TESTS OF FIT BASED ON THE CORRELATION COEFFICIENT

BY

RICHARD A. LOCKHART and M. A. STEPHENS

TECHNICAL REPORT NO. 436
OCTOBER 4, 1990

Prepared Under Contract
N00014-89-J-1627 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

Reproduction in Whole or in Part is Permitted
for any purpose of the United States Government

Approved for public release; distribution unlimited.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
1. **INTRODUCTION.**

1.1 **The regression model in goodness-of-fit.**

Suppose a random sample \( X_1, X_2, \ldots, X_n \) comes from distribution \( F_0(x) \) and let \( X_1, X_2, \ldots, X_n \) be the order statistics. \( F_0(x) \) may be of the form \( F(w) \) with \( w = (x - \alpha)/\beta \); \( \alpha \) is then the location parameter and \( \beta \) is the scale parameter of \( F_0(x) \). There may be other parameters in \( F(w) \), for example, a shape parameter; here we assume such parameters known, but \( \alpha \) and \( \beta \) are unknown. We can suppose the random sample of \( X \)-values to have been constructed from a random sample \( w_1, w_2, \ldots, w_n \) from \( F(w) \), by the transformation

\[
X_i = \alpha + \beta w_i .
\]  

(1)

If the order statistics of the \( w \)-sample are \( w_1 < w_2 < \ldots < w_n \), we have also

\[
X(i) = \alpha + \beta w(i) .
\]  

(2)

Let \( E(w(i)) \) be \( m_i \) and let \( v_{ij} \) be \( E(w(i) - m_i)(w(j) - m_j) \); let \( V \) be the \( n \times n \) matrix with entries \( v_{ij} \). \( V \) is the covariance matrix of the order statistics \( w(i) \). From (2) we have

\[
E(X(i)) = \alpha + \beta m_i
\]  

(3)

and a plot of \( X(i) \) against \( m_i \) should be approximately a straight line with intercept \( \alpha \) on the vertical axis and slope \( \beta \). The values \( m_i \) are the most natural values to plot along the horizontal axis to achieve a straight line plot, but for most distributions they are difficult to calculate.
Various authors have therefore proposed alternatives $T_i$ which are convenient functions of $i$; then (2) can be replaced by the model

$$X(i) = \alpha + \beta T_i + \epsilon_i$$  \hspace{1cm} (4)

where $\epsilon_i$ is an "error" which has mean zero only for $T_i = m_i$.

A common choice for $T_i$ is $H_i \equiv F^{-1}\{i/(n+1)\}$ or similar expressions which approximate $m_i$. A test of

$$H_0: \text{the X-sample comes from } F_0(x),$$  \hspace{1cm} (5)

can then be based on how well the data fits the line (3) or (4).

1.2 Example. As an example, suppose it is desired to test that the X-sample is normally distributed, with unknown mean $\mu$ and variance $\sigma^2$.

Then $F(w) = \frac{1}{\sqrt{2\pi}} \int_0^w e^{-t^2/2} \, dt$, and the w-sample is standard normal.

Then (1) becomes

$$X_i = \mu + \sigma \omega_i$$

and (3) is

$$E(X(i)) = \mu + \sigma m_i$$

where $m_i$ are the expected values of standard normal order statistics.

For this distribution, $\alpha = \mu$ and $\beta = \sigma$. 
1.3 Measures of fit.

The practice of plotting the \( X_{(1)} \) against \( m_i \) (or against another set of constants \( T_i \) which approximate the \( m_i \)-values) and looking to see if a straight line results, is time-honored as a quick technique for testing normality. An improvement on this procedure by eye, is to measure how well the data fits the line (3). Three main approaches to measuring the fit can be identified. The first is simply to measure the correlation coefficient \( R(X,T) \) between the paired sets \( X_i \) and \( T_i \). A second method is to estimate the line \( \alpha + \beta T_i \), using generalized least squares to take into account the covariance of the order statistics, and then to base the test of fit on the sum of squares of residuals. Finally, a third technique is to estimate \( \beta \) from (2) using generalized least squares, and to compare this estimate with the estimate of scale given by the sample standard deviation. In this article we explore the first two of these methods, which are often closely connected.

1.4 The correlation coefficient.

The simplest of the three methods above is to use the correlation co-efficient \( R(X,T) \). Here we extend the usual meaning of correlation, and also that of variance and covariance, to apply to constants as well as random variables. Thus let \( X \) refer to the vector \( X_{(1)}, \ldots, X_{(n)} \), and \( T \) to vector \( T_1, \ldots, T_n \); let \( \bar{X} = \frac{\Sigma X_{(i)}}{n} \) and \( \bar{T} = \frac{\Sigma T_i}{n} \), (all sums are for \( i = 1 \) to \( n \)) and define the sums
\[ S(X,T) = \sum (X_{(i)} - \bar{X})(T_i - \bar{T}) = \sum X_{(i)}T_i - n\bar{X}\bar{T}; \]

\[ S(X,X) = \sum (X_{(i)} - \bar{X})^2 = \sum (X_i - \bar{X})^2; \]

\[ S(T,T) = \sum (T_i - \bar{T})^2. \]

\[ S(X,X) \] will often be called \[ S^2. \]

The variance of \[ X \] is then \[ V(X,X) = \frac{1}{n-1} S(X,X), \] the variance of \[ T \] is \[ V(T,T) = \frac{1}{n-1} S(T,T), \] and the covariance of \[ X \] and \[ T \] is \[ V(X,T) = \frac{1}{n-1} S(X,T). \] The correlation coefficient between \[ X \] and \[ T \] is

\[ R(X,T) = \frac{V(X,T)}{\{V(X,X)V(T,T)\}^{\frac{1}{2}}} = \frac{S(X,T)}{\{S(X,X)S(T,T)\}^{\frac{1}{2}}}. \]

Statistics \[ R(X,m) \] (called sometimes \[ R \]) or \[ R^2(X,m) \] are attractive statistics for testing the fit of \[ X \] to the model (2), since if a "perfect" sample is given, that is, a sample whose ordered values fall exactly at their expected values, \[ R(X,m) \] will be 1, and the value of \[ R(X,m) \] can be interpreted as a measure of how closely the sample resembles a perfect sample. Then tests based on \[ R(X,m) \], or equivalently on \[ R^2(X,m) \] will be one-tailed; rejection of \[ H_0 \] occurs only for low values of \[ R \]. Suppose \[ \hat{X}_{(i)} = \hat{\alpha} + \hat{\beta}T_i \], where \[ \hat{\alpha} \] and \[ \hat{\beta} \] are the usual regression estimators of \[ \alpha \] and \[ \beta \] (ignoring the covariance between the \[ X_{(i)} \]). It is possible to set up the standard ANOVA table for straight line regression:
Regression \[ SS = \frac{s^2(X,T)}{s(T,T)}, \]

Error \[ SS = s^2 - \frac{s^2(X,T)}{s(T,T)} = \sum(X_{(i)} - \hat{X}_{(i)})^2 \]

Total \[ SS = s^2 = S(X,X) \]

and it is clear that

\[
\frac{\text{Error SS}}{\text{Total SS}} = 1 - R^2(X,T).
\]

Define, for any \( T \) vector,

\[ Z(X,T) = n[1 - R^2(X,T)]. \]

Then \( Z(X,T) \) is a test statistic equivalent to \( R^2(X,T) \), based on the sum of squares of the residuals after the line (3) has been fitted. \( Z(X,T) \) has, in common with many other goodness-of-fit statistics e.g., chi-square, and EDF statistics, the property that the larger \( Z(X,T) \) is, the worse the fit. Sarkadi [1975] and more recently Gerlach [1979] have shown consistency for correlation tests based on \( R(X,m) \), or equivalently \( Z(X,m) \), for a wide class of distributions including all the usual continuous distributions. This is to be expected, since for large \( n \) we expect a sample to become perfect in the sense above. We can expect the consistency property to extend to \( R(X,T) \) provided \( T \) approaches \( m \) sufficiently rapidly for large samples.

1.5 Censored data. \( R^2(X,T) \) can easily be calculated for censored data, provided the ranks of the available \( X_{(i)} \) are known. These
are paired with the appropriate \( T_i \) and \( R^2(X,T) \) is calculated using the same formula as above, with the sums running over the known \( i \).

For example if the data were right censored, so that only the \( r \) smallest values \( X(i) \) were available, the sums would run for \( i \) from 1 to \( r \); if the data were left-censored, with the first \( s \) values missing, the \( i \) would run from \( s+1 \) to \( n \). Tables of \( Z(X,T) \) for \( T = m \) or \( H \), for testing for the uniform, normal, exponential, logistic, or extreme-value distributions have been published by Stephens (1986).

2. **CORRELATION TESTS FOR THE UNIFORM DISTRIBUTION.**

For the uniform distribution for \( X \), between limits \((a,b)\), written \( U(a,b) \), we have \( F(w) = w, 0 < w < 1 \), and \( X_i = a + (b - a)W_i \); hence \( \alpha = a, \beta = b - a \). Then \( m_i = E(W(i)) = i/(m+1) \); also \( H_i = m_i \).

The order statistics \( X(i) \) could be plotted against \( i \) instead of against \( i/(n+1) \); the scale factor \( 1/(n+1) \) does not change the correlation coefficient, and \( R(X,m) = R(X,H) = R(X,T) \) where \( T_i = i \).

In discussing tests for the uniform distribution, we distinguish four cases:

- **Case 0:** \( a, b \) both known;
- **Case 1:** \( a \) unknown, but \((b-a)\) known;
- **Case 2:** \( a \) known, \((b-a)\) unknown;
- **Case 3:** both \( a \) and \( b \) unknown.

**Case 0.** Here \( a \) and \( b \) are both known, so that \( \alpha \) and \( \beta \) are known in (1). The transformation \( X' = (X-a)/(b-a) \) then reduces the problem to
a test that \( X' \) is \( U(0,1) \). There are of course many tests for this special case (see, e.g., Stephens, 1986). In the present context, the test will be based on the residuals from the known line \( F(x') = x', \ 0 < x' < 1; \) that is, on the statistic \( Z_0 = \sum (x' - i/(n+1))^2 \). It is clear that \( Z_0 \) has the same asymptotic distribution as the well-known Cramér-von Mises statistic \( W^2 = \sum (x' - (2i-1)/(2n))^2 + 1/(12n) \), and, for small samples, the two statistics will have much the same power properties.

**Case 1.** Here the model is \( X = a + \beta \bar{W}, \) with \( \beta = b-a \) known.

Substitute \( x' = X/\beta \); then the model becomes \( x' = a/\beta + W \), and \( E(x') = \alpha + (m_1 - \bar{m}) \), where \( \alpha = a/\beta + \bar{m} \). Ordinary least squares gives \( \hat{\alpha} = \bar{x}' \). Hence \( x' = \bar{x}' + m_1 - 0.5 \) and the test statistic based on residuals is \( Z_1 = \sum (x' - \bar{x}' - (m_1 - 0.5))^2 \). \( Z_1 \) has similar properties to the Watson \( U^2 \) statistic

\[
U^2 = \sum (x' - \bar{x} - (2i-1)/(2n) - 0.5)^2 + 1/(12n),
\]

and has the same asymptotic distribution.

**Case 2.** For Cases 2 and 3 the situation becomes much harder, and considerable analysis is required to obtain the asymptotic distributions of the test statistic \( nE(x - \hat{x}')^2 / \Sigma(x - \bar{x})^2 \), the denominators being necessary, and complicating the analysis, because for these two cases the scale must be estimated. We state the results and give proofs later. For Case 2, the model is \( E(x) = a + \beta m_i \), with \( \beta \) unknown and \( a \) known. Set \( x' = x - a \), so that \( E(x') = \beta m_i \), and estimate \( \beta \) by least squares; then

\[
\hat{\beta} = \Sigma x'_i m_i / \Sigma m_i^2.
\]

Thus \( \hat{x} = a + \hat{\beta} m_i \), and the test statistic is
\[ z_2 = n \sum (x_{(i)} - \hat{x}_{(i)})^2 / \sum (x_{(i)} - \bar{x})^2. \] 
\( z_2 \) has the same asymptotic distribution as \( z_2^* = \Sigma v_i / \lambda_i \) where \( v_i \), for \( i = 1, 2, \ldots \), are independent \( \chi_1^2 \) variables; \( \lambda_i \) is an infinite set of positive weights given by \( \lambda_i = \theta_i^2 \), where \( \theta_i \) are the solutions of \( \tan \theta_i = \theta_i \), \( \theta_i > 0 \) (see Section 3 below).

**Case 3.** For Case 3, the model is \( E(x_{(i)}) = \alpha + \beta (m_i - 0.5) \), with \( \alpha, \beta \) unknown, and least squares gives \( \hat{\alpha} = \bar{x} \) and \( \hat{\beta} = \Sigma [(x_{(i)} - \bar{x})m_i] / \Sigma (m_i - \bar{m})^2 \). The test statistic is now the correlation coefficient \( R(X,m) \) or equivalently \( z_3 = Z(X,m) \);

\[ z_3 = n(1 - R^2(X,m)) = n \sum (x_{(i)} - \hat{x}_{(i)})^2 / \sum (x_{(i)} - \bar{x})^2, \] 
where \( \hat{x}_{(i)} = \hat{\alpha} + \hat{\beta}(m_i - 0.5) \).

\( z_3 \) has the same asymptotic distribution as \( z_3^* = \Sigma v_i / \lambda_i \), where, as above, \( v_i \) are independent \( \chi_1^2 \) variables. The constants \( \lambda_i \) are positive weights given in two infinite sets:

Set 1: \( \lambda_i = 4 \pi^2 i^2, \ i = 1, 2, \ldots \).

Set 2: \( \lambda_k = 4 \phi_k^2, \ k = 1, 2, \ldots \), where \( \phi_k \) are the solutions of \( \tan \phi_k = \phi_k \), \( \phi_k > 0 \).

The derivation of the weights for Cases 2 and 3 will be given in the next section.

**3. ASYMPTOTIC PROPERTIES OF Z(X,m).**

**3.1. Case 3.** It is convenient to give the asymptotic results for Case 3 (the more difficult case) first. Suppose, without loss of generality, that the sample comes from \( U(0,1) \). However, the model is fitted without
this knowledge: thus the fitted model is
\[ x_{(i)} = \alpha + \beta (m_i - \bar{m}) + \epsilon_i. \] (6)

As stated in Section 2, this leads to the test statistic
\[ Z(X, m) = n(1 - R^2(X, m)) = \sum(x_{(i)} - \hat{X}_{(i)})^2 / \{ \sum(x_i - \bar{x})^2 / n \}. \]

Asymptotically, the denominator tends to 1/12; thus we must study
\[ x_{(i)} - \hat{x}_{(i)}. \]

This may be written
\[
x_{(i)} - \hat{\alpha} - \hat{\beta}(m_i - \bar{m}) = x_{(i)} - \bar{x} - (\hat{\beta} - 1)(m_i - \bar{m}) - (m_i - \bar{m})
\]
\[ = x_{(i)} - m_i - (\bar{x} - \bar{m}) - (\hat{\beta} - 1)(m_i - \bar{m}). \] (7)

The terms on the right hand side of (7) can be expressed in
terms of the quantile process \( Q_n(t) = X_{[nt]} - m_{[nt]}, 0 \leq t \leq 1 \), where
\([nt]\) is the greatest integer in \( nt \). For \( t \) given by \( i/n \), we have
\[ x_{(i)} - m_i = Q_n(t); \]
\[ \sqrt{n}(\bar{x} - \bar{m}) = \int_0^1 Q_n(s) ds + O_p(n^{-1/2}); \]
\[ \sqrt{n}(\hat{\beta} - 1) = n^{-1/2} \sum(x_{(i)} - \bar{x} - m_i + \bar{m})(m_i - \bar{m}) / \Sigma(m_i - \bar{m})^2 / n \]
\[ = 12 \int_0^1 (t - 1/2)Q_n(t) dt - \int_0^1 Q_n(s) ds dt + O_p(n^{-1/2}), \]
recalling that \( \bar{m} = \frac{1}{n} \) and \( \Sigma(m_i - \bar{m})^2 / n \rightarrow 1/12 \). It is convenient
to define the process \( Y_n(t) = Q_n(t) - \int_0^t Q_n(s) ds \). Then insertion
of the above expressions into (7) gives
\[ X(i) - \hat{X}(i) = Y_n(t) - \int_0^1 \int_0^{(u - \frac{1}{2})} 12(t - \frac{1}{2})Y_n(t) \, dt \, du + O_p(n^{-1}). \] (8)

As \( n \to \infty \), let \( Q(t), Y(t) \) be the limiting processes for \( Q_n(t) \) and \( Y_n(t) \) respectively. \( Q(t) \) is the well-known Brownian bridge with mean \( \mathbb{E}[Q(t)] = 0 \) and covariance \( \rho_0(s,t) = \min(s,t) - s \cdot t \). \( Y(t) \) then has mean 0 and covariance \( \rho_y(s,t) = \min(s,t) - \frac{1}{2}s(l - s) - \frac{1}{2}t(1-t) + 1/12. \) The process \( Y(t) \) has already been studied in connection with the Watson statistic \( U^2 \) (Watson, 1961; Stephens, 1976). For the asymptotic distribution of \( Z(X,m) \) we now need the distribution of

\[ Z^* = \int_0^1 W^2(t) \, dt, \] (9)

where, from (8), we have

\[ W(t) = Y(t) - \int_0^1 (u - \frac{1}{2}) \int_0^{12(t - \frac{1}{2})} Y(t) \, dt \, du. \] (10)

The covariance function of \( W(t) \) requires considerable algebra but the calculation is straightforward; the result may be expressed as

\[ \rho_w(s,t) = \rho_0(s,t) - \psi(s)' \Lambda \psi(t), \] (11)

where \( \psi(s)' \) is the transpose of \( \psi(s) \), and is the 2-component vector \( \{(s - \frac{1}{2}); s(1-s)(2s-1)\} \); \( \Lambda \) is the 2 x 2 matrix with rows \( (-\frac{1}{5}, 1) \) and \( (1,0) \). The calculation of the distribution of \( Z^* \) now follows well-known lines (see, for example, Durbin, 1973, or Stephens, 1976); \( Z^* \) has the same distribution as \( S = \sum_i v_i / \lambda_i \), where \( i \) runs from...
1 to \( \infty \), \( v_i \) are independent \( \chi_1^2 \) variables, and where \( \lambda_i \) are weights, found by solving the integral equation

\[
\frac{1}{\lambda} \int_0^1 f(s) \rho_w(s,t) \, ds = f(t) \tag{12}
\]

for eigenvalues \( \lambda_i \) and eigenfunctions \( f_i(t) \).

The solution of (12) is found as follows. The covariance \( \rho_w(s,t) \)
can be expressed as \( \rho_w(s,t) = \min(s,t) + g(s,t) \), with

\[
g(s,t) = \frac{6}{5} st - \frac{11}{10} s^2 + 2s^2 - s^3 - \frac{11}{10} t^2 + 2t^2 - t^3 + \frac{2}{15} - 3st^2 + 2st^3 - 3s^2t + 2s^3t.
\]

Differentiation of (12) twice with respect to \( t \) then gives

\[
-\frac{1}{\lambda} f''(t) + \frac{1}{\lambda} \int_0^1 f(s) ds - 6t \int_0^1 f(s) ds - \frac{1}{\lambda} \int_0^1 sf(s) ds + 12t \int_0^1 sf(s) ds = \frac{1}{\lambda} f'''(t) \tag{13}
\]

Differentiation again gives

\[
-\frac{1}{\lambda} f'(t) - 6 \int_0^1 f(s) ds + 12 \int_0^1 sf(s) ds = \frac{1}{\lambda} f'''(t) \tag{14}
\]

and finally

\[
- f''(t) = \frac{1}{\lambda} f'''(t) .
\]

Thus \( f(t) = A \cos \sqrt{\lambda} t + B \sin \sqrt{\lambda} t + Ct + D \).

Suppose \( f(s) \) is normalized, so that \( \int_0^1 f(s) ds = 1 \), and let \( K = \int_0^1 sf(s) ds \).

Set \( \theta = \sqrt{\lambda} \); then \( \int_0^1 f(s) ds = 1 \) gives

\[
\frac{A}{\theta} \sin \theta - \frac{B}{\theta} (\cos \theta - 1) + \frac{C}{2} + D = 1 \tag{16}
\]
and

\[ K = \int_0^1 s f(s) ds = A I_1 + B I_2 + \frac{C}{3} + \frac{D}{2}, \tag{17} \]

where

\[ I_1 = \int_0^1 s \cos \theta s ds = \frac{\theta \sin \theta + \cos \theta - 1}{\theta^2} \]

\[ I_2 = \int_0^1 s \sin \theta s ds = \frac{\sin \theta - \theta \cos \theta}{\theta^2}. \]

Substituting \( f(t) \) into (13) gives \(-C - D + 6t - 6K + 12Kt = 0\)

for all \( t \); thus, equating coefficients, we have

\[-C - 6 + 12K = 0 \quad \text{and} \quad -D + 4 - 6K = 0.\]

Hence \( \frac{C}{3} + \frac{D}{2} = K \), and \( C + 2D = 2 \).

Thus from (16) we have \( A \sin \theta - B(\cos \theta - 1) = 0 \), and from (17) we have

\[ AI_1 + BI_2 = 0. \] Hence \( \theta \) must satisfy

\[ \frac{\sin \theta}{\cos \theta - 1} = \frac{B}{A} = \frac{I_1}{I_2} = \frac{1 - \theta \sin \theta - \cos \theta}{\sin \theta - \theta \cos \theta}, \tag{18} \]

So \( \theta \) satisfies \( 2 - \theta \sin \theta - 2 \cos \theta = 0 \), by cross-multiplication of (18). Let \( \phi = \frac{\theta}{2} \); then \( 2 - 4\phi \sin \phi \cos \phi - 2[1 - 2 \sin^2 \phi] = 0 \), and hence

\[ \sin \phi = 0 \text{ or } \sin \phi - \phi \cos \phi = 0. \]

Then \( \phi_i = \pi i, i = 1, 2, \ldots \); or alternatively \( \phi_k \) is the solution of \( \tan \phi_k = \phi_k, k = 1, 2, \ldots \). Finally, \( \lambda_i = 4\phi_i^2 \), for the first \( \lambda \)-set, and \( \lambda_k = 4\phi_k^2 \), for the second \( \lambda \)-set.
3.2 Case 2. For Case 2 the test statistic is
\[ n\widehat{E}(X_{(i)} - \widehat{X})^2 / \widehat{E}(X_{(i)} - \overline{X})^2 = z_2. \]
We can take \( a = 0 \) in the model \( E(X_{(i)}) = a + \beta m_i \), so that this becomes \( E(X_{(i)}) = \beta m_i \), with
\[ \hat{\beta} = \frac{\sum X_{(i)} m_i}{\sum m_i^2}. \]
Hence \( \hat{\beta} - 1 = \frac{\sum (X_{(i)} - m_i)m_i}{\sum m_i^2} \). Similar reasoning to that for Case 3 gives the asymptotic distribution of \( z_2 \) to be that of
\[ \frac{1}{12} \int_0^1 W_2(t) dt \]
where
\[ W_2(t) = Q(t) - 3t \int_0^1 sQ(s) ds. \] (19)
\( Q(t) \) is as defined in the previous section, and then \( W_2(t) \) is a Gaussian process with mean 0; its covariance function (after some algebra) is
\[ \rho_2(s,t) = \min(s,t) - \frac{14}{5} st + \frac{s^3}{2} + \frac{s^3 t}{2}. \] (20)
Thus for the weights in the asymptotic distribution of \( z_2 \), we need eigenvalues of \( \lambda \int_0^1 \rho_2(s,t) f(s) ds = f(t) \). Similar steps to those for Case 3 give \( f(t) = A \cos \theta t + B \sin \theta t + Ct + D \) with \( \theta = \sqrt{\lambda} \), as before.
Also, \( f(0) \equiv 0 \), so \( D = -A \), and
\[ -f(t) = 3t \int_0^t sf(s) ds = \frac{f''(t)}{\lambda}. \] (21)
Thus \( f''(0) = 0 \), so \( D = A = 0 \). Then, from (21), we have
\[ -B \sin \theta t - Ct + 3t \left[ B \int_0^t s \sin \theta s ds + \int_0^1 Cs^2 ds \right] \equiv -B \sin \theta t. \]
Hence \( \int_0^1 s \sin \theta s \, ds = 0 \); thus \( \theta_j \) is the solution of
\[
\sin \theta_j - \theta_j \cos \theta_j = 0,
\]
that is, \( \tan \theta_j = \theta_j, \quad j = 1, 2, \ldots \). Finally,
\( \lambda_j = \theta_j^2 \). These are the weights given in Section 2.

### 3.3 Asymptotic percentage points

The next step is to calculate the percentage points of, say, \( Z_3^* = \Sigma v_i / \lambda_i \) where \( \lambda_i \) are the weights for Case 3. The mean \( \mu_3 \) of \( Z_3 \) is \( \int_0^1 \rho_3(s,s) \, ds = 1/15 \). The 80 largest \( \lambda_i \) were found, and \( Z_3^* \) was approximated by \( S_1 = S^* + T \), where
\[
S^* = \Sigma_{i=1}^{80} v_i / \lambda_i \quad \text{and} \quad T = \mu_3 - \Sigma_{i=1}^{80} \lambda_i^{-1}.
\]
\( S_1 \) differs from \( Z_3^* \) by \( \lambda_{81}^{-1} (v_i - 1) \) which is a random variable with mean 0 and variance
\[
2 \Sigma_{i=1}^{81} \lambda_i^{-2} = 2 \int_0^1 \rho_3(s,t) \, ds \, dt - \Sigma_{i=1}^{80} \lambda_i^{-2};
\]
this value is found to be negligibly small. Thus critical points of \( Z_j^* \) are found by finding those of \( S^* \), using Imhof's (1961) method for a finite sum of weighted \( \chi^2 \) variables, and then adding \( T \).

### 3.4 Tables

Tables 1 and 2 give percentage points for \( Z_2 \) and \( Z_3 \) respectively. Those for \( n \) finite have been obtained by Monte Carlo sampling. The last line in each table contains the asymptotic points. Table 1 also gives points for a modification of \( Z_2 \), called \( Z_{2A} \).

This is the statistic (using the terminology for Case 2 in Section 2)
\[
Z_{2A} = nE[X(i) - \hat{X}(i)]^2 / \Sigma X(i)^2 \quad \text{.}
\]
This uses the quantity \( \Sigma X(i)^2 / n \) to eliminate the square of the scale instead of the sample variance. This is a natural denominator in Case 2 with \( a = 0 \), where the model is \( E(X(i)) = \beta m_i \). (If \( a \) is
not zero, the new variable \( X'_{(i)} = X_{(i)} - a \) should be used instead of \( X_{(i)} \).

The asymptotic points for \( Z_{2A} \) are 0.25 times those for \( Z_2 \). An advantage in using \( Z_{2A} \) is that the statistic is much less variable for small \( n \). For \( Z_3 \), Table 2 has already been produced in Stephens (1986), although with less accurate points; there will be negligible difference in practical use.

3.5 Use of the Tables with censored data. Suppose origin and scale are both unknown (Case 3), and the data is censored at both ends. Thus

\[ n^* = r - k + 1 \]

observations are available, consisting of all those between \( X_{(k)} \) and \( X_{(r)} \). \( R(X,T) \) may be calculated, using the usual formula, but with sums for \( i \) from \( k \) to \( r \), and with \( T_i = i/(n+1) \) or \( T_i = i \), or even \( T_1, T_2, \ldots \) equal to \( 1, 2, \ldots, n^* \), these latter values for \( T_i \) being possibilities because \( R(X,m) \) is scale and location invariant. Then

\[ n^* \{ 1 - R^2(X,T) \} = Z(X,T) \]

will be referred to Table 3, using the values for sample size \( n^* \).

3.6 Example. It is well-known that if times \( Q_i; i = 1, 2, \ldots, n \) represent times of random events, occurring in order with the same rate, the \( Q(i) \) should be proportional to uniform order statistics \( U_{(i)} \). Thus the \( Q(i) \) may be regressed against \( i/(n+1) \) or equivalently against \( i \) as described above, to test that the events are random. Suppose

\( Q(9), Q(10), \ldots, Q(20) \) represent a subset of such times, denoting times of breakdown of an industrial process. We wish to test that these are uniform; times \( Q(1) \) to \( Q(8) \) have been omitted because the process took time to stabilize and these are not expected to have occurred at the same rate as the later times. The times \( Q(9), \ldots, Q(20) \) are 82, 93, 120, 135, 137, 142, 162, 163, 210, 228, 233, 261. The value of

\[ Z(Q,T) = 12 \{ 1 - R^2(Q,T) \} \]
= 0.464. Reference to Table 2 at line \( n = 12 \) show that there is not significant evidence, at the 10% level, to reject the hypothesis of uniformity.
4. THE CORRELATION COEFFICIENT: GENERAL CASE

4.1 The general case. We now discuss, in a non-rigorous fashion, the
distribution of \( Z(X,m) \) for the general test of \( H_0 \) given in (5). \( F_o(x) \)
is assumed to be a continuous distribution, and the sample can be left-
and right-censored. Thus we observe \( X_{(k)} < \ldots < X_{(r)} \) from a sample
of size \( n \) from the distribution \( F_o(x) \). We can assume the sample comes
from \( F_o(x) \) with \( \alpha = 0 \) and \( \beta = 1 \), that is, from \( F(*) \) although (3) is
fitted without this knowledge. Suppose \( f(x) \) is the density corresponding
to \( F(x) \). Then using \( H_{(i)} = F^{-1}\left(\frac{i}{n+1}\right) \) we have

\[
Z(X,m) = n\{1 - R^2(X,m)\} = \frac{\sum_k^r (X_{i,k} - H_i - \hat{\alpha} - (\hat{\beta} - 1))^2}{\frac{1}{n} \sum_k^r (H_i - \bar{H})^2}.
\]

Define \( p = (k-1)/n \) and \( q = r/n \) and let \( q^* = F^{-1}(q) \) and \( p^* = F^{-1}(p) \).

Also, let

\[
\gamma(t) = Q(t) - \int_p^q \mathcal{Q}(s) ds - \left\{ \frac{F^{-1}(t) - \mu}{\sigma} \right\} \int_p^q \left\{ \frac{F^{-1}(s) - \mu}{\sigma} \right\} \mathcal{Q}(s) ds,
\]

where

\[
Q(t) = \sqrt{n} \left\{ X_{[nt]} - F^{-1}(t) \right\}, \text{ and parameters } \mu \text{ and } \sigma \text{ are given by}
\]

\[
\mu = \int_p^{p^*} \frac{F^{-1}(s)}{\sigma} ds = \int_p^{p^*} xf(x) dx
\]

and
\[ \sigma^2 = \int_{p}^{q} (F^{-1}(s))^2 ds - \mu^2 = \int_{p}^{q^*} x^2 f(s)dx - \mu^2. \]

The process \( Q(t) \) is close to a Gaussian process with mean 0 and covariance

\[ \rho_Q(s, t) = \frac{\min(s, t)}{f(F^{-1}(s))f(F^{-1}(t))}. \]

The process \( Y(t) \) is then close to a Gaussian process with mean 0 and covariance

\[ \rho(s, t) = \rho_Q(s, t) - \psi(s) \int_{p}^{q} \rho_Q(u, t)du - \psi(t) \int_{p}^{q} \rho_Q(s, u)du \]

\[ - \int_{p}^{q} \rho_Q(u, t)du - \int_{p}^{q} \rho_Q(s, u)du + \int_{p}^{q} \int_{p}^{q} \rho_Q(u, v)dvdu \]

\[ + \psi(s) \psi(t) \int_{p}^{q} \int_{p}^{q} \rho_Q(u, v) \psi(v)dvdu + (\psi(s) + \psi(t)) \int_{p}^{q} \int_{p}^{q} \rho_Q(u, v) \psi(u)dvdu \]

where \( \psi(s) = \frac{F^{-1}(s) - \mu}{\sigma} \).

The denominator of \( Z = n(1 - R^2) \), where we write \( Z \) for \( Z(X, m) \) and \( R^2 \) for \( R^2(X, m) \), i.e. then close to \( \sigma^2 \), and the numerator is close to \( T = \int_{p}^{q} Y^2(t)dt. \) Thus the asymptotic theory now depends on the behaviour of \( T \). It appears generally...
that this behaviour is determined by that of \( \int_0^q Q^2(t) dt \). There are 3 cases in practice, which we label Cases A, B and C. Define

\[
J_1 = \int_0^q \int_0^q \rho_0^2(s,t) ds dt
\]

and

\[
J_2 = \int_0^q \rho_0(t,t) dt.
\]

**Case A.** In this case suppose \( J_1 < \infty \) and \( J_2 < \infty \). Then we have

\[
Z = n(1-R^2) = \frac{1}{\sigma^2} \sum_1^\infty \frac{v_i}{\lambda_i}
\]

where \( v_i \) are independent \( \chi^2 \) variables and \( \lambda_i \) are eigenvalues of

\[
f(s) = \lambda \int_0^q \rho(s,t)f(t) dt.
\]

(The sum \( \sum_1^{-1} \lambda_i \) will be < \( \infty \)).

**Case B.** Suppose \( J_1 < \infty \) but \( J_2 = \infty \). Then there exists \( a_n \rightarrow \infty \) such that

\[
Z - a_n = n(1-R^2) - a_n = \frac{1}{\sigma^2} \sum_1^\infty \frac{\lambda_i(v_i - 1)}{\lambda_i},
\]

where the \( \lambda_i \) and \( v_i \) are as defined above. (In this case \( \sum_1^{-1} \lambda_i = \infty \)).

**Case C.** For this case suppose both integrals \( J_1 \) and \( J_2 \) are infinite.

Then there exist constants \( a_n, b_n \), such that

\[
\frac{Z - a_n}{b_n} = \frac{n(1-R^2) - a_n}{b_n} = N(0,1).
\]
4.3 Examples.

1. The exponential distribution.

For \( q = 1 \) we have case C; \( a_n = \log n \), and \( b_n = (\log n)^{1/2} \), so that

\[
\frac{n(1 - R^2) - \log n}{2\sqrt{\log n}} \sim N(0,1).
\]

For \( q < 1 \) we are in Case A and the distribution is a sum of weighted chi-squared variables.

2. The uniform test (discussed above).

For any \( p \) or \( q \) Case A applies and \((r - k+1)(1 - R^2)\) has the same limiting distribution regardless of \( p, q \).

3. The normal test.

For \( p = 0 \) or \( q = 1 \) or both we get Case B.

For \( p > 0, q < 1 \) we get Case A.

4. The Logistic test: \( F(w) = 1/(1 + e^{-w}), -\infty < w < \infty \).

For \( p = 0 \) or \( q = 1 \) or both we get Case C.

For \( p > 0 \) and \( q < 1 \), we get Case A. The logistic test is thus similar to the exponential test.

5. Test for the Extreme Value distribution I: \( F(w) = 1 - e^{-e^w}, -\infty < w < \infty \).

For \( p = 0 \), we get Case C.

For \( p > 0 \) we get Case A.
6. **Test for the Extreme Value distribution II:** \( F(w) = e^{-e^{-w}}, -\infty < w < \infty \).

For \( q = 1 \), we get Case C.

For \( q < 1 \), we get Case A.

4.4 **Discussion.** The discussion above is somewhat imprecise. When \( p \) is 0 or \( q \) is 1 there are technical details which have been glossed over.

For the distributions we have studied however the criteria given in Cases A, B and C lead to the correct answer for asymptotic distributions of \( Z(X,m) = n(1 - R^2(X,m)) \).
Table 1. Critical Points for $Z_2$ and $Z_{2A}$.

Upper tail significance level (percent)

<table>
<thead>
<tr>
<th>n</th>
<th>50</th>
<th>25'</th>
<th>15</th>
<th>10</th>
<th>5</th>
<th>2.5</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.690</td>
<td>1.240</td>
<td>1.94</td>
<td>3.47</td>
<td>8.67</td>
<td>20.3</td>
<td>47.0</td>
</tr>
<tr>
<td>6</td>
<td>0.763</td>
<td>1.323</td>
<td>1.89</td>
<td>2.59</td>
<td>4.74</td>
<td>8.49</td>
<td>17.0</td>
</tr>
<tr>
<td>8</td>
<td>0.806</td>
<td>1.364</td>
<td>1.85</td>
<td>2.37</td>
<td>3.78</td>
<td>6.29</td>
<td>11.4</td>
</tr>
<tr>
<td>10</td>
<td>0.832</td>
<td>1.388</td>
<td>1.88</td>
<td>2.34</td>
<td>3.40</td>
<td>5.30</td>
<td>8.9</td>
</tr>
<tr>
<td>12</td>
<td>0.848</td>
<td>1.407</td>
<td>1.89</td>
<td>2.33</td>
<td>3.27</td>
<td>4.80</td>
<td>7.8</td>
</tr>
<tr>
<td>18</td>
<td>0.877</td>
<td>1.438</td>
<td>1.91</td>
<td>2.32</td>
<td>3.12</td>
<td>4.26</td>
<td>6.3</td>
</tr>
<tr>
<td>20</td>
<td>0.881</td>
<td>1.444</td>
<td>1.92</td>
<td>2.32</td>
<td>3.10</td>
<td>4.18</td>
<td>6.0</td>
</tr>
<tr>
<td>40</td>
<td>0.907</td>
<td>1.470</td>
<td>1.93</td>
<td>2.32</td>
<td>3.03</td>
<td>3.82</td>
<td>5.1</td>
</tr>
<tr>
<td>60</td>
<td>0.916</td>
<td>1.480</td>
<td>1.93</td>
<td>2.32</td>
<td>3.00</td>
<td>3.73</td>
<td>4.9</td>
</tr>
<tr>
<td>80</td>
<td>0.920</td>
<td>1.485</td>
<td>1.94</td>
<td>2.32</td>
<td>2.99</td>
<td>3.71</td>
<td>4.9</td>
</tr>
<tr>
<td>100</td>
<td>0.922</td>
<td>1.488</td>
<td>1.94</td>
<td>2.32</td>
<td>2.98</td>
<td>3.70</td>
<td>4.8</td>
</tr>
<tr>
<td>$\infty$</td>
<td>0.932</td>
<td>1.497</td>
<td>1.94</td>
<td>2.31</td>
<td>2.98</td>
<td>3.67</td>
<td>4.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>0.140</th>
<th>0.245</th>
<th>0.333</th>
<th>0.411</th>
<th>0.545</th>
<th>0.707</th>
<th>1.010</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.166</td>
<td>0.287</td>
<td>0.379</td>
<td>0.467</td>
<td>0.616</td>
<td>0.796</td>
<td>1.065</td>
</tr>
<tr>
<td>8</td>
<td>0.184</td>
<td>0.307</td>
<td>0.403</td>
<td>0.494</td>
<td>0.648</td>
<td>0.830</td>
<td>1.089</td>
</tr>
<tr>
<td>10</td>
<td>0.193</td>
<td>0.320</td>
<td>0.420</td>
<td>0.512</td>
<td>0.670</td>
<td>0.848</td>
<td>1.102</td>
</tr>
<tr>
<td>12</td>
<td>0.200</td>
<td>0.330</td>
<td>0.432</td>
<td>0.523</td>
<td>0.683</td>
<td>0.861</td>
<td>1.111</td>
</tr>
<tr>
<td>18</td>
<td>0.209</td>
<td>0.346</td>
<td>0.452</td>
<td>0.543</td>
<td>0.705</td>
<td>0.882</td>
<td>1.121</td>
</tr>
<tr>
<td>20</td>
<td>0.212</td>
<td>0.349</td>
<td>0.455</td>
<td>0.547</td>
<td>0.708</td>
<td>0.886</td>
<td>1.124</td>
</tr>
<tr>
<td>40</td>
<td>0.224</td>
<td>0.362</td>
<td>0.472</td>
<td>0.563</td>
<td>0.727</td>
<td>0.903</td>
<td>1.138</td>
</tr>
<tr>
<td>60</td>
<td>0.228</td>
<td>0.367</td>
<td>0.477</td>
<td>0.568</td>
<td>0.734</td>
<td>0.909</td>
<td>1.146</td>
</tr>
<tr>
<td>80</td>
<td>0.229</td>
<td>0.369</td>
<td>0.479</td>
<td>0.570</td>
<td>0.736</td>
<td>0.911</td>
<td>1.149</td>
</tr>
<tr>
<td>100</td>
<td>0.230</td>
<td>0.370</td>
<td>0.480</td>
<td>0.572</td>
<td>0.737</td>
<td>0.912</td>
<td>1.150</td>
</tr>
<tr>
<td>$\infty$</td>
<td>0.233</td>
<td>0.374</td>
<td>0.485</td>
<td>0.578</td>
<td>0.744</td>
<td>0.917</td>
<td>1.155</td>
</tr>
</tbody>
</table>
Table 2. Critical Points for $Z_3$.

<table>
<thead>
<tr>
<th>n</th>
<th>0.5</th>
<th>0.25</th>
<th>0.15</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.344</td>
<td>0.559</td>
<td>0.734</td>
<td>0.888</td>
<td>1.089</td>
<td>1.238</td>
<td>1.388</td>
</tr>
<tr>
<td>6</td>
<td>0.441</td>
<td>0.703</td>
<td>0.901</td>
<td>1.053</td>
<td>1.325</td>
<td>1.590</td>
<td>1.918</td>
</tr>
<tr>
<td>8</td>
<td>0.495</td>
<td>0.792</td>
<td>1.000</td>
<td>1.163</td>
<td>1.474</td>
<td>1.739</td>
<td>2.100</td>
</tr>
<tr>
<td>10</td>
<td>0.535</td>
<td>0.833</td>
<td>1.068</td>
<td>1.245</td>
<td>1.532</td>
<td>1.846</td>
<td>2.294</td>
</tr>
<tr>
<td>12</td>
<td>0.560</td>
<td>0.864</td>
<td>1.093</td>
<td>1.280</td>
<td>1.608</td>
<td>1.918</td>
<td>2.360</td>
</tr>
<tr>
<td>18</td>
<td>0.605</td>
<td>0.940</td>
<td>1.147</td>
<td>1.348</td>
<td>1.672</td>
<td>2.008</td>
<td>2.503</td>
</tr>
<tr>
<td>20</td>
<td>0.610</td>
<td>0.960</td>
<td>1.200</td>
<td>1.370</td>
<td>1.680</td>
<td>2.025</td>
<td>2.520</td>
</tr>
<tr>
<td>40</td>
<td>0.640</td>
<td>0.980</td>
<td>1.215</td>
<td>1.396</td>
<td>1.732</td>
<td>2.076</td>
<td>2.580</td>
</tr>
<tr>
<td>60</td>
<td>0.648</td>
<td>0.988</td>
<td>1.227</td>
<td>1.410</td>
<td>1.750</td>
<td>2.092</td>
<td>2.590</td>
</tr>
<tr>
<td>80</td>
<td>0.658</td>
<td>0.997</td>
<td>1.228</td>
<td>1.418</td>
<td>1.760</td>
<td>2.104</td>
<td>2.610</td>
</tr>
<tr>
<td>$\infty$</td>
<td>0.666</td>
<td>0.992</td>
<td>1.234</td>
<td>1.430</td>
<td>1.774</td>
<td>2.129</td>
<td>2.612</td>
</tr>
</tbody>
</table>
References


**Tests Of Fit Based On The Correlation Coefficient**

**Author(s)**
Richard A. Lockhart and M. A. Stephens

**Performing Organization Name and Address**
Department of Statistics
Stanford University
Stanford, CA 94305

**Controlling Office Name and Address**
Office of Naval Research
Statistics & Probability Program Code 1111

**Report Date**
October 4, 1990

**Number of Pages**
27

**Security Class. (of this report)**
UNCLASSIFIED

**Distribution Statement (of this report)**
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

**Key Words**
Goodness-of-fit, Correlation coefficient, Correlation tests.

**Abstract**
PLEASE SEE FOLLOWING PAGE.
20. ABSTRACT

In testing fit of a sample to a specific family of distributions \( F \), probability-plots use a plot of the order statistics of the sample against a set of suitable constants. These are chosen so that if the parent population is one of the family \( F \), the resulting plot should appear like a straight line. Members of \( F \) may contain unknown location and/or scale parameters. Historically, the judgment of a straight line fit was usually made by eye. In this article we propose the correlation coefficient \( R \) between the sample values and the constants as a test statistic. Asymptotic properties of \( R \) are derived in general, and tables produced to make the test for the case where the proposed distribution is uniform.