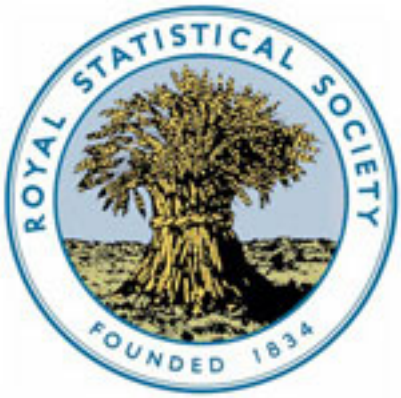

J. R. Statist. Soc. B (1984),
46, *No.* 2, *pp.* 228–237

Asymptotic Results on the Greenwood Statistic and some of its Generalizations

By J. S. RAO† and MORGAN KUO

University of California, USA *I.T.T., Vandenberg A.F.B., Lompoc*

[Received October 1982. Revised March 1983]

SUMMARY

There has been a renewed interest in the Greenwood statistic for goodness of fit and its percentage points. See Burrows (1979), Currie (1981) and Stephens (1981). This statistic, which is based on the sum of squares of the sample spacings, is known to be locally most powerful among all tests based symmetrically on the simple one-step spacings. This paper discusses some generalizations of this statistic based on higher-order or m-step spacings and shows them to be asymptotically more efficient. It is also shown that the limiting efficiency increases further if the test is based on overlapping m-step spacings as opposed to the disjoint (or non-overlapping) m-step spacings.

Keywords: GOODNESS OF FIT; SAMPLE SPACINGS; HIGHER-ORDER SPACINGS; GREENWOOD'S STATISTIC; ASYMPTOTIC EFFICIENCY

1. INTRODUCTION

Let $X_1, \ldots, X_{n-1}$ be independently and identically distributed (iid) with a given continuous cumulative distribution function (cdf) F. The goodness-of-fit problem is to test if this df is equal to a specified one. A simple probability integral transformation on the random variables reduces the support of F to $[0, 1]$ and permits us to equate the specified df to the uniform distribution on $[0, 1]$. Thus from now on we shall assume that this reduction has been effected and that the null hypothesis of interest is

$$H_0: F(x) = x, \quad 0 \leqslant x \leqslant 1. \tag{1.1}$$

Among the many procedures available are those based on sample spacings. See, for instance, Pyke (1965) and references contained therein. Let $X_1' \leqslant \ldots \leqslant X_{n-1}'$ be the order statistics. The simple or one-step spacings are defined by

$$D_i = (X_{i+1}' - X_i'), \quad i = 0, \ldots, n-1, \tag{1.2}$$

where we put $X_0' = 0$ and $X_n' = 1$. It may be noted that the order statistics as well as the spacings should have an extra subscript n to denote the sample size but it will be suppressed throughout for notational simplicity. Most common among test statistics based on spacings are $n^{-1} \Sigma |D_i - n^{-1}|$, $n^{-1} \Sigma \log (nD_i)$ and $n^{-1} \Sigma (nD_i)^r$ for $r \geqslant -\frac{1}{2}, r \neq 0, 1$. The last statistic for $r = 2$, namely,

$$G_{1n} = n^{-1} \sum_{0}^{n-1} (nD_i)^2 \tag{1.3}$$

was proposed by Greenwood (1946) and will be called throughout the Greenwood statistic. The asymptotic normality of these statistics and in fact, more generally, that of statistics of the form

† *Present address*: Statistics Dept, University of California, Santa Barbara, CA., USA

 0035–9246–84–46228 $2.00

$$T_{1n} = n^{-1} \sum_{i=0}^{n-1} h(nD_i), \tag{1.4}$$

where $h(\cdot)$ is a function that satisfies some mild regularity conditions, has been established for instance in Rao and Sethuraman (1975). The exact distributions are often more difficult and therefore tables of percentage points are hard to construct. It is this latter aspect of the Greenwood statistic G_{1n} that Burrows (1979), Currie (1981) and Stephens (1981) address.

The importance attached to G_{1n} is somewhat justified in view of the result established in Sethuraman and Rao (1970) that among all symmetric statistics of the form T_{1n} in (1.4), the Greenwood statistic G_{1n} has the maximum Asymptotic Relative Efficiency (ARE). See also Lemma 2.4. However, recent results on the asymptotic theory of spacing statistics suggest that there are asymptotically more efficient procedures than the Greenwood statistic if one takes into account test statistics based on higher-order spacings. Two types of higher-order or m-step spacings ($m \geqslant 1$, fixed) should be distinguished. The *overlapping m-step spacings* are defined by

$$D_i^{(m)} = X'_{i+m} - X'_i, \quad i = 0, 1, \ldots, n-1. \tag{1.5}$$

The non-overlapping or *disjoint m-step spacings*, on the other hand, are defined by

$$D_{i\cdot m}^{(m)} = X'_{(i+1)\cdot m} - X'_{i\cdot m}, \quad i = 0, 1, \ldots, \left[\frac{n}{m}\right] - 1, \tag{1.6}$$

where $[n/m]$ denotes the integer part of (n/m). Since we will only be concerned with asymptotic properties, it is convenient to assume, without loss of generality, that $n = N \cdot m$ so that $(n/m) = N$ is an integer in (1.6) and also to define circularly $X'_k = 1 + X'_{k-n}$ in (1.5). Del Pino (1979) studied symmetric test statistics based on disjoint m-spacings defined in (1.6) and using methods developed in Rao and Sethuraman (1975), establishes the asymptotic normality of statistics of the form

$$T_{2n}(m) = N^{-1} \sum_{i=0}^{N-1} h(nD_{i\cdot m}^{(m)}). \tag{1.7}$$

He also shows that a statistic analogous to Greenwood's, namely,

$$G_{2n}(m) = N^{-1} \sum_{i=0}^{N-1} (nD_{i\cdot m}^{(m)})^2 \tag{1.8}$$

based on the sum of squares of these disjoint m-spacings has the maximum ARE among statistics of the form T_{2n}. Also the statistic $G_{2n}(m)$ is asymptotically more efficient than Greenwood's statistic G_{1n}, and the disparity grows with increasing m, the length of the step (cf. Table 1). Cressie (1976) studied the statistic $\Sigma \log (nD_i^{(m)})$ based on the overlapping m-spacings.

Kuo and Rao (1981) consider again the general class of symmetric statistics based on overlapping m-spacings

$$T_{3n}(m) = n^{-1} \sum_{i=0}^{n-1} h(nD_i^{(m)}) \tag{1.9}$$

and establish their asymptotic normality for a wide class of functions $h(\cdot)$. They also investigate the ARE's of such statistics and show again that a statistic analogous to Greenwood's, namely

$$G_{3n}(m) = n^{-1} \sum_{i=0}^{n-1} (nD_i^{(m)})^2 \tag{1.10}$$

is asymptotically the most efficient and that this efficiency again increases with m.

In Section 2, we quote some relevant results on the classes of statistics T_{1n}, T_{2n} and T_{3n} and give a simple proof why Greenwood-type statistics G_{1n}, G_{2n} and G_{3n} are best in the respective categories. We conclude Section 2 with a short table of efficacies of G_{1n}, $G_{2n}(m)$ and $G_{3n}(m)$ for different m and questions about choice of m.

In Section 3, the advantage of considering statistics based on overlapping m-spacings, as compared to those based on disjoint m-spacings, is investigated.

2. SOME RELEVANT RESULTS

Although the exact calculation of power, when it can be done, is preferable for small sample comparisons of two competing tests, it is not often that this can be done in practice. In such a case, one resorts to asymptotic measures of test efficiency as an indication of the relative local powers, i.e. for alternatives close to the null hypothesis. An introduction to various measures of asymptotic efficiencies of tests, including Pitman's Asymptotic Relative Efficiency (ARE), can be found for instance in C. R. Rao (1973, pp. 464-470). In general, since the power at any fixed alternative, approaches one as the sample size increases to ∞ for any reasonable test (implying consistency), such asymptotic comparisons can only be made for an appropriate sequence of alternatives which converge to the null hypothesis as $n \to \infty$. In this situation, the ARE of one test with respect to another, may be interpreted as the inverse ratio of sample sizes needed by the tests to have the same power at such a sequence of alternatives (cf. Rao, 1973, p. 469). For statistics based symmetrically on spacings, namely T_{1n}, T_{2n} and T_{3n}, the appropriate sequence of alternatives is given by the cdf (cf. Rao and Sethuraman, 1975; Del Pino, 1979; Kuo and Rao, 1981)

$$F_n(x) = x + \frac{L_n(x)}{n^{1/4}}, \quad 0 \leqslant x \leqslant 1, \tag{2.1}$$

where $L_n(0) = L_n(1) = 0$. We assume that $L_n(x)$ is continuously differentiable with derivative $l_n(x)$. Let $L(x)$ be a twice continuously differentiable function defined on $[0, 1]$ with the first and second derivatives denoted by $l(x)$ and $l'(x)$ respectively such that $L_n(x)$ converges to $L(x)$ uniformly on $[0, 1]$, so that

$$\sup_{0 \leqslant x \leqslant 1} | l_n(x) - l(x) | = o(1). \tag{2.2}$$

We now quote three results about the asymptotic distributions of the general statistics T_{1n}, T_{2n} and T_{3n} respectively, when the observations $X_1, \ldots, X_{n-1}$ come from the alternative cdf (2.1). These results give the limit distributions under the null hypothesis by putting $L_n(x) = L(x) = 0$, $0 \leqslant x \leqslant 1$, and also allow us to compute the ARE.s of such statistics.

Let $Z_0, Z_1, \ldots, Z_{n-1}$ be a sequence of iid exp (1) random variables with pdf e^{-z} for $z > 0$. Define the partial sums

$$S_k = Z_0 + \ldots + Z_{k-1}, \quad k = 1, 2, \ldots, n. \tag{2.3}$$

Define also the rotating partial sums (of m terms at a time)

$$S_k^{(m)} = \sum_{j=0}^{m-1} Z_{k+j}, \quad k = 0, 1, \ldots, n-1 \tag{2.4}$$

with the convenient notation $Z_j = Z_{j-n}$ for $j \geqslant n$. Finally let Z stand as a generic symbol for exp (1) random variable and S for a Gamma $(m, 1)$ random variable with pdf

$$s^{m-1}e^{-s}/\Gamma m \quad \text{for } s \geqslant 0. \tag{2.5}$$

Rao and Sethuraman (1970, 1975) express statistics of the form T_{1n} in (1.4) as a functional of the empirical spacings process and derive their limit distributions. Using a similar approach Del Pino (1979) obtains the limiting distributions of statistics of the form T_{2n} in (1.7). Kuo and Rao (1981), using a more direct approach based on Taylor expansions, study statistics of the form T_{3n} defined in (1.9). We now quote the main results of each of these papers in terms of the present terminology and notations. The detailed regularity conditions on the class of functions $h(\cdot)$ are given in the respective papers and are omitted here. We are content to remark that the classes are quite broad and include all the statistics we mentioned in Section 1.

Theorem 2.1 (Sethuraman and Rao, 1970, pp. 405-416, Theorem 3). Under the alternatives (2.1), the statistic T_{1n} in (1.4), with $h(\cdot)$ satisfying some regularity conditions (ibid.), has the asymptotic distribution

$$\sqrt{n}(T_{1n} - Eh(Z)) = n^{-\frac{1}{2}} \sum_{i=0}^{n-1} [h(nD_i) - Eh(Z)] \xrightarrow{d} N(\mu_1, \sigma_1^2),$$

where

$$\mu_1 = \frac{1}{2} \left(\int_0^1 l^2(u)\, du \right) \cdot \operatorname{cov}(h(Z), (Z-2)^2), \tag{2.6}$$

and

$$\sigma_1^2 = \operatorname{Var}(h(Z)) - \operatorname{cov}^2(h(Z), Z) \tag{2.7}$$

and $\xrightarrow{d}$ denotes convergence in distribution. □

Theorem 2.2 (Del Pino, 1979). Under the alternatives (2.1), the statistic T_{2n} in (1.7) with $h(\cdot)$ satisfying some regularity conditions (ibid.), has the asymptotic distribution

$$\sqrt{N}(T_{2n} - Eh(S)) = N^{-\frac{1}{2}} \sum_{i=0}^{N-1} [h(nD_{i \cdot m}^{(m)}) - Eh(S)] \xrightarrow{d} N(\mu_2, \sigma_2^2),$$

where

$$\mu_2 = \left(\int_0^1 l^2(u)\, du \right) \cdot \operatorname{cov}(h(S), (S-m-1)^2) \Big/ 2\sqrt{m} \tag{2.8}$$

and

$$\sigma_2^2 = \operatorname{var}(h(S)) - \operatorname{cov}^2(h(S), S)/m. \quad \square \tag{2.9}$$

Finally,

Theorem 2.3 (Kuo and Rao, 1981). Under the alternatives (2.1), the statistic T_{3n} in (1.9) with $h(\cdot)$ satisfying some regularity conditions (ibid.), has the asymptotic distribution

$$\sqrt{n}(T_{3n} - Eh(S)) = n^{-\frac{1}{2}} \sum_{i=0}^{n-1} [h(nD_i^{(m)}) - Eh(S)] \xrightarrow{d} N(\mu_3, \sigma_3^2),$$

where

$$\mu_3 = \left(\int_0^1 l^2(u)\, du \right) \cdot \operatorname{cov}\left(h(S), (S-m-1)^2\right) \Big/ 2 \tag{2.10}$$

and

$$\sigma_3^2 = \sum_{-m+1}^{m-1} \operatorname{cov}\left(h(S_0^{(m)}), h(S_k^{(m)})\right) - (\operatorname{cov}(h(S), S))^2 . \quad \square \tag{2.11}$$

Clearly, the distribution under the null hypothesis (1.1) is obtained by putting $l(u) \equiv 0$, $0 \leqslant u \leqslant 1$ in any of these theorems. This reduces the mean of the normalized statistics to zero with the variance remaining the same. But the main reason for considering distributions under the alternatives (2.1) is that this allows computation of ARE's, which we now describe very briefly.

Let $\mu(h)$ and $\sigma^2(h)$ denote the asymptotic mean and variance of the test statistic $T_n(h)$ based on the function $h(\cdot)$ under the sequence of alternatives (2.1). Here it is assumed that the test statistic $T_n(h)$ has been normalized to have asymptotic mean zero and finite variance under the null hypothesis. Then under certain standard regularity conditions, which include a condition on the nature of alternatives and asymptotic normal distribution of $T_n(h)$ under these alternatives, the Pitman asymptotic relative efficiency of $T_n(h_1)$ with respect to $T_n(h_2)$, denoted by ARE (h_1, h_2), can be calculated by

$$\text{ARE}\,(h_1, h_2) = \left(\frac{\mu^2(h_1)}{\sigma^2(h_1)} \right)^2 \Big/ \left(\frac{\mu^2(h_2)}{\sigma^2(h_2)} \right)^2 \tag{2.12}$$

(see, for example, Fraser, 1957). The value $e^4(h) = \mu^4(h)/\sigma^4(h)$ is called the "efficacy". The test with maximum efficacy has asymptotically maximum local power. To find such a test, against the specific alternatives (2.1), we need to find a function $h(\cdot)$ which maximizes $e(h)$. The following lemma shows that in all these cases, the locally optimal test is obtained by taking $h(x) = x^2$, i.e. the Greenwood-type statistic based on the sum of squares of spacings.

Lemma 2.4. The value of $e_i(h) = \mu_i(h)/\sigma_i(h)$ for any of the three classes of statistics $(i = 1, 2, 3)$ is maximized by taking the function $h(x) = x^2$, i.e. among the classes of statistics T_{1n}, T_{2n} and T_{3n}, the locally optimal test is provided by G_{1n}, G_{2n} and G_{3n} respectively (cf. equations (1.3), (1.8) and (1.10)).

Proof. Within any specific class, say i, consider all non-degenerate statistics (i.e. with non-zero variance) $T_{in} = T_{in}(h)$ obtained by varying over all $h(\cdot)$, which satisfy the regularity conditions. Since the efficacy is unaffected by a linear transformation of the statistic, we may assume without any loss of generality that $\sigma_i^2(h) = 1$. Thus the problem of finding a $h(\cdot)$ for which the maximum efficacy is obtained is the same as finding a $h(\cdot)$ which maximizes the numerator in $e_i(h)$, namely $\mu_i(h)$. Thus by the Cauchy-Schwartz inequality,

$$e_1(h) = \mu_1(h) = \frac{1}{2} \left(\int_0^1 l^2(u)\, du \right) \cdot \operatorname{cov}\left((h(Z)), (Z-2)^2\right)$$

$$\leqslant \frac{1}{2} \left(\int_0^1 l^2(u)\, du \right) \cdot (\operatorname{var}(h(Z)))^{1/2} \cdot (\operatorname{var}(Z-2)^2)^{1/2}. \tag{2.14}$$

Similarly,

$$e_2(h) = \mu_2(h) = \frac{1}{2\sqrt{m}} \left(\int_0^1 l^2(h)\, du \right) \cdot \operatorname{cov}\left(h(S), (S-m-1)^2\right)$$

$$\leqslant \frac{1}{2\sqrt{m}} \left(\int_0^1 l^2(h)\, du \right) \cdot (\operatorname{var}(h(S)))^{1/2} \cdot (\operatorname{var}(S-m-1)^2)^{1/2} \tag{2.15}$$

and

$$e_3(h) = \mu_3(h) = \frac{1}{2} \left(\int_0^1 l^2(u)\, du \right) \cdot \operatorname{cov}\left(h(S), (S-m-1)^2\right)$$

$$\leqslant \frac{1}{2} \left(\int_0^1 l^2(u)\, du \right) \cdot (\operatorname{var}(h(S)))^{1/2} \cdot (\operatorname{var}(S-m-1))^{1/2}. \tag{2.16}$$

These inequalities become equalities if and only if $h(x) = a(x-2)^2 + b$ in (2.14) and $h(x) = a(x-m-1)^2 + b$ in (2.15) and (2.16) for some real numbers $a \neq 0$ and b. Since the sum over the x term results in a constant, the optimal statistic is equivalent to the one obtained by putting $h(x) = x^2$. This gives (1.3), (1.8) and (1.10) as the asymptotically locally most powerful tests in the respective classes T_{1n}, $T_{2n}(m)$ and $T_{3n}(m)$. □

Specializing Theorems 2.1, 2.2 and 2.3 to the case $h(x) = x^2$, one gets the asymptotic distributions of the three Greenwood-type statistics G_{1n}, $G_{2n}(m)$ and $G_{3n}(m)$. The asymptotic means and variances are presented in Table 1 below. Also since $[\int_0^1 l^2(u)\, du]^4$ enters as a multiplicative factor in the definition of efficacy in (2.13), we tabulate "modified efficacies", namely $[\mu^4/\sigma^4(\int_0^1 l^2(u)\, du)^4]$.

It is clear that the efficacies of G_{2n} and G_{3n} increase with m, the length of the step and exceed that of the Greenwood statistic G_{1n} which corresponds to $m = 1$. Table 2 gives some numerical values of modified efficacies of G_{2n} and G_{3n} as a function of m.

From this it follows that both G_{2n} corresponding to the non-overlapping spacings and G_{3n} corresponding to the overlapping spacings are clearly superior to the Greenwood statistic and this superiority increases with the length of the step m. The optimal choice of m is an open question at this point and it appears one should allow m to increase unboundedly with n. This is presently under investigation. For moderate sample sizes Monte Carlo studies might provide an answer. The superiority of $G_{3n}(m)$ over $G_{2n}(m)$ for any given m is further investigated in some generality in the next section.

TABLE 1

Statistic, G_i	*Mean* μ_i	*Variance* σ_i^2	*Modified efficacy*
$G_{1n} = n^{-\frac{1}{2}} \sum_{i=0}^{n-1} [(nD_i)^2 - 2]$	$2\left(\int_0^1 l^2(u)\,du\right)$	4	1
$G_{2n} = N^{-\frac{1}{2}} \sum_{i=0}^{N-1} [(nD_{i\cdot m}^{(m)})^2 - m(m+1)]$	$m^{\frac{1}{2}}(m+1)\left(\int_0^1 l^2(u)\,du\right)$	$2m(m+1)$	$\left(\frac{m+1}{2}\right)^2$
$G_{3n} = n^{-\frac{1}{2}} \sum_{i=0}^{n-1} [(nD_i^{(m)})^2 - m(m+1)]$	$m(m+1)\left(\int_0^1 l^2(u)\,du\right)$	$2m(m+1)(2m+1)/3$	$[3m(m+1)/(4m+2)]^2$

TABLE 2

m	G_{1n}	$G_{2n}(m)$	$G_{3n}(m)$
1	1.000	1.000	1.000
2	–	2.250	3.240
3	–	4.000	6.610
4	–	6.250	11.111
5	–	9.000	16.736
10	–	30.250	61.732
20	–	110.250	236.114
50	–	650.250	1434.213

3. SUPERIORITY OF THE CLASS OF STATISTICS BASED ON OVERLAPPING SPACINGS

Recall the definitions (1.5) and (1.6) of the overlapping m-spacings and the non-overlapping m-spacings. The latter is a subset of the former with the ith disjoint m-spacing corresponding to the $(i \cdot m)$th overlapping m-spacing. In this section, we compare the ARE of statistics of the form T_{3n} in (1.9) with the corresponding statistic T_{2n} in (1.7), i.e. for the same function $h(\cdot)$. As may be expected, it is shown that overlapping provides higher efficacies (see Theorem 3.2). We shall assume as before that $N = (n/m)$ is an integer without any loss of generality and define

$$V_{j,n} = N^{-1} \sum_{i=0}^{N-1} h(nD^{(m)}_{i \cdot m + j}), \quad j = 0, 1, \ldots, m-1. \tag{3.1}$$

This $V_{j,n}$ is based on non-overlapping or disjoint m-spacings starting from the jth order statistic and $V_{0,n} = T_{2n}$ defined in (1.7). On the other hand, T_{3n} in (1.9) is based on all the overlapping m-spacings and one may write

$$T_{3n} = m^{-1} \sum_{j=0}^{m-1} V_{j,n}. \tag{3.2}$$

Thus T_{3n} is a simple average of the $\{V_{j,n}, j = 0, \ldots, m-1\}$ each of which are based on disjoint m-spacings. Clearly $\{V_{j,n}\}$ are exchangeable random variables and from Theorem 2.2, $V_{j,n} \xrightarrow{d} V_j$, say which has an $N(\mu_2, \sigma_2^2)$ distribution under the alternatives (2.1). The proof of the present Theorem 2.3 as given in Kuo and Rao (1981) can be easily adapted to establish the asymptotic normality of the weighted average $\Sigma_{j=0}^{m-1} a_j \cdot V_{j,n}$ (instead of the unweighted average T_{3n} in (3.2)) for any set of real numbers $(a_0, \ldots, a_{m-1})$. This implies the limiting multivariate normality of $\{V_{j,n}, j = 0, \ldots, m-1\}$. Denoting an m-variate normal distribution by $N_m(\cdot, \cdot)$, we thus have the following

Theorem 3.1. If the assumptions on $L_n(x)$ and $h(\cdot)$ hold as in Theorem 2.3, then under the close alternatives (2.1), the vector $W_n = \{V_{0,n}, \ldots, V_{m-1,n}\}$ converges in distribution to the vector $W = \{V_0, \ldots, V_{m-1}\}$, say which has an $N_m(\mu, \tau)$ distribution with

$$\mu = (\mu_2, \ldots, \mu_2)' \tag{3.3}$$

with μ_2 as defined in (2.8) and covariance matrix

$$\begin{pmatrix} \tau_0 & \tau_1 & \cdots & \tau_{m-2} & \tau_{m-1} \\ \tau_{m-1} & \tau_0 & \cdots & \tau_{m-3} & \tau_{m-2} \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \tau_1 & \tau_2 & \cdots & \tau_{m-1} & \tau_0 \end{pmatrix} \tag{3.4}$$

with $\tau_0 = \sigma_2^2$ defined in (2.9) and

$$\tau_j = \operatorname{cov}\left(h(S_0^{(m)}), h(S_j^{(m)}) + h(S_{m-j}^{(m)})\right) - \operatorname{cov}^2(h(S), S)/m. \quad \square \tag{3.5}$$

Observe that Theorem 2.2 is a special case of this and refers to the marginal distribution of $V_{0,n} = T_{2n}$. Setting $l = 0$ in the above result, specifically in the component μ_2 of μ, gives the asymptotic null distribution of W_n which is $N_m(0, \tau)$. Since the covariance matrix $\tau = ((\tau_{ij}))$ has the property that $\tau_{ij} = \tau_k$ if $|j - i| = k$, it is called a "circulant". The following theorem shows that the optimal (locally most powerful) test among all possible linear combinations of $\{V_{j,n}\}$ is obtained by taking their simple average, i.e. T_{3n}, and that it is always more efficient than tests based only on $T_{2n} = V_{0,n}$.

Theorem 3.2. Let $\{V_{j,n}\}$ and $\{T_{3n}\}$ be as defined in (3.1) and (3.2) (or (1.9)) respectively. Among all possible test statistics which are linear combinations of $\{V_{j,n}\}$, the maximum efficacy is attained for $T_{3n} = m^{-1} \sum_{j=0}^{m-1} V_{j,n}$. The ARE of the overlapping m-spacing test T_{3n} with respect to the corresponding disjoint m-spacing test $T_{2n} = V_{0,n}$ is $(m\tau_0/\sum_{j=0}^{m-1} \tau_j)^2$ which exceeds 1 for any $h(\cdot)$ except in the trivial case $m = 1$, in which case they coincide.

Proof. From Theorem 3.1, for any real vector $b' = (b_0, \ldots, b_{m-1})$, the linear combination $b'W_n$ has asymptotically an $N(b'\mu, b'\tau b)$ distribution. The problem is to find a vector b for which the efficacy $e_b = (b'\mu)^2/(b'\tau b)$ is a maximum. This is an easy problem (cf. Rao, 1973, p. 60) if the covariance matrix τ is positive definite. If not, for some $\epsilon > 0$ define the positive definite matrix $\tau_\epsilon = \tau + \epsilon I_m$ where I_m is the $(m \times m)$ identity matrix. Let

$$e(\epsilon) = \sup_b e_b(\epsilon) = \sup_b \frac{(b'\mu)^2}{(b'\tau_\epsilon b)}. \tag{3.6}$$

This supremum is attained when $b = b^* = \lambda(\tau_\epsilon^{-1}\mu)$ for some real number $\lambda \neq 0$ (cf. Rao, 1973, p. 60). Since the inverse of a circulant matrix is a circulant (see Good, 1950) and μ has all its component equal, it follows that the maximizing vector $b^* = \lambda(\tau_\epsilon^{-1}\mu)$ has all equal components, say

$$b^* = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right).$$

Then

$$e(\epsilon) = (b^{*\prime}\mu)^2/(b^{*\prime}\tau_\epsilon b^*) = (m\mu_2)^2 \Big/ \left(\sum_{j=0}^{m-1} \tau_j + \epsilon\right). \tag{3.7}$$

Since $\sum_{j=0}^{m-1} \tau_j$ in the denominator corresponds to the variance of T_{3n}, it is non-zero except for the degenerate case, i.e. when $h(x) = c_1 x + c_2$ for some constants c_1 and c_2. Thus we may let ϵ tend to zero so that the efficacy of T_{3n} is given by

$$e = \sup_b e_b = \lim_{\epsilon \to 0} e(\epsilon) = (m\mu_2)^2 \Big/ \left(\sum_{j=0}^{m-1} \tau_j \right). \tag{3.8}$$

Finally to compare the ARE's of T_{3n} and T_{2n}, we have from Theorem 2.2 and (3.8)

$$\text{ARE}(T_{3n}, T_{2n}) = \left(\frac{m\mu_2^2}{\sum_{j=0}^{m-1} \tau_j} \right)^2 \div \left(\frac{\mu_2^2}{\tau_0} \right)^2 = \left(\frac{m\tau_0}{\sum_{j=0}^{m-1} \tau_j} \right)^2.$$

This ratio is, of course, equal to 1 if $m = 1$, in which case the statistics T_{3n} and T_{2n} coincide and are the same as T_{1n}. If $m \geqslant 2$, notice that by Cauchy-Schwartz inequality

$$\tau_j = \text{cov}(V_0, V_j) \leqslant \text{var} \cdot V_0 = \tau_0 \text{ for any } j = 1, \ldots, m-1.$$

Equality holds, i.e. $\tau_j = \tau_0$ if and only if $V_j = \lambda \cdot V_0$ with probability one for some real number λ. But in view of the fact that V_j and V_0 have identical distributions, this implies $\lambda = 1$. This is impossible except in the degenerate case, thus proving that T_{3n} is strictly superior to T_{2n} for any choice of $h(\cdot)$. □

4. CONCLUSIONS

For any choice of the function $h(\cdot)$ and size of the step m, Theorem 3.2 indicates that the test statistic $T_{3n}(m)$ which makes use of overlapping spacings is superior to the corresponding test statistic $T_{2n}(m)$ which uses only the disjoint spacings. In particular the Greenwood-type statistic $G_{3n}(m)$ given in (1.10) is preferable to $G_{2n}(m)$ in (1.8) and both of them have much higher asymptotic relative efficiencies than the Greenwood statistic G_{1n}. From Table 1, the ARE of G_{3n} relative to G_{2n} for any fixed m is seen to be $9/(2 + (1/m))^2$ which is approximately 9/4 for large m. One can reduce the required sample size by approximately 4/9 for comparable power by using G_{3n} instead of G_{2n}. Thus even in moderately large samples, there is every reason to prefer G_{3n} over G_{2n} or G_{1n}. Even though the limiting theory suggests larger m values are always better, the choice of m is an important open question. One could choose m as high as the integer part of $n/2$ (beyond which it corresponds to the complement of a smaller than m-step spacing) although for practical applications, a rule of thumb about the order of m in relation to n, may be obtained from Monte Carlo studies. These and some related problems are presently under investigation.

REFERENCES

Burrows, P. M. (1979) Selected percentage points of Greenwood's statistic. *J. R. Statist. Soc.* A, **142**, 256–258.

Cressie, N. (1976) On the logarithms of high-order spacings. *Biometrika*, **63**, 343–355.

Currie, I. D. (1981) Further percentage points of Greenwood's statistic. *J. R. Statist. Soc.* A, **144**, 360–363.

Del Pino, G. E. (1979) On the asymptotic distribution of k-spacings with applications to goodness of fit tests. *Ann. Statist.*, 7, 1058–1065.

Fraser, D. A. S. (1957) *Nonparametric Methods in Statistics*. New York: Wiley.

Good, I. J. (1950) On the inversion of circulant matrices. *Biometrika*, **37**, 185–186.

Greenwood, M. (1946) The statistical study of infectious diseases. *J. R. Statist. Soc.* A, **109**, 105–109.

Kuo, M. and Rao, J. S.(1981) Limit theory and efficiencies for tests based on higher order spacings. In *Statistics–Applications and New Directions*, Proceedings of the Golden Jubilee Conference of the Indian Statistical Institute, Statistical Publishing Society, Calcutta (to appear).

Pyke, R. (1965) Spacings. *J. R. Statist. Soc.* B, **27**, 395–449.

Rao, C. R. (1973) *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.

Rao, J. S. and Sethuraman, J. (1975) Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors. *Ann. Statist.*, **3**, 299–313.

Sethuraman, J. and Rao, J. S. (1970) Pitman efficiencies of tests based on spacings. *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.). Cambridge: Cambridge University Press.

Stephens, M. A. (1981) Further percentage points of Greenwood's statistic. *J. R. Statist. Soc.* A, **144**, 364–366.