TECHNICAL REPORT NO. 405

January 1975

ON THE ASYMPTOTIC DISTRIBUTION OF
SOME GOODNESS OF FIT TESTS BASED
ON SPACINGS

by

Guido E. Del Pino

University of Wisconsin

Typist: Candy Smith

This report was supported by the Air Force Office of Scientific
Research under Grant No. AFOSR 72-2363B.
Abstract

Let $X_1, \ldots, X_{n-1}$ be independent and identically distributed random variables with common distribution $F$ on $[0,1]$ and let $X'_1, \ldots, X'_{n-1}$ be the corresponding ordered observations. The problem of testing the hypothesis that $F$ is uniform in $[0,1]$ based on the statistic $T_n = \frac{1}{N} \sum_{i=1}^{N} m[n(X'_{ik} - X'_{(i-1)k})]$ is considered, where $k$ is a fixed number independent of $n$, $m$ is a function satisfying some regularity conditions and $X'_0 = 0$, $X'_n = 1$. The asymptotic distribution of $T_n$ under a sequence of local alternatives is obtained and it is used to compute Pitman efficiencies of tests based on different functions $m$. This generalizes results of Rao and Sethuraman [in Nonparametric Techniques in Statistical Inference, Ed. M.L. Puri] who only study the case $k = 1$. Examples are given where, for a given $m$, the Pitman efficiency increases arbitrarily with $k$.


Key words and phrases: Goodness of fit tests, spacings, weak convergence, Pitman efficiency.
On the asymptotic distribution of some goodness of fit
tests based on spacings

Guido E. del Pino

1. Introduction

Let \( X_1, X_2, \ldots, X_{(n-1)} \) be (n-1) independent and identically
distributed random variables with common distribution
\( A_n \) on \([0,1]\) and let \( X'_1, X'_2, \ldots, X'_{(n-1)} \) be the corresponding
order statistics. For any fixed \( k \) the sample spacings
\( D_1, \ldots, D_n \) are defined as

\[
D_i = X'_{ik} - X'_{(i-1)k} \quad i = 1, 2, \ldots, n
\]  

(1.1)

where we assume \( n = Nk \) and take \( X'_0 = 0 \) and \( X'_n = 1 \).
For the case \( k = 1 \), Rao and Sethuraman ([8], [9]) have
studied the asymptotic distribution of

\[
T_n = \frac{1}{N} \sum_{i=1}^{N} m(nD_i)
\]  

(1.2)

under the sequence

\[
A_n(x) = x + \frac{L_n(x)}{n} \delta = x + \left( \frac{1}{k \delta} L_n(x) \right) / N^\delta \quad 0 \leq x \leq 1
\]  

(1.3)

where \( L_n(0) = L_n(1) = 0 \) and \( \delta \leq \frac{1}{4} \). It is also assumed that
\( L_n \) is twice differentiable in \([0,1]\) and there is a function
\( L \) which is twice continuously differentiable and such that
\[ L(0) = L(1) = 0 \]
\[ n^{\delta^*} \sup_{0 \leq x \leq 1} |L_n(x) - L(x)| = o(1) \]
\[ n^{\delta^*} \sup_{0 \leq x \leq 1} |L'_n(x) - \ell(x)| = o(1) \quad (1.4) \]
and
\[ n^{\delta^*} \sup_{0 \leq x \leq 1} |L''_n(x) - \ell'(x)| = o(1) \]

where \( \ell \) and \( \ell' \) are the first and second derivatives of \( L \) and \( \delta^* = \max(0, \frac{1}{2} - \delta) \). The function \( m(x) \) must satisfy some regularity conditions which are analyzed in the next section.

The statistic \( T_n \) can be used to perform a goodness of fit test for the uniform distribution on \([0,1]\) (which corresponds to \( L_n = 0 \) and will be denoted by \( U \)). Knowledge of the asymptotic distribution of \( T_n \) under the sequence of alternatives \( A_n \) allows us to compute Pitman's Asymptotic Relative Efficiencies (A.R.E.) for tests based on different functions \( m \).

The object of this note is to extend the above to the case of any fixed \( k \) and show that for some functions \( m \) the efficiency can be greatly improved by taking a large value of \( k \).

2. Asymptotic distribution of \( T_n \)

Let \( Z_1, \ldots, Z_n \) be independent and identically distributed random variables with density

\[ f_k(z) = \frac{1}{\Gamma(k)} z^{k-1} e^{-z} \]
and corresponding distribution function $H_k$. It is well known that under $U$
\[
(nD_i, i=1, \ldots, N) \overset{d}{=} \{\frac{Z_i}{\tilde{Z}_n}, i=1, \ldots, N\}
\]
where
\[
\tilde{Z}_n = \frac{1}{n} \sum_{i=1}^{N} Z_i = \frac{1}{k} \left( \frac{1}{N} \sum_{i=1}^{N} Z_i \right)
\]
d and $\equiv$ stands for "has the same distribution as".

We have
\[
\xi_n \equiv \sqrt{N}(\tilde{Z}_n - 1) \Rightarrow \xi \sim N(0, \frac{1}{k}) \quad (2.1)
\]
with $\Rightarrow$ denoting weak convergence (i.e., convergence in distribution in this case).

It is easy to verify that conditions for the applicability of Theorem 2.8 in [9] are satisfied. Thus we have obtained the following

**Theorem 2.1**

Let $\tilde{F}_n$ be the empirical distribution function (e.d.f.) of $(nD_i, i=1, \ldots, N)$ and let $G_n$ be defined as follows
\[
G_n(x) = \frac{1}{N} \sum_{i=1}^{N} H_k(\alpha_{Ni} x)
\]
with
\[
\alpha_{Ni} = 1 + \left( \frac{1}{k^{\delta}} L(\frac{i}{N}) \right) / N^{\delta} - \left( \frac{1}{k^{\delta}} L(\frac{i}{N}) \right) \left( \frac{1}{k^{\delta}} L'(\frac{i}{N}) \right) / N^{2\delta} \quad (2.3)
\]
Similarly let $F_n$ be the e.d.f. of $(Z_i, i=1, \ldots, N)$. 
Define

\[ \tilde{\eta}_n(x) = \sqrt{N} (\tilde{F}_n(x) - G_n(x)) \]  
\[ \eta_n(x) = \sqrt{N} (F_n(x) - H_k(x)) . \]  

Then

\[ \sup_{0 < X < \infty} |\tilde{\eta}_n(x) - \eta_n(x) - xf_k(x)\xi_n| \xrightarrow{P} 0 . \]  

Remarks

I) The smoothness of \( f_k \) allows us to replace \( G_n(x) \) defined in (2.2) by

\[ G_n(x) = H_k(x) + (xf_k(x) + \frac{x^2}{2f_k(x)})(\int_0^x f_k(x)dx) / n^{2\delta} \]  

which differs from the previous definition by terms smaller than \( n^{-2\delta} \) uniformly in \( x \).

II) The finite-dimensional distributions of \( ((\eta_n(x), 0 < X < \infty), \xi_n) \) converge to the finite-dimensional distributions of a Gaussian process \( ((\eta(x), 0 < X < \infty), \xi) \), where \( \eta(x) \) has mean function zero and covariance function

\[ K(x, y) = \min(H_k(x), H_k(y)) - H_k(x)H_k(y) \]

\( \xi \) is defined in (2.1) and

\[ \text{Cov} (\xi, \eta(x)) = \text{Cov}(\xi_n, \eta_n(x)) = \frac{1}{k} \int_0^x (t-k)f_k(t)dt . \]

Since we also have \( \eta_n \xrightarrow{\text{S}} \eta \) (in the Skorohod topology) and
\( \xi_n \Rightarrow \xi \), then \(((\eta_n(x), 0 \leq x \leq \infty), \xi_n)\) is tight ([1], p. 41).

Therefore

\[
((\eta_n(x), 0 \leq x \leq \infty), \xi_n) \Rightarrow ((\eta(x), 0 \leq x \leq \infty), \xi)
\]

and consequently

\[
(\eta_n(x) + x f_k(x) \xi_n, 0 \leq x \leq \infty) \Rightarrow (\eta(x) + x f_k(x) \xi, 0 \leq x \leq \infty). \tag{2.8}
\]

III) \( \text{Cov}(\xi, \eta(x)) = \frac{1}{\sqrt{k}} \int_0^x (t-k) f_k(t) \, dt = -\frac{1}{\sqrt{k}} f_k(x) \).

From this we see that

\[
\eta(x) + x f_k(x) \xi = \eta(x) - (\frac{1}{\sqrt{k}} x f_k(x)) \sqrt{k} \xi
\]

is the component of \( \eta(x) \) orthogonal to \( \xi \). As a consequence of this, \( \eta(x) + x f_k(x) \xi \) is independent of \( \xi \) and has the same distribution as that of \( \eta(x) \) conditioned on \( \xi = 0 \).

This is merely reflecting the fact that

\[
(Z_1/\tilde{Z}_n, i=1, \ldots, N) \overset{d}{=} (Z_1, i=1, \ldots, N | \tilde{Z}_n = 1).
\]

By using this construction, the same results can be readily obtained through a trivial extension of a Theorem of LeCam [6] together with a method suggested by Pyke ([7], p. 414). From this remarks and Theorem 2.1 we have the following analog of Corollary 2.9 in [9].

**Theorem 2.2**

The processes \((\tilde{\eta}_n(x), 0 \leq x \leq \infty)\) converge weakly to a Gaussian process \((\tilde{\eta}(x), 0 \leq x \leq \infty)\) with mean zero and covariance function
\[ \hat{\kappa}(x,y) = \min(H_k(x),H_k(y)) - H_k(x)H_k(y) - \frac{1}{k}\ xyf_k(x)f_k(y). \]  

(2.9)

Returning now to our original problem we can write

\[ T_n' = \sqrt{N} \left( T_n - \int_0^\infty m(x)dH_k(x) \right) = C_k(n) + W_n \]  

(2.10)

where

\[ C_k(n) = \frac{\sqrt{N}}{n^{\frac{1}{2}}} \left( \int_0^2 (x^2 dx) \int_0^\infty m(x)d(\nabla f_k(x)) + \frac{x^2}{2} f_k'(x) \right) \]  

(2.11)

and

\[ W_n = \int_0^\infty m(x)d\bar{\eta}_n(x)dx. \]  

(2.12)

To be able to make efficiency comparisons we require

\[ 0 < \lim_{n \to \infty} C_k(n) < \infty \]  

which is equivalent to the conditions

\[ \int_0^\infty m(x)d(\nabla f_k(x)) + \frac{x^2}{2} f_k'(x) < \infty \]  

(2.13)

and \( \delta = \frac{1}{4} \) since the analysis is only valid for fixed \( k \). We assume \( \delta = \frac{1}{4} \) in all what follows. Integrating (2.12) by parts we get

\[ W_n = - \int_0^\infty m'(x)\bar{\eta}_n(x)dx \]

provided

\[ m \text{ is absolutely continuous in } (0,\infty) \]  

(2.14)

\[ \lim_{x \to \infty} m(x)(1-H_k(x)) = 0 \]  

(2.15)

and

\[ \lim_{x \to 0} m(x)H_k(x) = 0 \]  

(2.16)
The weak convergence of the processes \( \tilde{\eta}_n \) implies that the limit distribution of \( W_n \) will be the same as that of
\[
- \int_0^\infty \tilde{m}'(x)\tilde{\eta}(x)dx
\]
if the linear functional
\[
L: y \rightarrow \int_0^\infty \tilde{m}'(x)y(x)dx, \ y \in \mathcal{D}[0, \infty]
\] (2.17)
is continuous (i.e. bounded) with probability one under the probability measure generated by \( \tilde{\eta} \). It is easy to check that, with probability one, all trajectories of \( \tilde{\eta} \) are continuous ([3], p. 183) and that they tend to zero as \( x \) tends to \( \infty \).

Therefore we can work with the uniform instead of the Skorohod topology. For the continuity of \( L \) we assume as in [8] that \( m \) is bounded in every closed interval in \([0, \infty)\). To relax the boundness condition near zero and infinity one must make better use of the fact that it is only necessary to consider functions \( y(x) \) that are trajectories of the Gaussian process \( \tilde{\eta} \). First we notice that \( \eta(x) \) can be written as \( \eta^*(H_k(x)) \), where \( \eta^* \) is a tied-down Brownian motion in \([0, 1]\). By applying the law of the iterated logarithm as in [8], it is easy to see that the contribution of \( x f_k(x) \xi \) is neglectable and arrive to the conditions
\[
\int_0^\varepsilon |m(x)|\sqrt{2H_k(x)} \log \log(H_k(x))^{-1}dx < \infty
\] (2.18)
\[
\int_M^\infty |m(x)|\sqrt{2(1-H_k(x))} \log \log(1-H_k(x))^{-1}dx < \infty
\] (2.19)

We remark that the variance of the normal random variable
\[ \int_0^\infty m'(x)\tilde{\eta}(x)dx \] is automatically finite given these conditions.

For practical purposes it is more convenient to have sufficient conditions for (2.13)-(2.16) and (2.18)-(2.19) that can be easily checked. One possible set of such conditions is (*) below:

i) \( m \) is absolutely continuous in \((0,\infty)\) and \( m' \) is bounded on any closed interval in \((0,\infty)\)

ii) \( m \) is monotone in the neighborhood of 0 and \( \infty \)

iii) \( \lim_{x \to \infty} e^{-\alpha x} m^2(x) = 0 \) for some \( \alpha < 1 \)

iv) \( \lim_{x \to 0} x^\beta m^2(x) = 0 \) for some \( \beta < k \)

We assume in what follows that (*) is satisfied.

Let \( \mu \) and \( \sigma^2 \) be the asymptotic mean and variance of \( T_n' \). Then

\[ \sigma^2 = \int_0^{\infty} \int_0^y m'(x)m'(y)\tilde{\eta}(x,y)dx\,dy \quad (2.20) \]

\[ = \int_0^{\infty} m^2(x)f_k(x)dx - \left( \int_0^{\infty} m(x)f_k(x)dx \right)^2 - \frac{1}{k} \left( \int_0^{\infty} m(x)(x-k)f_k(x)dx \right)^2 \quad (2.21) \]

Under the null hypothesis (2.21) can be obtained directly from LeCam's theorem.

\[ \mu = \frac{1}{\sqrt{k}} \int_0^{\infty} \int_0^{\infty} x^2 m(x)dx dy \quad (2.22) \]

\[ = - \frac{(k+1)}{2} \sqrt{k} \int_0^{\infty} x^2 (x)dx \int_0^{\infty} (f_{k+1}(x)-f_{k+2}(x))dx \quad (2.23) \]

\[ = \frac{1}{2\sqrt{k}} \int_0^{\infty} x^2 m''(x)f_k(x)dx \quad (2.24) \]
the last equality being true if the integration by parts of (2.23) can be justified.

3. Asymptotic relative efficiency

We turn now to the question of computing the A.R.E. of one test with respect to another when we consider different functions $m$. Let $\mu_i, \sigma_i^2$ be the asymptotic mean and variance of the statistic $T_{ni}$ corresponding to the $i$-th test with associated function $m_i$. Due to the asymptotic normality of $T_{ni}$ the A.R.E. of test 1 with respect to test 2 is simply

$$\text{ARE}(1,2) = \frac{\frac{1}{\delta}}{\frac{1}{\sigma_1^2}}$$

with $\delta = \frac{1}{4}$ in our case (see [5]). In [8] an exponent 2 instead of 4 was wrongly used. We will compute

$$e_m(k) = \frac{\mu_{m,k}^2}{\sigma_{m,k}^2} \sqrt{\frac{1}{\int_0^k \ell^2(x)dx}^2}$$

for different functions $m$ and different $k$. Then the ARE can be obtained as

$$\text{ARE}(1,2) = \frac{\frac{e_{m_1}^2(k_1)}{e_{m_2}^2(k_2)}}{.}$$

Examples

1) $m(x) = x^\alpha, \ 2\alpha > -k, \ \alpha(\alpha-1) \neq 0$

From (2.21) and (2.24) we get
\[ e(k) = \frac{1}{4k} \left( \frac{\alpha^2 (\alpha-1)^2 \Gamma^2(\alpha+k)}{\Gamma(2\alpha+k) \Gamma(k) - \Gamma^2(\alpha+k)(1+\alpha^2/k)} \right) \]

In particular for \( \alpha = 2 \)
\[ e(k) = \frac{k+1}{2} \]

In general we can prove
\[ \lim_{k \to \infty} \frac{2e(k)}{k} = 1 \text{ for } \alpha(\alpha-1) \neq 0 \]

by using the following formula ([4], p. 245):
\[ \frac{\Gamma(n)}{\Gamma(n+\beta)} = \frac{1}{n^\beta} - \beta(\beta+1) \frac{1}{2n^\beta+1} + o\left(\frac{1}{n^{\beta+1}}\right) \quad \beta > 0 \quad n \to \infty \]

2) \( m(x) = \log x \)

From (2.21) and (2.24)
\[ e(k) = \frac{1}{4k \left( \sum_{j=1}^{\infty} \frac{1}{j^2} - \frac{1}{k} \right)} \]

Again
\[ \lim_{k \to \infty} \frac{2e(k)}{k} = 1 \]

since
\[ \sum_{k} \frac{1}{k j^2} = \frac{1}{k} + \frac{1}{2k^2} + o\left(\frac{1}{k^2}\right) \quad (k \to \infty) \]

3) \( m(x) = |x-k| \)

From (2.21) and (2.23)
\[ e(k) = \frac{(k+1)^2 f_{k+2}(k)}{1 - 4k f_{k+1}(k) - [2(1 - F_{k+2}(k)) - \frac{2k}{k+1} f_{k+1}(k) - 1]^2}. \]

In particular,

\[ e(1) = \frac{1}{8e-20} \approx 0.5726. \]

Because of the convergence of \( f_k \) suitably normalized to a normal density, it is easy to see that

\[ f_{k+1}(k) \approx f_{k+2}(k) \approx \frac{1}{\sqrt{2\pi k}} \text{ for large } k \]

\[ F_{k+2}(k) \rightarrow \frac{1}{2} \text{ as } k \rightarrow \infty \]

and from this we get

\[ \lim_{k \rightarrow \infty} \frac{2e(k)}{k} = \frac{1}{11-2} < 1. \]

To find the asymptotically most efficient test for any fixed \( k \), within the class of tests we are considering, we must maximize \( e_m(k) \). From (2.20) and (2.23) this is equivalent to the maximization of

\[ \int_0^{\infty} (f_m'(x) \frac{k(k+1)}{4} (f_{k+1}(x) - f_{k+2}(x)) dx)^2 \]

\[ \int_0^{\infty} \int_0^{\infty} (x)m'(y)K(x,y) dx \]

where this expression is taken to be zero if the numerator vanishes.

Let \( P \) be the integral operator defined by

\[ (Ph)(x) = \int_0^{\infty} h(y)K(x,y) dy \]
where $h$ belongs to a Hilbert space containing the derivatives of the functions satisfying (*) and with inner product defined by

$$<h,f> = \int \int K(x,y)h(x)f(y)dxdy$$

(3.5)

It is easy to check that the null space of $P$ is the set of constant functions in $(0,\infty)$. Let

$$h = m$$

$$g(x) = \frac{k(k+1)}{4}(f_{k+1}(x) - f_{k+2}(x))$$

and assume there exists $g^*$ in the domain of $P$ such that $Pg^* = g$. Then the problem is equivalent to the maximization of

$$\frac{<h,g^*>^2}{<h,h>}$$

which has the solution $h = \lambda g^*$. It can be verified that $g^*(x) = 2x$ satisfies $P g^* = g$. Therefore the solution to the maximization problem is

$$m'(x) = 2\lambda x + B$$

or

$$m(x) = \lambda x^2 + Bx + C$$

Since

$$\frac{1}{N} \sum_{i=1}^{N} (\lambda (nD_i)^2 + B nD_i + C) = \frac{1}{N} \sum_{i=1}^{N} (nD_i)^2 + Bk + C$$
it is clearly enough to consider \( m(x) = x^2 \). This result has been proved by Sethuraman and Rao in the particular case \( k = 1 \). As a by-product of our analysis, the following theorem is easily obtained.

**Theorem 3.1**

Assume (*) holds. Then the ARE of a test with statistic \( T_{n1} \) with respect to another with statistic \( T_{n2} \) is equal to the ratio \( \rho_1^4 / \rho_2^4 \), where \( \rho_i \) is the asymptotic correlation coefficient of \( T_{n1} \) and the statistic \( T_n \) corresponding to \( m(x) = x^2 \).

4) **Commentary**

Although the analysis is only valid for fixed \( k \), the examples in the last section suggest that we should let \( k \) increase with \( n \). *Formally* putting \( k = k(n) \sim n^{1-4\beta} \), \( 0 < \beta \leq \frac{1}{4} \) in (2.10) we could take \( \delta = \frac{1}{2} - \beta \) and still have limiting power strictly greater than the significance level in all three examples. The case \( \beta = 0 \) corresponds to \( N \) fixed and no version of the central limit theorem could possibly be applied. For \( N \) fixed there are relations between the tests with \( m(x) = x^2 \) or \( m(x) = x^{-1} \) and \( \chi^2 \) tests with cells determined by sample quantiles. Also let \( -2\log W \) be the Bartlett statistic (See [2]) for the comparison of the variance of \( N \) normal populations when independent random samples of size
2k+1 are available. Then \(-2\log W\) and
\[
-2k\sum_{i=1}^{N} \log(nD_i) + 2n\log k
\]
have the same null distribution. In these two cases, the convenience of keeping \(N\) small and \(k\) large is obvious. Research on the asymptotic distribution of \(T'_n\) (suitably normalized) when \(k\) depends on \(n\) is currently being done by the author.

Acknowledgement

I would like to express my gratitude to Professor Grace Wahba for suggesting the problem and her encouragement in the course of this work. I would also like to thank Professors J. Sethuraman and J.S. Rao for making available some of their unpublished material.
REFERENCES


**Title:** On the asymptotic distribution of some goodness of fit tests based on spacings

**Author:** Guido E. Del Pino

**Performing Organization:**
Department of Statistics  
University of Wisconsin  
Madison, Wisconsin

**Controlling Office:**
Air Force Office of Scientific Research  
1400 Wilson Blvd.  
Arlington, Virginia

**Distribution Statement:** Unclassified

**Abstract:** Let $X_1, \ldots, X_{(n)}$ be independent and identically distributed random variables with common distribution $F$ on $[0,1]$ and let $X_1', \ldots, X_{(n)}'$ be the corresponding ordered observations. The problem of testing the hypothesis that $F$ is...
uniform in [0,1] based on the statistic \( T_n = \frac{1}{N} \sum_{i=1}^{N} \frac{m[n(X_{ik} - X'_{i-1} + 1)]}{N} \)

is considered, where \( k \) is a fixed number independent of \( n \), \( m \) is a function satisfying some regularity conditions and \( X_0' \equiv 0, X_n' \equiv 1 \). The asymptotic distribution of \( T_n \) under a sequence of local alternatives is obtained and it is used to compute Pitman efficiencies of tests based on different functions \( m \). This generalizes results of Rao and Sethuraman [in Nonparametric Techniques in Statistical Inference, Ed. M.L. Puri] who only study the case \( k = 1 \). Examples are given where, for a given \( m \), the Pitman efficiency increase arbitrarily with \( k \).