m^2, W_u; \alpha) = R^2_{\omega_1 \ldots \omega_m, \beta} \frac{1-\alpha-\Pr\{\chi^2_{m+2} \le \chi^2_m(1-\alpha)\}}{1-\alpha-\Pr\{\chi^2_3 \le \chi^2_1(1-\alpha)\}} . \tag{6.6}$$

Then the relative efficiency of the size-α test based on ϕ_m^2 to the test based on ψ_k^2, with respect to β-departures, is

$$\begin{aligned} e(\phi_m^2, \psi_k^2; \alpha) &= e(\phi_m^2, W_u; \alpha)/e(\psi_k^2, W_u; \alpha) \\ &= \frac{R^2_{\omega_1 \ldots \omega_m, \beta}[1-\alpha-\Pr\{\chi^2_{m+2} \le \chi^2_m(1-\alpha)\}]}{R^2_{\theta_1 \ldots \theta_k, \beta}[1-\alpha-\Pr\{\chi^2_{k+2} \le \chi^2_k(1-\alpha)\}]} . \end{aligned} \tag{6.7}$$

With the usual modification, (6.7) can be applied to tests of composite hypotheses.

The relative efficiency in (6.7) permits us to make an interesting comparison. A test for normality with the mean and variance unknown can be based on the statistic

$$T = n\left[\frac{\hat{\beta}_1^2}{6} + \frac{(\hat{\beta}_2-3)^2}{24}\right]$$

where $\hat{\beta}_1$, $\hat{\beta}_2$ are the usual estimators of the skewness and kurtosis. In fact this is just the quadratic score statistic for testing $H_0: \beta_1 = \beta_2 = 0$ with $\lambda = (\mu, \sigma)'$ unknown for the family of densities $h(x|\beta_1, \beta_2, \lambda)$ given by (5.7). Clearly, the partial multiple correlation of $U_{\beta_1|\lambda}$ and $U_{\beta_2|\lambda}$ with $U_{\beta_j|\lambda}$ is unity for $j = 1, 2$, so from (6.7) we find that for a β_j departure

$$e(\psi_k^2(\hat{\lambda}_0), T; \alpha) = R^2_{\theta_1 \ldots \theta_k, \beta_j|\lambda} \frac{1-\alpha-\Pr\{\chi^2_{k+2} \leq \chi^2_k(1-\alpha)\}}{1-\alpha-\Pr\{\chi^2_4 \leq \chi^2_2(1-\alpha)\}}$$

For example, with $\alpha = 0.10$,

$$e(\psi_k^2(\hat{\lambda}_0), T; .10) = R^2_{\theta_1 \ldots \theta_k, \beta_j|\lambda} \frac{1-\alpha-\Pr\{\chi^2_{k+2} \leq \chi^2_k(1-\alpha)\}}{0.230}$$

$$= e(\psi_k^2(\hat{\lambda}_0), W_{uj}(\hat{\lambda}_0); .10)(\frac{0.339}{0.230})$$

$$= 1.47e(\psi_k^2(\hat{\lambda}_0), W_{uj}(\hat{\lambda}_0); .10)$$

where $W_{uj}(\hat{\lambda}_0)$ denotes the quadratic score statistic for testing β_j, $j = 1, 2$. The efficiencies $e(\psi_k^2(\hat{\lambda}_0) W_{uj}(\hat{\lambda}_0); .10)$ are given in Tables 10a and 10b. This result once again demonstrates the effect of increasing the degrees of freedom of a test. For a β_j departure, T is poorer than $W_{uj}(\hat{\lambda}_0)$: $\psi_k^2(\hat{\lambda}_0)$ is nearly 50% more efficient relative to T than to $W_{uj}(\hat{\lambda}_0)$.

III. 7. Comparison of Pearson's χ^2 with the Smooth Tests

The remarks at the end of the previous section lead us to a method of determining the relative efficiency of Pearson's χ^2 test with respect to the smooth goodness-of-fit tests. Pearson's χ^2 test for a simple hypothesis is in fact the quadratic score statistic for a family of alternatives quite similar to the Neyman alternatives (II. 2. 1). We suppose the test to be based on m intervals which partition the range of X into cells of probability $1/m$. Let $I_{m,j}(y)$ denote the indicator function of the interval $((j-1)/m, j/m)$ for $j = 1, \ldots, m-1$, that is

$$I_{m,j}(y) = \begin{cases} 1 & \text{if} \quad (j-1)/m < y < j/m \\ 0 & \text{otherwise} \end{cases} .$$

We consider the alternatives, indexed by the $(m-1)$-dimensional parameter $\omega = (\omega_1, \ldots, \omega_{m-1})'$, for the null density $f(x)$,

$$g(x|\omega) = f(x)\exp\left\{\sum_{j=1}^{m-1} \omega_j I_{m,j}[F(x)] - K(\omega)\right\}$$

where $F(x)$ is the cumulative distribution function under the null hypothesis and $K(\omega)$ is the normalizing constant. For a sample of size n, let n_j denote the number of observations for which $(j-1)/m < F(x) < j/m$, for $j = 1, \ldots, m$, i. e., the number in the jth cell. Then it is easy to show that, for testing $H_0 : \omega = 0$,

$$U'_{\omega} = [n_1 - n/m, \ldots, n_{m-1} - n/m]$$

and

$$i_{\omega\omega} = m^{-2} \begin{bmatrix} m-1 & -1 & \cdots & -1 \\ -1 & m-1 & \cdots & -1 \\ \vdots & \vdots & & \vdots \\ -1 & -1 & \cdots & m-1 \end{bmatrix} .$$

Thus

$$i_{\omega\omega}^{-1} = m \begin{bmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 2 \end{bmatrix} \tag{7.1}$$

and

$$nU'_{\omega} i_{\omega\omega}^{-1} U_{\omega} = \sum_{j=1}^{m} \frac{(n_j - n/m)^2}{n/m}$$

which is, of course, Pearson's χ^2 statistic.

Now we consider, as in Section 2, parametric alternatives to $f(x)$ indexed by a parameter β. To calculate powers and efficiencies of the Pearson test with respect to β alternatives, we must find $i_{\omega\beta}$. But since $U_{\omega_j}(\omega=0, \beta=0) = I_j[F(x|\beta=0)] - m^{-1}$, it follows that

$$(7.2) \qquad i_{\omega_j \beta} = \int_{F^{-1}[(j-1)/m|0]}^{F^{-1}(j/m|0)} U_\beta f(x|\beta=0)dx$$

$$= \int_{F^{-1}[(j-1)/m|0]}^{F^{-1}(j/m|0)} [\partial f(x|\beta)/\partial\beta]_{\beta=0} dx$$

for $j = 1, \ldots, m-1$. Note that typically $F^{-1}(0|0)$ is either 0 or $-\infty$. For the calculation of multiple correlations, we find from (7.1) and (7.2) that

$$(7.3) \qquad i_{\beta\omega} i_{\omega\omega}^{-1} i_{\omega\beta} = m \left\{ \sum_{j=1}^{m-1} i_{\omega_j \beta}^2 + \left[\sum_{j=1}^{m-1} i_{\omega_j \beta} \right]^2 \right\}$$

and

$$\sum_{j=1}^{m-1} i_{\omega_j \beta} = \int_{F^{-1}(0|0)}^{F^{-1}[(m-1)/m|0]} [\partial f(x|\beta)/\partial\beta]_{\beta=0}\, dx.$$

In some cases a closed expression can be found for the integral in (7.2), but numerical integration may be required. For example, for skewness or kurtosis departures from normality, it follows from (3.1) that

$$[\partial h(x|\beta_1, 3)/\partial\beta_1]_{\beta_1=0} = \frac{1}{6} H_3(x)\phi(x),$$

$$[\partial h(x|0, \beta_2)/\partial\beta_2]_{\beta_2=3} = \frac{1}{24} H_4(x)\phi(x).$$

Then, since

$$(7.4) \qquad \int_{-\infty}^{a} \frac{1}{6} H_3(x)\phi(x)dx = \frac{1-a^2}{6}\,\phi(a)$$

$$\int_{-\infty}^{a} \frac{1}{24} H_4(x)\phi(x)dx = \frac{3a-a^2}{24}\,\phi(a),$$

the covariances $i_{\omega_j\beta_\ell}$ for $\ell = 1, 2$ given by (7.2) may be easily calculated. With these one can then calculate $R^2_{\omega_1 \ldots \omega_{m-1}, \beta_\ell}$ using (7.3), and then $e(\chi^2_{m-1}, W_u; \alpha)$ and $e(\chi^2_{m-1}, \psi^2_k; \alpha)$ using (6.2) and (6.7). These results, for $m = 3, 6, 11,$ and 15 cells, for $k = 1, \ldots, 4,$ and for $\alpha = 0.10$ are summarized in Table 12. The relative efficiencies of the χ^2_{m-1} and ψ^2_k tests may be interpreted as the ratio of the ψ^2_k test sample size to the χ^2_{m-1} test sample size necessary to insure that the asymptotic power functions of the tests at $\alpha = 0.10$ have equal slope as functions of β_1^2 or $(\beta_2 - 3)^2$. Notice that for $k = 1$ and kurtosis departures, the relative efficiency is undefined since ψ^2_1 has asymptotically no local power against β_2.

For Weibull and Gamma departures from the exponential distribution we have

$$[\partial f(x|\beta)/\partial\beta]_{\beta=0} = e^{-x}[1+(1-x)\ln x]$$

$$[\partial f(x|\beta)/\partial\beta]_{\beta=0} = e^{-x}[\gamma + \ln x],$$

Table 12. Efficiencies of χ^2 tests for normality: simple hypothesis.

m-1	$R^2_{\omega_1 \ldots \omega_{m-1}, \beta_1}$		$R^2_{\omega_1 \ldots \omega_{m-1}, \beta_2}$	
a. Multiple correlations				
2	0.088		0.146	
5	0.179		0.223	
10	0.337		0.233	
14	0.424		0.249	
m-1	Skewness (ℓ = 1)		Kurtosis (ℓ = 2)	
b. $e(\chi^2_{m-1}, W_u; .10)$ for W_u based on U_{β_ℓ} and β_ℓ departure				
2	0.060		0.099	
5	0.072		0.089	
10	0.091		0.063	
14	0.095		0.056	
	k			
m-1	1	2	3	4
c. $e(\chi^2_{m-1}, \psi^2_k; .10)$ for skewness departure				
2	1.500	2.222	0.242	0.286
5	1.800	2.667	0.290	0.343
10	2.275	3.370	0.367	0.433
14	2.375	3.519	0.383	0.452
d. $e(\chi^2_{m-1}, \psi^2_k; .10)$ for kurtosis departure				
2	-	0.832	1.042	0.733
5	-	0.748	0.937	0.659
10	-	0.529	0.663	0.467
14	-	0.471	0.589	0.415

respectively, where γ is Euler's constant. The covariances (7.2) must be evaluated by numerical integration, but after that the efficiencies are calculated as before; see Table 13.

The relative efficiencies $e(\chi^2_{m-1}, \psi^2_k; .10)$ listed in Tables 12 and 13 vary over a wide range, and this suggests that consideration of the efficiencies relative to W_u based on the parametric score U_β may be more informative. So for example, while $e(\chi^2_2, \psi^2_1; .10) = 5.133$ for a Weibull departure from the exponential distribution implies that ψ^2_1 requires more than five times as many observations as χ^2_2 to insure comparable local asymptotic power, it is perhaps more instructive to note that neither test is highly efficient relative to the quadratic score statistic test based on the Weibull score. From Tables 13b and 10c the efficiencies of χ^2_2 and ψ^2_1 relative to W_u are seen to be 0.154 and 0.030, respectively.

Unfortunately the straightforward analysis described above does not carry over to the more practical case of composite goodness-of-fit hypotheses. The most widely used procedure consists of calculating Pearson's χ^2 statistic with the nuisance parameter, say $\lambda \in \mathbb{R}^s$, replaced by an estimate, often the maximum likelihood estimator $\hat{\lambda}_0$. The resulting "Chernoff-Lehmann" $\chi^2_{CL}(\hat{\lambda}_0)$, if based on m cells, is taken to have a χ^2_{m-s-1} distribution, although Chernoff and Lehmann (1954) showed that $\chi^2_{CL}(\hat{\lambda}_0)$ is distributed as

Table 13. Efficiencies of χ^2 tests for the exponential distribution: simple hypothesis.

m-1	Weibull	Gamma
a. Multiple correlations $R^2_{\omega_1 \ldots \omega_{m-1}, \beta}$		
2	0.227	0.729
5	0.514	0.873
10	0.691	0.934
14	0.758	0.952
b. $e(\chi^2_{m-1}, W_u; .10)$ for W_u based on U_β		
2	0.154	0.495
5	0.206	0.350
10	0.188	0.253
14	0.170	0.213

m-1	k = 1	2	3	4
c. $e(\chi^2_{m-1}, \psi^2_k; .10)$ for Weibull departure				
2	5.133	0.323	0.391	0.392
5	6.867	0.432	0.523	0.524
10	6.267	0.394	0.477	0.478
14	5.667	0.356	0.431	0.433
d. $e(\chi^2_{m-1}, \psi^2_k; .10)$ for gamma departure				
2	0.565	0.795	0.950	1.115
5	0.400	0.562	0.672	0.788
10	0.289	0.406	0.486	0.570
14	0.243	0.342	0.409	0.480

$$\sum_{j=1}^{m-s-1} Z_j^2 + \sum_{j=m-s}^{m-1} \alpha_j Z_j^2$$

where $Z_1, \ldots, Z_{m-1}$ are independently distributed as $N(0, 1)$, and $0 \leq \alpha_j \leq 1$ for $j = m-s, \ldots, m-1$. See also Watson (1959). In terms of the notation of this thesis,

$$\chi^2_{CL}(\hat{\lambda}_0) = U'_{\cdot\omega}(\hat{\lambda}_0) i^{-1}_{\cdot\omega\omega} U_{\cdot\omega}(\hat{\lambda}_0) .$$

Thus the "wrong" covariance matrix has been used, and the statistic is not asymptotically distributed as χ^2. The correct quadratic score statistic

$$\chi^2_K(\hat{\lambda}_0) = U'_{\cdot\omega}(\hat{\lambda}_0) i^{-1}_{\cdot\omega\omega|\lambda}(\hat{\lambda}_0) U_{\cdot\omega}(\hat{\lambda}_0)$$

is attributed by Moore and Spruill (1975) to the unpublished 1971 thesis of C. Khambhampati (K. C. Rao); see also Rao and Robson (1974). From the results of Section II. 1, this statistic will have the asymptotic χ^2_{m-1} distribution, however it is no longer a simple sum of squares.

The comparisons of Pearson's χ^2_{m-1} and ψ^2_k for simple hypotheses can easily be generalized, by using partial multiple correlation coefficients, to handle $\chi^2_K(\hat{\lambda}_0)$ and $\psi^2_k(\hat{\lambda}_0)$. Although it would be of interest to compare $\psi^2_k(\hat{\lambda}_0)$ with the widely used $\chi^2_{CL}(\hat{\lambda}_0)$, this is not possible by the means presented in this paper.

Nevertheless it is of value to compare $\psi_k^2(\hat{\lambda}_0)$ with the statistic that should be used, in order to see how well $\psi_k^2(\hat{\lambda}_0)$ performs. It should be noted that simulation results of Rao and Robson (1974) and Kang (1977) show that for testing normality, $\chi_K^2(\hat{\lambda}_0)$ has consistently greater power than $\chi_{CL}^2(\hat{\lambda}_0)$ against a wide range of alternatives, and Spruill (1976) showed that the approximate Bahadur efficiency of $\chi_K^2(\hat{\lambda}_0)$ is not less than that of $\chi_{CL}^2(\hat{\lambda}_0)$.

The calculations proceed as usual. For example, for testing normality, for the density (5.7) the nuisance parameter scores, where $\lambda' = (\mu, \sigma)$, are

$$U_\mu(0,\lambda) = (x-\mu)/\sigma^2 = \sigma^{-1}H_1[(x-\mu)/\sigma]$$

$$U_\sigma(0,\lambda) = (x-\mu)^2/\sigma^3 - \sigma^{-1} = \sigma^{-1}H_2[(x-\mu)/\sigma]$$

where $H_1(z) = z$ and $H_2(z) = z^2 - 1$ are the first two Hermite polynomials. Also

$$U_{\beta_1}(0,\lambda) = H_3[(x-\mu)/\sigma]/6$$

$$U_{\beta_2}(0,\lambda) = H_4[(x-\mu)/\sigma]/24$$

and, since Hermite polynomials are uncorrelated, it follows that

$$\begin{bmatrix} i_{\beta\beta} & i_{\beta\lambda} \\ & i_{\lambda\lambda} \end{bmatrix} = \mathrm{diag}(1/6, 1/24, 1/\sigma^2, 2/\sigma^2).$$

Thus

$$i_{\beta\beta|\lambda} = i_{\beta\beta} = \mathrm{diag}(1/6, 1/24)$$

$$i_{\omega\beta|\lambda} = i_{\omega\beta}.$$

Also, (7.2) generalizes readily to

$$i_{\omega_j\beta_\ell}(\lambda) = \int_{F^{-1}[(j-1)/m|0,\lambda]}^{F^{-1}(j/m|0,\lambda)} U_{\beta_\ell}(0,\lambda)f(x|0,\lambda)dx$$

and

$$i_{\omega_j\lambda_\ell}(\lambda) = \int_{F^{-1}[(j-1)/m|0,\lambda]}^{F^{-1}(j/m|0,\lambda)} U_{\lambda_\ell}(0,\lambda)f(x|0,\lambda)dx .$$

For the normal example, it follows easily that $i_{\omega\beta}$ is the same as for the simple hypothesis (i. e. is free of λ) and the matrix $i_{\omega\lambda}$ may be calculated directly since

$$\int_{-\infty}^{a} \sigma^{-1}H_1(x)\phi(x)dx = -\sigma^{-1}\phi(a)$$

$$\int_{-\infty}^{a} \sigma^{-1}H_2(x)\phi(x)dx = -a\sigma^{-1}\phi(a).$$

As usual, $i_{\omega\omega}$ does not depend on λ. So $i_{\omega\omega|\lambda}$ can be calculated, and then

$$R^2_{\omega_1 \ldots \omega_{m-1}, \beta_1|\lambda} = 6 i_{\beta_1\omega} i^{-1}_{\omega\omega|\lambda} i_{\omega|\beta_1}$$

$$R^2_{\omega_1 \ldots \omega_{m-1}, \beta_2|\lambda} = 24 i_{\beta_2\omega} i^{-1}_{\omega\omega|\lambda} i_{\omega\beta_1} .$$

These squared partial multiple correlation coefficients, along with several efficiencies, are presented in Table 14. The efficiencies of $\chi^2_K(\hat{\lambda}_0)$ relative to $\psi^2_k(\hat{\lambda}_0)$ are uniformly rather low for $k \leq 4$ and are quite stable in comparison to the entries of Table 12. Notice that for a kurtosis alternative the efficiency with respect to $\psi^2_1(\hat{\lambda}_0)$ is not defined, that statistic having, asymptotically, no local power. Since the entries increase in k, the low order smooth tests are recommended.

For Weibull and gamma departures from the exponential distribution, the calculations are somewhat simplified by the fact that $i_{\omega\beta}$ is obtained by multiplying the matrix obtained for the simple hypothesis by λ^{-1}, where λ as usual denotes the nuisance scale parameter. Moreover, since $U_\lambda(\lambda) = \lambda^{-1}[(x/\lambda)-1]$ is the same for both the Weibull and gamma cases, and since covariances are calculated under the null exponential distribution, it follows that

$$i_{\omega_j\lambda}(\lambda) = \int_{F^{-1}[(j-1)/m|0,\lambda]}^{F^{-1}(j/m|0,\lambda)} \lambda^{-1}[(x/\lambda)-1]\lambda^{-1} \exp(-x/\lambda)dx$$

$$= \lambda^{-1}\{(\frac{m-j}{m})\ln(\frac{m-j}{m})-(\frac{m-j-1}{m})\ln(\frac{m-j-1}{m})\}$$

Table 14. Efficiencies of χ^2 tests for normality: composite hypothesis.

m-1	$R^2_{\omega_1 \ldots \omega_{m-1}, \beta_1 \mid \lambda}$	$R^2_{\omega_1 \ldots \omega_{m-1}, \beta_2 \mid \lambda}$
a. Multiple correlations ($\lambda' = (\mu, \sigma)$)		
2	0.424	0.187
5	0.643	0.471
10	0.786	0.654
14	0.840	0.726
b. $e(\chi^2_K(\hat{\lambda}_0), W_u(\hat{\lambda}_0); .10)$ for $W_u(\hat{\lambda}_0)$ based on $U_{\beta_\ell}(\hat{\lambda}_0)$ and β_ℓ departure		
2	0.288	0.127
5	0.258	0.189
10	0.213	0.177
14	0.188	0.163

m-1	k = 1	2	3	4
c. $e(\chi^2_K(\hat{\lambda}_0), \psi^2_k(\hat{\lambda}_0); .10)$ for skewness departure				
2	0.326	0.481	0.535	0.633
5	0.292	0.431	0.480	0.567
10	0.241	0.356	0.396	0.468
14	0.213	0.314	0.349	0.413
d. $e(\chi^2_K(\hat{\lambda}_0), \psi^2_k(\hat{\lambda}_0); .10)$ for kurtosis departure				
2	-	0.256	0.321	0.322
5	-	0.380	0.477	0.478
10	-	0.356	0.447	0.448
14	-	0.328	0.412	0.413

applies in both cases. The other matrices needed have been calculated in Sections 3 and 4. The partial multiple correlations and efficiencies calculated from these are presented in Table 15. As for the normal case, $\psi_k^2(\hat{\lambda}_0)$ uniformly dominates $\chi_K^2(\hat{\lambda}_0)$ in terms of relative efficiency, and low values of k are preferred.

Table 15. Efficiencies of χ^2 tests for the exponential distribution: composite hypothesis.

m-1	Weibull	Gamma
a. Multiple correlations $R^2_{\omega_1 \ldots \omega_{m-1}, \beta \mid \lambda}$ (λ = scale parameter)		
2	0.580	0.454
5	0.811	0.718
10	0.906	0.846
14	0.937	0.887
b. $e(\chi^2_K(\hat{\lambda}_0), W_u(\hat{\lambda}_0); .10)$ for $W_u(\hat{\lambda}_0)$ based on $U_\beta(\hat{\lambda}_0)$		
2	0.394	0.308
5	0.325	0.288
10	0.246	0.230
14	0.210	0.199

	k			
m-1	1	2	3	4
c. $e(\chi^2_K(\hat{\lambda}_0), \psi^2_k(\hat{\lambda}_0); .10)$ for Weibull departures				
2	0.450	0.632	0.756	0.887
5	0.371	0.522	0.624	0.732
10	0.281	0.395	0.472	0.554
14	0.240	0.337	0.403	0.473
d. $e(\chi^2_K(\hat{\lambda}_0), \psi^2_k(\hat{\lambda}_0); .10)$ for gamma departures				
2	0.444	0.540	0.627	0.718
5	0.415	0.505	0.587	0.671
10	0.331	0.404	0.468	0.536
14	0.287	0.349	0.405	0.464

IV. CONCLUSIONS

IV. 1. Discussion of the Smooth Tests

In Section III. 1 two questions were posed regarding the smooth goodness-of-fit tests: how do they perform relative to other tests, and what choice of k is best? The results of the previous chapter provide us with partial answers to both questions. In the comments that follow, attention will be restricted to the more useful case of composite goodness-of-fit hypotheses, rather than simple hypotheses. Moreover, the conclusions strictly apply to tests for the normal and exponential distributions, since these are the only cases investigated in depth in this paper.

How does $\psi_k^2(\hat{\lambda}_0)$ compare with other goodness-of-fit statistics? Tables 4 and 5 show that, for a wide range of alternatives to normality, a smooth test of some order has at least as much power as the Chernoff-Lehmann χ^2 and Kolmogorov-Smirnov tests, and most often the smooth test has substantially greater power. Moreover, Tables 14c, 14d, 15c, and 15d show that the smooth tests of orders one through four are uniformly more efficient locally against the departures considered than $\chi_K^2(\hat{\lambda}_0)$ for a wide range of degrees of freedom. Thus it appears that the smooth tests are perhaps to be preferred to the other omnibus tests.

The situation for comparing the smooth tests to parametric tests with highly focused power is not quite so clear. For simple hypotheses, W_u represents the locally best test. For composite hypotheses, however, $W_u(\hat{\lambda}_0)$ is in general not locally best or even locally best among similar tests; see Cox and Hinkley (1974, p. 146). $W_u(\hat{\lambda}_0)$ does have the optimality properties described by Wald (1943) and Roussas (1972), as discussed in the Appendix, and is optimal in the class $C(\alpha)$ of tests described by Neyman (1959). Nevertheless, Tables 10 and 11 show that $\psi_k^2(\hat{\lambda}_0)$ is always at least moderately efficient locally relative to the statistic $W_u(\hat{\lambda}_0)$ appropriate to the alternative distributions in question. Since the power of a particular $W_u(\hat{\lambda}_0)$ may be very poor for some other alternative--for example, the skewness test for normality has, locally, no power against kurtosis departures and vice versa, since $i_{\beta_1\beta_2} = 0$ --the smooth test is preferred on the basis of giving better protection against a wider range of alternatives.

How should k be chosen? Neyman (1937) felt that k need not be taken greater than four in most practical problems, but clearly warned that this was not a mathematical result. The evidence in the present work suggests that a value of two or perhaps even one may serve very adequately. Table 5 strongly supports the use of $k = 2$ for testing normality, and this is born out by the efficiencies in Table 10 and 11. Although $\psi_2^2(\hat{\lambda}_0)$ is somewhat less efficient (both locally and at moderate power) than $\psi_1^2(\hat{\lambda}_0)$ for skewness departures, $\psi_2^2(\hat{\lambda}_0)$ is preferred since $\psi_1^2(\hat{\lambda}_0)$ provides no protection against heavy- or light-tailed alternatives. Tables 10 and 11 also support the use of $\psi_1^2(\hat{\lambda}_0)$ for the exponential distribution. This conclusion would be much stronger, however, if a larger class of alternatives was considered.

IV. 2. Discussion of the χ^2 Tests

The results of Section III. 7 permit us to extend the discussion above to the χ^2 tests of goodness of fit. The tables in that section provide further evidence of a fact already widely known: Pearson's χ^2 (and its proper generalization $\chi_K^2(\hat{\lambda}_0)$) have relatively little power against any particular alternative. In addition, however, Tables 14b and 15b may aid in the selection of m, the number of cells on which the test is based. This problem, which is analogous to the problem of selecting k in the smooth test, has received wide attention. Mann and Wald (1942) based the choice of m on the sample size and size of the test, but Williams (1950) suggested that their solution could be safely halved. For further discussion see Dahiya and Gurland (1973) and Hamdan (1963). The efficiencies in Tables 14 and 15 suggest that the degrees of freedom should be rather smaller: perhaps less than ten as a general rule. Since these efficiencies change rather slowly as the degrees of freedom vary, an exact determination of the number of cells does not seem necessary. If the ranges of best efficiencies with respect to different alternatives roughly coincide, which Tables 14b and 15b show to be the case for the normal and exponential cases, then the choice of a suitable range of degrees of freedom on the basis of efficiency is straightforward. That the selection of the number of cells can possibly depend on the alternative distribution may, however, make efficiency-based choice either more or less worthwhile than a choice which ignores the alternative, depending on the statistician's level of knowledge of possible alternatives.

BIBLIOGRAPHY

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. New York: Wiley.

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

Bartlett, M. S. (1953). Approximate Confidence Intervals. II. More Than One Unknown Parameter. Biometrika 40:306-317.

Barton, D. E. (1953). On Neyman's Smooth Test of Goodness of Fit and Its Power with Respect to a Particular System of Alternatives. Skandinavsk Aktuarietidskrift 36:24-64.

Barton, D. E. (1955). A Form of Neyman's ψ_k^2 Test of Goodness of Fit Applicable to Grouped and Discrete Data. Skandinavsk Aktuarietidskrift 38:1-16.

Barton, D. E. (1956). Neyman's ψ_k^2 Test of Goodness of Fit When the Null Hypothesis is Composite. Skandinavsk Aktuarietidskrift 39:216-246.

Bhat, B. R., and Nagnur, B. N. (1965). Locally Asymptotically Most Stringent Tests and Lagrangian Multiplier Tests of Linear Hypotheses. Biometrika 52:459-468.

Blomqvist, N. (1950). On a Measure of Dependence between Two Random Variables. Annals of Mathematical Statistics 21:593-600.

Bowman, K. O. (1973). Power of the Kurtosis Statistic, b_2, in Tests of Departures from Normality. Biometrika 60:623-628.

Bowman, K. O., and Shenton, L. R. (1973). Notes on the Distribution of $\sqrt{b_1}$ in Sampling from Pearson Distributions. Biometrika 60:155-167.

Bowman, K. O., and Shenton, L. R. (1973). Remarks on the Distribution of $\sqrt{b_1}$ in Sampling from a Normal Mixture, and a Normal Type A Distribution. Journal of the American Statistical Association 68:998-1003.

Bowman, K.O., and Shenton, L.R. (1975). Omnibus Test Contours for Departures from Normality Based on $\sqrt{b_1}$ and b_2. Biometrika 62:243-251.

Chernoff, H., and Lehmann, E.L. (1954). The Use of Maximum Likelihood Estimates in χ^2 Tests of Goodness of Fit. Annals of Mathematical Statistics 25:579-586.

Cox, D.R., and Hinkley, D.V. (1974). Theoretical Statistics. London: Chapman and Hall.

D'Agostino, R.B. (1970). Transformation to Normality of the Null Distribution of g_1. Biometrika 57:679-681.

D'Agostino, R.B. (1971). An Omnibus Test of Normality for Moderate and Large Sample Sizes. Biometrika 58:341-348.

D'Agostino, R.B. (1973). Monte Carlo Power Comparison of the W' and D Tests of Normality for N = 100. Communications in Statistics 1:545-551.

D'Agostino, R.B., and Pearson, E.S. (1973). Tests for Departure from Normality. Empirical Results for the Distributions of b_2 and $\sqrt{b_1}$. Biometrika 60:613-622.

D'Agostino, R.B., and Rosman, B. (1974). The Power of Geary's Test of Normality. Biometrika 61:181-184.

D'Agostino, R.B., and Tietjen, G.L. (1971). Simulation Probability Points of b_2 for Small Samples. Biometrika 58:669-672.

Dahiya, R.C., and Gurland, J. (1973). How Many Classes in the Pearson Chi-square Test? Journal of the American Statistical Association 68:707-712.

David, F.N. (1939). On Neyman's 'Smooth' Test of Goodness of Fit. Biometrika 31:191-199.

Durbin, J. (1975). Kolmogorov-Smirnov Tests when Parameters are Estimated with Applications to Tests of Exponentiality and Tests on Spacings. Biometrika 62:5-22.

Dyer, A.R. (1974). Comparisons of Tests for Normality with a Cautionary Note. Biometrika 61:185-189.

Gnanadesikan, R. (1977). Statistical Data Analysis of Multivariate Observations. New York: Wiley.

Hājek, J. (1962). Asymptotically Most Powerful Rank-Order Tests. Annals of Mathematical Statistics 33:1124-1147.

Hājek, J., and Šidāk, Z. (1967). Theory of Rank Tests. New York: Academic Press.

Hamdan, M. A. (1962). The Powers of Certain Smooth Tests of Goodness of Fit. Australian Journal of Statistics 4:25-40.

Hamdan, M. A. (1963). The Number and Width of Classes in the Chi-square Test. Journal of the American Statistical Association 58:678-689.

Hamdan, M. A. (1964). A Smooth Test of Goodness of Fit Based on the Walsh Functions. Australian Journal of Statistics 6:130-136.

Hannan, E. J. (1956). The Asymptotic Powers of Certain Tests Based on Multiple Correlations. Journal of the Royal Statistical Society-Series B 18:227-233.

Harter, H. L., and Owen, D. B. (1970). Selected Tables in Mathematical Statistics-Volume I. Chicago: Markham.

Hensler, G. L., Mehrota, K. G., and Michalek, J. E. (1977). A Goodness of Fit Test for Multivariate Normality. Communications in Statistics-Theory and Methods A6:33-41.

Javitz, H. S. (1975). Generalized Smooth Tests of Goodness of Fit, Independence, and Equality of Distributions. Unpublished thesis: University of California, Berkeley.

Kang, S. (1977). Performance of Some Neyman Smooth Goodness of Fit Tests. Thesis research in progress: Oregon State University Department of Statistics.

Kendall, M. G., and Stuart, A. (1969). The Advanced Theory of Statistics-Volume I (3rd edition). New York: Hafner.

Kendall, M. G., and Stuart, A. (1973). The Advanced Theory of Statistics-Volume II (3rd edition). New York: Hafner.

Kowalski, C. J. (1970). The Performance of Some Rough Tests for Bivariate Normality before and after Coordinate Transformations to Normality. Technometrics 12:517-544.

Lehmann, E. L. (1959). Testing Statistical Hypotheses. New York: Wiley.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Journal of the American Statistical Association 62:399-402.

Lilliefors, H. W. (1969). On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. Journal of the American Statistical Association 64:387-389.

Lind, B., and Roussas, G. (1972). A Remark on Quadratic Mean Differentiability. Annals of Mathematical Statistics 43:1030-1034.

Malkovich, J. F., and Afifi, A. A. (1973). On Tests for Multivariate Normality. Journal of the American Statistical Association 68:176-179.

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

Mardia, K. V. (1974). Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies. Sankhya-Series B 36:115-128.

Margolin, B. H., and Maurer, W. (1976). Tests of the Kolmogorov-Smirnov Type for Exponential Data with Unknown Scale, and Related Problems. Biometrika 63:149-160.

Moore, D. S., and Spruill, M. C. (1975). Unified Large-sample Theory of General Chi-squared Statistics for Tests of Fit. Annals of Statistics 3:599-616.

Moran, P. A. P. (1970). On Asymptotically Optimal Tests of Composite Hypotheses. Biometrika 57:47-55.

Neyman, J. (1937). 'Smooth' Test for Goodness of Fit. Skandinavsk Aktuarietidskrift 20:149-199.

Neyman, J. (1959). Optimal Asymptotic Tests of Composite Statistical Hypotheses, in Probability and Statistics (U. Grenander, ed.) Stockholm: Almquist and Wiksell.

Noether, G.E. (1955). On a Theorem of Pitman. Annals of Mathematical Statistics 26:64-68.

Pearson, E.S. (1938). The Probability Integral Transformation for Testing Goodness of Fit and Combining Independent Tests of Significance. Biometrika 30:134-148.

Pearson, E.S., and Hartley, H.O. (1972). Biometrika Tables for Statisticians-Volume II. Cambridge: Cambridge University Press.

Peers, H.W. (1971). Likelihood Ratio and Associated Test Criteria. Biometrika 58:577-587.

Prescott, P. (1976). Comparison of Tests of Normality Using Stylized Sensitivity Surfaces. Biometrika 63:285-289.

Rao, C.R. (1948). Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation. Proceedings of the Cambridge Philosophical Society 44: 50-57.

Rao, C.R. (1973). Linear Statistical Inference and Its Applications. New York: Wiley.

Rao, K.C., and Robson, D.S. (1974). A Chi-square Statistic for Goodness-of-fit Tests within the Exponential Family. Communications in Statistics 3:1139-1153.

Roussas, G.G. (1968). Asymptotic Normality of the Maximum Likelihood Estimate in Markov Processes. Metrika 14:62-70.

Roussas, G.G. (1972). Contiguity of Probability Measures: Some Applications in Statistics. Cambridge: Cambridge University Press.

Shapiro, S.S., Wilk, M.B., and Chen, H.J. (1968). A Comparative Study of Various Tests for Normality. Journal of the American Statistical Association 63:1343-1372.

Shenton, L.R., and Bowman, K.O. (1975). Johnson's S_u and the Skewness and Kurtosis Statistics. Journal of the American Statistical Association 70:220-228.

Shirahata, S. (1976). Tests for Independence in Infinite Contingency Tables. Annals of Statistics 4:542-553.

Spruill, M.C. (1976). A Comparison of Chi-square Goodness-of-fit Tests Based on Approximate Bahadur Slope. Annals of Statistics 4:409-412.

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

Thomas, D.R. (1977). On Neyman Smooth Goodness of Fit Tests for Censored Data. Unpublished research in progress.

Thomas, D.R., and Pierce, D.A. (1977). Neyman's Smooth Goodness of Fit Test When the Hypothesis is Composite. Technical Report 59. Oregon State University Department of Statistics.

van Eeden, C. (1963). The Relation between Pitman's Asymptotic Relative Efficiency of Two Tests and the Correlation Coefficient between their Test Statistics. Annals of Mathematical Statistics 34:1442-1451.

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large. Transactions of the American Mathematical Society 54:426-482.

Watson, G.S. (1959). Some Recent Results in Chi-square Goodness-of-fit Tests. Biometrics 15:440-468.

Williams, C.A. (1950). On the Choice of the Number and Width of Classes for the Chi-square Test of Goodness of Fit. Journal of the American Statistical Association 45:77-86.

APPENDIX

APPENDIX

This appendix contains a discussion of the asymptotic distributional and optimality theory of tests of simple and composite hypotheses based on quadratic score statistics, with special reference to smooth goodness-of-fit tests. First the asymptotic equivalence of the quadratic score statistics with likelihood ratio and Wald's statistics is considered. Little mathematical detail is provided at this point. Sets of regularity conditions for simple and composite hypotheses are then listed, expressed in a general terminology, followed by a discussion of the distributional and optimality results they guarantee. Finally, simplifications and restatements of the regularity conditions suitable for smooth goodness-of-fit tests are presented.

Rao (1948) suggested the use of quadratic score statistics for testing simple and composite hypotheses. We consider $H_0 : \theta = \theta_0$ in either case, with λ denoting the nuisance parameter for composite H_0. Bartlett (1953) employed the statistic to obtain confidence regions for θ in the presence of an unknown λ. Neyman (1959) gave a detailed derivation of the statistic for composite H_0, as well as an optimality property, to be mentioned below, for the case of real-valued θ, and Bhat and Nagnur (1965) generalized that work to deal with vector-valued θ.

A test of $H_0:\theta = \theta_0$ can also be based on the likelihood ratio statistic or on Wald's statistic, the latter, due to Wald (1943), being a quadratic form in $(\hat{\theta}-\theta_0)$ with a χ^2 distribution, where $\hat{\theta}$ denotes the maximum likelihood estimator of θ. Let W and W_e, respectively, denote the two test statistics in the case of a simple hypothesis, and $W(\hat{\lambda}_0)$ and $W_e(\hat{\lambda}_0)$ the same for a composite hypothesis; this notation is due to Cox and Hinkley (1974; Section 9.3). Wald (1943; Section 13) demonstrated the asymptotic equivalence of tests based on $W(\hat{\lambda}_0)$ and $W_e(\hat{\lambda}_0)$ under regularity conditions. In fact, under regularity conditions the quadratic score statistic, W_u or $W_u(\hat{\lambda}_0)$, is asymptotically equivalent to W and W_e or $W(\hat{\lambda}_0)$ and $W_e(\hat{\lambda}_0)$, as the case may be; see, for example Cox and Hinkley (1974; Section 9.3), Moran (1970), or Javitz (1975; Section 13). Peers (1971) considered approximations of the power functions of W, W_e, and W_u and found that none of the statistics dominated either of the others, even when the approximate distributions of the statistics are improved by the inclusion of terms of higher order. The following outline of a proof of the asymptotic equivalence of $W(\hat{\lambda}_0)$ and $W_u(\hat{\lambda}_0)$ will ignore questions of regularity conditions. Consider a family of densities $g(x|\theta,\lambda)$ and the composite hypothesis $H_0:\theta = \theta_0$, where λ is unknown. Let

$$\ell.(\theta,\lambda) = \sum_{i=1}^{n} \ln g(X_i \mid \theta,\lambda).$$

The unrestricted joint maximum likelihood estimator is denoted $(\hat{\theta},\hat{\lambda})$, while $\hat{\lambda}_0$ is the maximum likelihood estimator of λ under the assumption $H_0:\theta = \theta_0$. Then

$$\text{(A. 1)} \qquad \frac{1}{2} W(\hat{\lambda}_0) = \ell.(\hat{\theta},\hat{\lambda}) - \ell.(\theta_0,\hat{\lambda}_0)$$
$$= [\ell.(\hat{\theta},\hat{\lambda}) - \ell.(\theta_0,\lambda)] - [\ell.(\theta_0,\hat{\lambda}_0) - \ell.(\theta_0,\lambda)]$$

where λ denotes the true, unknown value of the nuisance parameter. We suppose $\theta \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}^s$, and let

$$U._{\theta_i}(\theta^*,\lambda^*) = [\partial\ell.(\theta,\lambda)/\partial\theta_i]_{\theta^*,\lambda^*}$$
$$U._{\lambda_j}(\theta^*,\lambda^*) = [\partial\ell.(\theta,\lambda)/\partial\lambda_j]_{\theta^*,\lambda^*}$$

for $i = 1,\ldots,k$ and $j = 1,\ldots,s$. The score vectors are then

$$U._{\theta}(\theta,\lambda) = [U._{\theta_1}(\theta,\lambda),\ldots,U._{\theta_k}(\theta,\lambda)]'$$
$$U._{\lambda}(\theta,\lambda) = [U._{\lambda_1}(\theta,\lambda),\ldots,U._{\lambda_s}(\theta,\lambda)]'$$

The information matrix, with obvious partitioning, may be written

$$\mathrm{Cov}_{\theta,\lambda}\begin{bmatrix} U._{\theta}(\theta,\lambda) \\ U._{\lambda}(\theta,\lambda) \end{bmatrix} = \begin{bmatrix} i._{\theta\theta}(\theta,\lambda) & i._{\theta\lambda}(\theta,\lambda) \\ & i._{\lambda\lambda}(\theta,\lambda) \end{bmatrix} .$$

Approximating each term in brackets in (A. 1) to first order we obtain, suppressing the dependence on (θ,λ)

(A. 2) $$W(\hat{\lambda}_0) \doteq [U.'_{\theta}, U.'_{\lambda}]\begin{bmatrix} i._{\theta\theta} & i._{\theta\lambda} \\ & i._{\lambda\lambda} \end{bmatrix}^{-1}\begin{bmatrix} U._{\theta} \\ U._{\lambda} \end{bmatrix} - U.'_{\lambda} i.^{-1}_{\lambda\lambda} U._{\lambda}$$

The first quadratic form in (A. 2) may be written in terms of $T[U.'_{\theta}, U.'_{\lambda}]'$ rather than $[U.'_{\theta}, U.'_{\lambda}]'$, and an efficacious choice of T can lead to a simpler expression for the right-hand side. We take

$$T = \begin{bmatrix} I_k & -i._{\theta\lambda} i.^{-1}_{\lambda\lambda} \\ 0 & I_s \end{bmatrix}$$

where I_k denotes the $k \times k$ identity matrix. Let

(A. 3) $$\begin{bmatrix} U._{\theta|\lambda} \\ U._{\lambda} \end{bmatrix} = T\begin{bmatrix} U._{\theta} \\ U._{\lambda} \end{bmatrix}$$

so that $U._{\theta|\lambda} = U._{\theta} - i._{\theta\lambda} i.^{-1}_{\lambda\lambda} U._{\lambda}$ and

$$\mathrm{Cov}_{\theta_0,\lambda}\begin{bmatrix} U._{\theta|\lambda} \\ U._{\lambda} \end{bmatrix} = \begin{bmatrix} i._{\theta\theta|\lambda} & 0 \\ & i._{\lambda\lambda} \end{bmatrix}$$

where

(A.4) $$i._{\theta\theta|\lambda} = i._{\theta\theta|\lambda}(\theta_0, \lambda)$$
$$= i._{\theta\theta}(\theta_0, \lambda) - i._{\theta\lambda}(\theta_0, \lambda) i.^{-1}_{\lambda\lambda}(\theta_0, \lambda) i._{\lambda\theta}(\theta_0, \lambda).$$

Substituting into (A.2) we then find that

(A.5) $$W(\hat{\lambda}_0) \doteq U.'_{\theta|\lambda}(\theta_0, \lambda) i.^{-1}_{\theta\theta|\lambda}(\theta_0, \lambda) U._{\theta|\lambda}(\theta_0, \lambda)$$

which is just the desired approximate equivalence, when λ on the right-hand side is replaced by $\hat{\lambda}_0$.

By considering the residuals, $U._{\theta|\lambda}$, from the regression of $U._\theta$ on $U._\lambda$, we use only that part of the θ-scores which are orthogonal to the λ-scores. Another way of viewing the test is as an approximately similar test. This follows from the asymptotic normality of the efficient score vectors, from which $[U.'_\theta, U.'_\lambda]'$ is approximately sufficient for (θ, λ), and $U._\lambda$ is approximately a complete sufficient statistic for λ under H_0; see, for example, Cox and Hinkley (1974; Section 9.2v). Thus the test for $H_0: \theta = \theta_0$ should be based on the conditional distribution of $U._\theta$ given $U._\lambda$, which is approximately $N_k(i._{\theta\lambda} i.^{-1}_{\lambda\lambda} U._\lambda, i._{\theta\theta} - i._{\theta\lambda} i.^{-1}_{\lambda\lambda} i._{\lambda\theta})$.

The removal of the unknown true value of λ from the right-hand side of (A.5) is a crucial step in obtaining a usable statistic. To eliminate the dependence of the quadratic score statistic on λ, the unknown parameter value is replaced by an estimate $\hat{\lambda}_0$ which is guaranteed to be "close" to the true value. Then, under regularity conditions, the statistic evaluated at $(\theta_0, \hat{\lambda}_0)$ is asymptotically equivalent to the same evaluated at (θ_0, λ) as in (A. 2), thus providing a usable test statistic. Neyman (1959) showed that weak root-n consistency of the estimator provides such a result. Consider a sequence $\{\hat{\lambda}_j^{(n)}\}$ of estimators for the jth component of λ. $\{\hat{\lambda}_j^{(n)}\}$ is said to be root-n consistent if, for all θ and λ, $n^{1/2}|\hat{\lambda}_j^{(n)} - \lambda_j|$ remains bounded in probability as $n \to \infty$, and is said to be locally root-n consistent if there exists a nonzero $A_j \in \mathbb{R}^k$ such that, for all θ and λ, $n^{1/2}|\hat{\lambda}_j^{(n)} - \lambda_j - A_j'(\theta - \theta_0)|$ remains bounded in probability as $n \to \infty$. Finally if $\{\hat{\lambda}^{(n)}\}$ is a sequence of componentwise (locally) root-n consistent estimators, and if for every $a \in \mathbb{R}^s$ such that $a'a > 0$, the assumption that $\theta = \theta_0$ implies that $n^{1/2} a'(\hat{\lambda}^{(n)} - \lambda)$ does not tend in probability to zero, then $\{\hat{\lambda}^{(n)}\}$ is said to be weakly root-n consistent. Notice that a sufficient condition for weak root-n consistency is that the joint limiting distribution of $n^{1/2}(\hat{\lambda}^{(n)} - \lambda)$ be nondegenerate, which is true of maximum likelihood estimators under standard regularity conditions.

Neyman (1959; Theorem 1) and Bhat and Nagnur (1965; Equation (2. 9)) show that

$$\text{(A. 6)} \qquad i_{\cdot\theta\theta|\lambda}^{-1/2}(\theta_0, \hat{\lambda}_0^{(n)}) U_{\cdot\theta|\lambda}(\theta_0, \hat{\lambda}_0^{(n)}) - i_{\cdot\theta\theta|\lambda}^{-1/2}(\theta_0, \lambda) U_{\cdot\theta|\lambda}(\theta_0, \lambda) \xrightarrow{P} 0$$

if $\{\hat{\lambda}_0^{(n)}\}$ is a weakly root-n consistent estimator, under the assumption that $i_{\cdot\theta\theta|\lambda}^{-1/2}(\theta_0, \lambda)U_{\cdot\theta|\lambda}(\theta_0, \lambda)$ is a normed Cramer function. This argument is essentially reproduced by Javitz (1975), and the following discussion of conditions under which quadratic score statistics have asymptotic χ^2 distributions and provide optimal tests is derived from the work of Roussas (1972) and of Javitz.

We now consider sets of regularity conditions under which distributional and optimality results may be obtained, chiefly by means of arguments based on contiguous probability measures. The conditions will be expressed in a general notation, and later will be applied to smooth tests of both simple and composite hypotheses. We begin with the sample space $(\mathbb{R}, \overline{\mathcal{B}}(\mathbb{R}))$, where $\overline{\mathcal{B}}(\mathcal{A})$ in general denotes the σ-field of Lebesgue measurable subsets of $\mathcal{A}$, for a subset of some finite dimensional Euclidean space. On $(\mathbb{R}, \overline{\mathcal{B}}(\mathbb{R}))$ is a family $\{P(\xi); \xi \in \Xi\}$ of probability measures indexed by a vector parameter ξ ranging over a bounded parameter set $\Xi \subset \mathbb{R}^k$. Further, we assume that each $P(\xi)$ is absolutely continuous with respect to Lebesgue measure μ, and let $g(x|\xi) = [dP(\xi)/d\mu]$ for all $\xi \in \Xi$. For each ξ, we then can define the infinite product probability space $(\mathbb{R}^{\infty}, \overline{\mathcal{B}}^{\infty}(\mathbb{R}), P^{\infty}(\xi))$ corresponding to the infinite sequence $\{X_1, X_2, \ldots\}$ of independent, identically distributed X's, and let $P^{(n)}(\xi)$ denote the restriction of $P^{\infty}(\xi)$ to the sub-σ-field of $\overline{\mathcal{B}}^{\infty}(\mathbb{R})$ generated by $\{X_1, \ldots, X_n\}$ for $n = 1, 2, \ldots$.

We first consider regularity conditions for a simple hypothesis. For notational consistency we let $\theta = \xi$, $\Theta = \Xi$. The regularity conditions are listed below.

[S1] The probability measures $P(\theta)$ are mutually absolutely continuous for $\theta \in \Theta$.

[S2] For all $\theta \in \Theta$ and for $j = 1, \ldots, k$, $U_{\theta_j}(\theta)$ exists and is continuous in θ.

[S3] For all $\theta \in \Theta$ and for $j = 1, \ldots, k$, $E\{U^2_{\theta_j}(\theta) | \theta\} < \infty$.

[S4] For all $\theta \in \Theta$ and for $j = 1, \ldots, k$ there exist a neighborhood $\mathcal{N}_j(\theta)$ containing θ and a function $H_j(x|\theta)$ such that for all $\theta^* \in \mathcal{N}_j(\theta)$

$$|U_{\theta_j}(\theta^*) - U_{\theta_j}(\theta)| \leq H_j(x|\theta) \quad \text{a.e. } P(\theta)$$

and such that $E\{H_j^2(X|\theta)|\theta\} < \infty$.

[S5] $U_\theta(\theta)$ is $\overline{\mathcal{B}}(\mathbb{R}) \times \overline{\mathcal{B}}(\Theta)$-measurable.

[S6] $i_{\theta\theta}(\theta)$ is positive definite for all $\theta \in \Theta$.

[S7] For each fixed $\theta_0 \in \Theta$, $[g(x|\theta^*)/g(x|\theta)]$, considered as a function of x and θ^*, is $\overline{\mathcal{B}}(\mathbb{R}) \times \overline{\mathcal{B}}(\Theta)$-measurable.

Conditions [S5]-[S7] are simply restatements of (A2')(ii)-(A2')(iv) in Roussas (1972; page 46). Also, [S1] implies Roussas' (A1'). Finally, from Lind and Roussas (1972; Theorem 2.2), it follows that [S2]-[S4] imply Roussas' (A2')(i). Thus [S1]-[S7] give us Roussas'

(A1') and (A2') and therefore by his Theorems 4.2 and 4.6 (page 54) we can conclude that

$$\mathcal{L}\ [n^{-1/2}U_{\cdot\theta}(\theta)\,|\,P^{(n)}(\theta)] \to N_k(0, i_{\theta\theta}(\theta))$$

for all $\theta \in \Theta$, and for $\theta^{(n)} = \theta + n^{-1/2}h$ where $h \in \mathbb{R}^k$,

$$\mathcal{L}\ [n^{-1/2}U_{\cdot\theta}(\theta)\,|\,P^{(n)}(\theta^{(n)})] \to N_k(i_{\theta\theta}(\theta)h, i_{\theta\theta}(\theta))$$

and therefore

$$\mathcal{L}\ [W_u\,|\,P^{(n)}(\theta)] \to \chi^2_k$$

$$\mathcal{L}\ [W_u\,|\,P^{(n)}(\theta^{(n)})] \to \chi^2_{k,\,\delta(\theta)}$$

where $\delta(\theta) = h' i_{\theta\theta}(\theta)h$. In addition, under (A1') and (A2') of Roussas, three optimality properties of the test based on W_u are obtained. These will be discussed in detail following the presentation of regularity conditions for composite hypotheses. At the end of this appendix it will be seen that [S1]-[S7] are satisfied by the smooth goodness-of-fit test for any simple hypothesis specifying a continuous distribution.

We now turn our attention to composite hypotheses. Let $\theta \in \Theta$ be as before and take $\lambda \in \Lambda$ to be the nuisance parameter, where $\Lambda \subset \mathbb{R}^s$. Now $\xi = (\theta', \lambda)'$ and $\Xi = \Theta \times \Lambda$. The regularity conditions are now listed.

[C1] The probability measures $P(\theta,\lambda)$ are mutually absolutely continuous for $(\theta',\lambda')' \in \Theta \times \Lambda$.

[C2] The density $g(x|\theta,\lambda)$ is $\bar{\mathcal{B}}(\mathbb{R}) \times \bar{\mathcal{B}}(\Theta \times \Lambda)$-measurable.

[C3] For all $(\theta',\lambda')' \in \Theta \times \Lambda$, the first and second partial derivatives of $g(x|\theta,\lambda)$ with respect to all combinations of $\theta_1,\ldots,\theta_k,\lambda_1,\ldots,\lambda_s$, evaluated at any $(\theta',\lambda')'$ exist and are finite, and all the first partial derivatives are $\bar{\mathcal{B}}(\mathbb{R}) \times \bar{\mathcal{B}}(\Theta \times \Lambda)$-measurable and continuous in $(\theta',\lambda')'$.

[C4] The covariance matrix

$$\mathrm{Cov}\left\{\begin{bmatrix} U_\theta(\theta,\lambda) \\ U_\lambda(\theta,\lambda) \end{bmatrix} \middle| \theta,\lambda \right\} = \begin{bmatrix} i_{\theta\theta}(\theta,\lambda) & i_{\theta\lambda}(\theta,\lambda) \\ & i_{\lambda\lambda}(\theta,\lambda) \end{bmatrix}$$

is finite, positive definite, and continuous in $(\theta',\lambda')'$ for all $(\theta',\lambda')' \in \Theta \times \Lambda$.

[C5] For all $(\theta',\lambda')' \in \Theta \times \Lambda$,

$$\{[\partial g(x|\theta^*,\lambda^*)/\partial\theta_j^*]_{\theta^*,\lambda^*}\}/g(x|\theta,\lambda)$$

and

$$\{[\partial g(x|\theta^*,\lambda^*)/\partial\lambda_\ell^*]_{\theta^*,\lambda^*}\}/g(x|\theta,\lambda)$$

are continuous in $(\theta^{*\prime},\lambda^{*\prime})'$, for $j = 1,\ldots,k$ and $\ell = 1,\ldots,s$.

[C6] For every $(\theta',\lambda')' \in \Theta \times \Lambda$ there exist a neighborhood $\eta_1(\theta,\lambda)$ containing $(\theta',\lambda')'$ and functions $H_{1,j}(x|\theta,\lambda)$

and $H_{2,\ell}(x|\theta,\lambda)$ such that

$$|\{[\partial g(x|\theta^*,\lambda^*)/\partial\theta_j^*]_{\theta^*,\lambda^*}\}/g(x|\theta,\lambda)| \leq H_{1,j}(x|\theta,\lambda)$$

a. e. $P(\theta,\lambda)$

and

$$|\{[\partial g(x|\theta^*,\lambda^*)/\partial\lambda_\ell^*]_{\theta^*,\lambda^*}\}/g(x|\theta,\lambda)| \leq H_{2,j}(x|\theta,\lambda)$$

a. e. $P(\theta,\lambda)$

for $j = 1,\ldots,k$ and $\ell = 1,\ldots,s$, and for $(\theta^{*\prime},\lambda^{*\prime})' \in \mathcal{N}_1(\theta,\lambda)$ and such that for $m = 1,2$ and $a = j$ or ℓ,

$$E\{H_{m,a}^2(X|\theta,\lambda)|\theta,\lambda\} < \infty.$$

[C7] For every $(\theta',\lambda')' \in \Theta \times \Lambda$ there exist a neighborhood $\mathcal{N}_2(\theta,\lambda)$ containing $(\theta',\lambda')'$ and functions $G_1(x|\theta,\lambda)$ and $G_2(x|\theta,\lambda)$ such that

$$\int_{\mathbb{R}} G_1(x|\theta,\lambda)d\mu(x) < \infty$$

and

$$\int_{\mathbb{R}} G_2(x|\theta,\lambda)g(x|\theta,\lambda)d\mu(x) < \infty$$

and such that

$$\left.\begin{array}{l} |[\partial g(x|\theta,\lambda)/\partial\theta_j]_{\theta*,\lambda*}| \\ |[\partial g(x|\theta,\lambda)/\partial\lambda_\ell]_{\theta*,\lambda*}| \\ |[\partial^2 g(x|\theta,\lambda)/\partial\theta_i\partial\theta_j]_{\theta*,\lambda*}| \\ |[\partial^2 g(x|\theta,\lambda)/\partial\theta_j\partial\lambda_\ell]_{\theta*,\lambda*}| \\ |[\partial^2 g(x|\theta,\lambda)/\partial\lambda_\ell\partial\lambda_m]_{\theta*,\lambda*}| \end{array}\right\} \leq G_1(x|\theta,\lambda)$$

and

$$\left.\begin{array}{l} |[\partial \ln g(x|\theta,\lambda)/\partial\theta_j]_{\theta*,\lambda*}| \\ |[\partial \ln g(x|\theta,\lambda)/\partial\lambda_\ell]_{\theta*,\lambda*}| \\ |[\partial^2 \ln g(x|\theta,\lambda)/\partial\theta_i\partial\theta_j]_{\theta*,\lambda*}| \\ |[\partial^2 \ln g(x|\theta,\lambda)/\partial\theta_j\partial\lambda_\ell]_{\theta*,\lambda*}| \\ |[\partial^2 \ln g(x|\theta,\lambda)/\partial\lambda_\ell\partial\lambda_m]_{\theta*,\lambda*}| \end{array}\right\} \leq G_2(x|\theta,\lambda)$$

for all $(\theta^{*\prime},\lambda^{*\prime})' \in \mathcal{N}_2(\theta,\lambda)$ and for $i,j = 1,\ldots,k$ and $\ell,m = 1,\ldots,s$.

[C8] The sequence $\{\hat{\lambda}_0^{(n)}\}$ is a system of weakly root-n consistent estimators of λ under $P^{(n)}(0,\lambda)$ for all $\lambda \in \Lambda$. Here we are tacitly assuming that the null hypothesis of interest specifies $\theta = 0$.

It is easy to show that under these assumptions the conditions [A4. 1]-[A4. 8], [A5. 1]-[A5. 5], and [AII. 1]-[AII-2] of Javitz (1975) are satisfied. Notice that [C8] makes no reference to alternative distributions $P(\theta,\lambda)$ with $\theta \neq 0$. Now under [C1]-[C6] we may apply

Javitz' Theorem 4.1 (page 21) to conclude that

$$\mathcal{L}\,[n^{-1/2}U_{\cdot\theta|\lambda}(0,\lambda)\,|\,P^{(n)}(\theta^{(n)},\lambda) \to N_k(i_{\theta\theta|\lambda}(0,\lambda)\theta,\, i_{\theta\theta|\lambda}(0,\lambda))$$

under alternatives $\theta^{(n)} = n^{-1/2}\theta$ for some $\theta \in \Theta$ and any $\lambda \in \Lambda$, where $U_{\theta|\lambda} = U_\theta - i_{\theta\lambda} i_{\lambda\lambda}^{-1} U_\lambda$. Moreover, under [C1]-[C8], via Javitz' Lemmas 5.1 and 5.2 (page 28) and Theorem 5.1 (page 29), we can conclude that

$$U_{\cdot\theta|\lambda}(0,\hat{\lambda}_0^{(n)}) - U_{\cdot\theta|\lambda}(0,\lambda) \xrightarrow{P} 0$$

under the alternatives $P^{(n)}(\theta^{(n)},\lambda)$. This is the necessary analogue of (A.6). In view of [C4] it follows from the Central Limit Theorem that

$$\mathcal{L}\,[n^{-1/2}U_{\cdot\theta|\lambda}(0,\lambda)\,|\,P^{(n)}(0,\lambda)] \to N_k(0, i_{\theta\theta|\lambda}(\theta))$$

so we find that

$$\mathcal{L}\,[W_u(\hat{\lambda}_0)\,|\,P^{(n)}(0,\lambda)] \to \chi_k^2$$

$$\mathcal{L}\,[W_u(\hat{\lambda}_0)\,|\,P^{(n)}(\theta^{(n)},\lambda)] \to \chi^2_{k,\,\delta(\theta|\lambda)}$$

where

$$\delta(\theta|\lambda) = \theta' i_{\theta\theta|\lambda}(\theta,\lambda)\theta.$$

In his proof of Theorem 4.1, Javitz (page 91) notes that his [A4.1]-[A4.8] and [AII.1]-[AII.2] (and therefore our [C1]-[C6]) imply that (A1') and (A2') of Roussas (1972; page 46) are satisfied. So, as mentioned above for [S1]-[S7], the optimality properties of Roussas' Theorems 2.1 (page 170), 2.2 (page 171), and 6.1 (page 191) will also be obtained. For simple hypotheses, these theorems parallel Theorems I-III in Wald (1943) and for composite hypotheses, they parallel Wald's Theorems IV-VI. Of course Wald's theorems apply to W_e and $W_e(\hat{\lambda}_0)$, and carry over to W_u and $W_u(\hat{\lambda}_0)$ because of the asymptotic equivalence. The results of Roussas' three theorems will be listed here.

To describe the optimality results we require some additional notation, which will be very similar to that of Roussas. In the following, $i_{\xi\xi}(\xi_0)$ will represent $i_{\theta\theta}(\theta_0)$ for simple hypotheses and $i_{\theta\theta|\lambda}(\theta_0, \lambda)$ for composite hypotheses. We define the ellipsoidal surfaces

$$E(c) = \{\xi \in \mathbb{R}^k \,|\, (\xi-\xi_0)' i_{\xi\xi}(\xi_0)(\xi-\xi_0) = c\}$$

for $c \in (0, \infty)$, and the function

$$a(\xi) = [|i_{\xi\xi}(\xi_0)|(\xi-\xi_0)' i_{\xi\xi}(\xi_0)(\xi-\xi_0)]^{1/2} / \| i_{\xi\xi}(\xi_0)(\xi-\xi_0)\|$$

for $\xi \in \mathbb{R}^k - \{\xi_0\}$. Further we define a weight function

$$\zeta(c,\xi) = a(\xi)/A(c)$$

where A is defined by the surface integral

$$A(c) = \int_{E(c)} a(\xi)dA.$$

Notice that $\int_{E(c)} \zeta(c,\xi)dA = 1$ for all $c \in (0,\infty)$. We will deal with alternatives $\xi^{(n)} = \xi_0 + hn^{-1/2}$ for $h \in \mathbb{R}^k$, so let

$$E(n,c) = E(cn^{-1}) = \{\xi \in \mathbb{R}^k \mid (\xi-\xi_0)' i_{\xi\xi}(\xi_0)(\xi-\xi_0) = cn^{-1}\}$$

and

$$\zeta_n(c,\xi) = \zeta(cn^{-1},\xi).$$

In the results that follow we will take $c \in K$, a compact subset of $(0,\infty)$, so for sufficiently large n, $E(n,c) \subset \Xi$ for all $c \in K$.

We restrict our attention to tests of $H_0: \xi = \xi_0$ which depend on the data only through $n^{-1/2}U_{\cdot\xi}(\xi_0)$, where $U_\xi(\xi_0) = U_\theta(\theta_0)$ or $U_{\theta|\lambda}(\theta_0,\lambda)$ as needed, and let $\phi[n^{-1/2}U_{\cdot\xi}(\xi_0)]$ denote such a test: ϕ gives the probability of rejecting H_0 as a function of the data through $n^{-1/2}U_{\cdot\xi}(\xi_0)$. Let $I_A(\cdot)$ denote the indicator function of the set $A \in \overline{\mathcal{B}}^k(\mathbb{R})$, and define

$$\overline{\mathcal{F}} = \{\phi \mid \phi = I_{C^c} \text{ for closed, convex } C \in \overline{\mathcal{B}}^k(\mathbb{R})\}.$$

For the test based on W_u (which denotes W_u or $W_u(\lambda)$ as needed)

$$C_{W_u} = \{z \in \mathbb{R}^k \mid z' i_{\xi\xi}^{-1}(\xi_0) z \leq \chi_k^2(1-\alpha)\}$$

so, clearly, for ϕ_{W_u} defined accordingly, $\phi_{W_u} \in \tilde{\mathcal{F}}$. We also will restrict our attention to sequences of tests in $\tilde{\mathcal{F}}$ which are asymptotically of size α, i.e., $\{\phi_n\} \subset \tilde{\mathcal{F}}$ for which

$$E\{\phi_n \mid \xi_0\} = E\{I_{C_n^c} \mid \xi_0\} \rightarrow \alpha$$

as $n \rightarrow \infty$. Because of the asymptotic χ^2 distribution of W_u, ϕ_{W_u} clearly satisfies this condition.

Finally, let $\pi_n(\xi, \phi)$ denote the power of the test ϕ with sample size n, under $\xi \in \Xi$,

$$\pi_n(\xi, \phi) = E\{\phi[n^{-1/2} U_{\cdot\xi}(\xi_0)] \mid \xi\}$$

and for clarity let $\pi_n(\xi, W_u) = \pi_n(\xi, \phi_{W_u})$. Then (A1') and (A2') of Roussas (1972; page 46) imply the following results (pages 170, 171):

[R1] For any sequence of tests $\{\phi_n\} \subset \tilde{\mathcal{F}}$ asymptotically of size α, for an arbitrary compact set $K \subset (0, \infty)$, and for testing $H_0: \xi = \xi_0$ against $H_1: \xi \neq \xi_0$, we have

$$\liminf_{n} \left\{ \inf_{c \in K} \left[\int_{E(n,c)} \pi_n(\xi, W_u) \zeta_n(c, \xi) dA - \int_{E(n,c)} \pi_n(\xi, \phi_n) \zeta_n(c, \xi) dA \right] \right\} \geq 0.$$

[R2] Under the same assumptions as [R1], we have

(a) $$\lim_{n \to \infty} \left\{ \sup_{c \in K} \left[\sup_{\xi \in E(n,c)} \pi_n(\xi, W_u) - \inf_{\xi \in E(n,c)} \pi_n(\xi, W_u) \right] \right\} = 0$$

and for any $\{\phi_n\}$ as in [R1] for which $\pi_n(\xi, \phi_n)$ satisfies (a),

(b) $$\liminf_{n} \left\{ \inf_{c \in K} \left[\inf_{\xi \in E(n,c)} \{\pi_n(\xi, W_u) - \pi_n(\xi, \phi_n)\} \right] \right\} \geq 0.$$

These results may be roughly interpreted as follows. [R1] simply states that, approximately, W_u or $W_u(\lambda)$ has asymptotically the best average power on the surfaces $E(n, c)$ of tests in $\tilde{\mathcal{F}}$, where the average is taken with respect to the weight function ζ_n. The first part of [R2] implies that, approximately, the power of the test based on W_u or $W_u(\lambda)$ is asymptotically constant on the surfaces $E(n, c)$, and the second part of [R2] implies that, approximately, W_u or $W_u(\lambda)$ gives the test which is asymptotically most powerful among tests in $\tilde{\mathcal{F}}$ with power constant on the surfaces $E(n, c)$.

The third optimality result requires some additional notation. We define the class of tests $\tilde{\mathcal{F}}_0 \subset \tilde{\bar{\mathcal{F}}}$ by

$$\tilde{\mathcal{F}}_0 = \{\phi \mid \phi = I_{C^c} \text{ for closed, convex } C \in \bar{\mathcal{B}}^k(\mathbb{R}) \text{ such that } \Pr[C \mid \xi_0] = 1-\alpha\}.$$

This permits us to work with the envelope power function $\pi_n(\xi, \alpha)$ defined by

$$\pi_n(\xi, \alpha) = \sup_{\phi \in \tilde{\mathcal{F}}_0} \{\pi_n(\xi, \phi)\}.$$

Using the comments above and Theorem 6.1 of Roussas (1972; page 191), it follows that (A1') and (A2') insure the following result.

[R3] For any sequence of tests $\{\phi_n\} \in \tilde{\mathcal{F}}_0$, for an arbitrary compact set $K \subset (0, \infty)$, and for testing $H_0 : \xi = \xi_0$ against $H_1 : \xi \neq \xi_0$, we have

$$\limsup_n \left\{ \sup_{c \in K} \left[\sup_{\xi \in E(n,c)} \{\pi_n(\xi, \alpha) - \pi_n(\xi, W_u)\} - \sup_{\xi \in E(n,c)} \{\pi_n(\xi, \alpha) - \pi_n(\xi, \phi_n)\} \right] \right\} \leq 0.$$

Thus the test based on W_u or $W_u(\lambda)$ is asymptotically most stringent in the class $\tilde{\mathcal{F}}_0$, i.e., if we call the difference between a given test's power at some ξ and $\pi_n(\xi, \alpha)$ the shortcoming of the test at ξ, then W_u minimizes the maximum shortcoming.

To conclude the discussion of optimality properties we note that Neyman (1959) and Bhat and Nagnur (1965) showed that the test based on $W_u(\hat{\lambda}_0)$ is locally asymptotically most powerful or stringent in the class $C(\alpha)$ of tests; their regularity conditions are of the traditional Cramér type.

To conclude this appendix, we consider the evaluation of [S1]-[S7] and [C1]-[C8] for smooth goodness-of-fit tests. First we consider simple hypotheses. Since the multiplying factor

$$\exp\left\{\sum_{i=1}^{k} \theta_i F^i(x) - K(\theta)\right\}$$

is strictly positive, [S1] is true. Since

$$K(\theta) = \ln \int_0^1 \exp\left\{\sum_{i=1}^{k} \theta_i y^i\right\} dy$$

we may apply the Dominated Convergence Theorem to show that

$$\partial g(x|\theta)/\partial\theta_j = F^j(x) - \partial K(\theta)/\partial\theta_j$$

exists and is continuous in θ; hence [S2] is true. Also, [S3] is satisfied because of the boundedness of F, and for [S4], simply note that $|U_{\theta_j}(\theta^*) - U_{\theta_j}(\theta)|$ is constant in x. [S6] follows from simple calculations, and [S5] and [S7] may easily be seen to be true. Thus the regularity conditions, and therefore the asymptotic distributional and optimality theory, are always obtained for the smooth test of a simple goodness-of-fit hypothesis under smooth alternatives.

Of course the situation is more complicated for composite hypotheses. However substantial simplifications occur under the following conditions on the null family with densities $[dP(0, \lambda)/d\mu] = f(x|\lambda)$.

[C1'] The probability measures $P(0, \lambda)$ are mutually absolutely continuous for $\lambda \epsilon \Lambda$.

[C2'] $\partial f(x|\lambda)/\partial\lambda_\ell$, $\partial F(x|\lambda)/\partial\lambda_\ell$, and $\partial^2 F(x|\lambda)/\partial\lambda_\ell\partial\lambda_m$ exist and, evaluated at any $\lambda \epsilon \Lambda$, are finite, for $\ell, m = 1, \ldots, s$.

[C3'] $f(x|\lambda)$, $F(x|\lambda)$, $\partial f(x|\lambda)/\partial\lambda_\ell$, $\partial F(x|\lambda)/\partial\lambda_\ell$ are $\overline{\mathcal{B}}(\mathbb{R}) \times \overline{\mathcal{B}}(\Lambda)$-measurable and continuous in λ, for $\ell = 1, \ldots, s$.

[C4'] For all $\lambda \epsilon \Lambda$ there exist a neighborhood $\mathcal{N}(\lambda)$ containing λ and functions $H_{1,\ell}(x|\lambda)$ and $H_{2,\ell}(x|\lambda)$ such that

$$|U_{\lambda_\ell}(\lambda^*)| \leq H_{1,\ell}(x|\lambda) \qquad \text{a.e. } P(0,\lambda)$$

and

$$|[\partial F(x|\lambda)/\partial\lambda_\ell]_{\lambda^*}| \leq H_{2,\ell}(x|\lambda) \qquad \text{a.e. } P(0,\lambda)$$

for all $\lambda^* \epsilon \mathcal{N}(\lambda)$, and such that

$$E\{H^2_{m,\ell}(X|\lambda)|\lambda\} < \infty$$

for $m = 1, 2$ and $\ell = 1, \ldots, s$.

By reasoning analogous to the verification of [S1] for simple goodness-of-fit hypotheses, we find that [C1'] implies [C1]. Also [C2] follows from [C2']. The first and second partial derivatives of [C3] can all be written in terms of $F(x|\lambda)$, $g(x|\theta,\lambda)$, the first and second partial λ-derivatives of f and F and the first and second partial θ-derivatives of $K(\theta)$. So [C3] can be shown to be true under [C2'] and [C3']. For [C5], we have

$$\{[\partial g(x|\theta^*,\lambda^*)/\partial\theta_j^*]_{\theta^*,\lambda^*}\}/g(x|\theta,\lambda)$$

$$= [f(x|\lambda^*)/f(x|\lambda)] \exp\left\{\sum_{i=1}^{k}[\theta_i^* F^i(x|\lambda^*) - \theta_i F^i(x|\lambda)] - [K(\theta^*) - K(\theta)]\right\}$$

$$\times \{F^j(x|\lambda^*) - [\partial K(\theta)/\partial\theta_j]_{\theta^*}\}$$

and continuity can be obtained using [C3'] and the fact that

$$K(\theta) = \int_0^1 \exp\left\{\sum_{i=1}^{k}\theta_i y^i\right\} dy.$$

Similarly

$$[\partial g(x|\theta^*, \lambda^*)/\partial\lambda_\ell^*]_{\theta^*,\lambda^*}/g(x|\theta,\lambda)$$

$$= [f(x|\lambda^*)/f(x|\lambda)] \exp\left\{\sum_{i=1}^{k}[\theta_i^* F^i(x|\lambda^*) - \theta_i F^i(x|\lambda)] - [K(\theta^*) - K(\theta)]\right\}$$

$$\times \left\{U_{\lambda_\ell}(\lambda^*) + [\partial F(x|\lambda)/\partial\lambda_\ell]_{\lambda^*}\sum_{i=1}^{k} i\theta_i^* F^{i-1}(x|\lambda^*)\right\}.$$

From this, [C3'], and the nature of $K(\theta)$ we see that [C5] is satisfied. Moreover, with [C4'] parts of [C6] and [C7] may be satisfied. Further conditions on the null densities $f(x|\lambda)$ might be found to bring about additional simplifications.

It should be noted that [C8] does not require $\{\hat{\lambda}_0^{(n)}\}$ to be a sequence of maximum likelihood estimators. However, as mentioned above, if $\hat{\lambda}_0^{(n)}$ is the maximum likelihood estimator of λ under the null hypothesis, and if $n^{1/2}(\hat{\lambda}_0^{(n)} - \lambda)$ is asymptotically normal, which happens under well-known regularity conditions, for example those of the Cramer type, then $\{\hat{\lambda}_0^{(n)}\}$ is a system of weakly root-n consistent estimators.

The null χ^2 distribution of $\psi^2(\hat{\lambda}_0)$ can also be obtained under relatively weak assumptions. We assume that $U_\lambda(\lambda)$ is a vector-valued Cramer function as defined by Neyman (1959). If, in addition, there exist functions $H_\ell(x|\lambda)$, $H_{\ell,m}(x|\lambda)$, and $H_{\ell,m,t}(x|\lambda)$ for all $\lambda \in \Lambda$ and for $\ell, m, t = 1, \ldots, s$ such that

$$|\partial F(x|\lambda)/\partial\lambda_\ell| \leq H_\ell(x|\lambda)$$

$$|\partial^2 F(x|\lambda)/\partial\lambda_\ell\partial\lambda_m| \leq H_{\ell,m}(x|\lambda)$$

$$|\partial^3 F(x|\lambda)/\partial\lambda_\ell\partial\lambda_m\partial\lambda_t| \leq H_{\ell,m,t}(x|\lambda)$$

a. e. μ and such that

$$E[H_\ell^3(X|\lambda)|\lambda] < \infty$$

$$E[H_{\ell,m}^2(X|\lambda)|\lambda] < \infty$$

$$E[H_{\ell,m,t}(X|\lambda)|\lambda] < \infty,$$

then $U_\theta(\lambda)$ is also a vector-valued Cramer function, and as Neyman notes (1959; Section 8(i)), $U_{\theta|\lambda}(\lambda)$ is therefore also a Cramer function. Finally, if $\{\hat{\lambda}_0^{(n)}\}$ is the sequence of maximum likelihood estimators and the null distributions are sufficiently regular so that $n^{1/2}(\hat{\lambda}_0^{(n)}-\lambda)$ is asymptotically normal--see, for example, Cox and Hinkley (1974; Chapter 9)--then by Neyman (1959; Theorem 1) and the multivariate central limit theorem, the asymptotic null χ^2 distribution is obtained.