TECHNICAL REPORT NO. 441

February 1976

A SIMPLE TEST FOR GOODNESS-OF-FIT BASED ON SPACINGS WITH SOME EFFICIENCY COMPARISONS

by

M. L. Puri
Indiana University

J. S. Rao
University of Wisconsin

and

Youngjoo Yoon
Blue Cross & Blue Shield of Texas

A simple test for goodness-of-fit based on spacings with some efficiency comparisons

by

M. L. Puri
Indiana University
Bloomington

J. S. Rao
University of Wisconsin
Madison

Youngjoo Yoon
Blue Cross & Blue Shield
of Texas

1. Introduction and Summary:

Let $X_1, \ldots, X_{n-1}$ be independently and identically distributed random variables with a common distribution function (d. f.). The goodness-of-fit problem is to test if this d. f. is equal to a specified one. A simple probability integral transformation on the random variables (r. v. 's) would permit us to equate the specified d. f. to the uniform distribution on [0, 1]. Thus from now on, we shall assume that this reduction has been effected and under the hypothesis, the observations have a uniform distribution on [0, 1]. The original problem thus, is equivalent to one of testing for uniformity viz. whether a given random sample of observations come from a uniform distribution on [0,1].

Let $X'_1 \leq X'_2 \leq \ldots \leq X'_{n-1}$ be the order statistics. The sample spacings $(D_1, \ldots, D_n)$ are defined by

$$D_i = X'_i - X'_{i-1}, \quad i = 1, \ldots, n$$

where we put $X'_o = 0$, $X'_n = 1$. Tests for goodness-of-fit (or equivalently uniformity) based on spacings have been proposed by several authors. See for instance Pyke (1965) or Rao and Sethuraman (1975) and the references contained therein. It can be seen (see e. g. Pyke (1965) Section 2. 1) that the distribution of $(T_1, \ldots, T_n)$ under the hypothesis of uniformity is Dirichlet $D(1, 1 \ldots 1; 1)$ distribution with any subset $(T_{i_1}, \ldots, T_{i_k})$ of them having $D(1, \ldots, 1; n-k)$ distribution. See Wilks (1962) pp 177-182 for an elementary discussion on Dirichlet distributions.

In analysing circularly distributed data, testing for uniformity i. e. deciding whether a given set of observations on the circumference of a unit circle indicate a preferred direction, is a very basic problem. This is a necessary preliminary step before estimating or making inferences on the mean direction. Also the goodness-of-fit problem on the circle is equivalent to this just as on the line. In the circular case, the spacings may be defined as the arc-lengths between successive observations on the circumference, ignoring the zero-direction. Apart from the minor difference that n observations on the circle lead to n circular spacings while on the line (n-1) observations make n spacings, the distribution of the spacings in either case is the same (see for instance Rao (1969) pp. 63-67 or Mardia (1972) p. 172. For purposes of inference on the circle, one requires a statistic that is invariant under changes of the origin and a general invariant statistic is of the form $h(T_1, T_2, \ldots, T_n)$ where $h(\cdot)$ is a function that remains invariant under cyclical permutations of the arguments. For instance functions symmetric in all the arguments may be considered though they are not asymptotically efficient. See Sethuraman and Rao (1970). Thus the spacings $\{T_i\}$ play a crucial role in testing goodness-of-fit on the circle whereas for the linear case, one has tests that are not necessarily based on spacings. Therefore all our further discussion on spacings can be related also to the circular case and is indeed more important in that context.

In Section 2, we propose a simple class of tests $R_n = R_n(n\delta_n)$ based on spacings and obtain the exact distribution under the hypothesis of uniformity. Section 3, deals with the asymptotic distribution of R_n while sections 4 and 5 respectively discuss the Asymptotic Relative Efficiency (ARE) and Bahadur Efficiency (BE) of R_n relative to U_n, another spacings test discussed by Rao (1969). Since the limiting efficiencies of a number of test-statistics including U_n have already been investigated by Sethurmaman and Rao (1970) and Rao (1972), the results of sections 4 and 5 provide a basis for comparing R_n with any of those tests. Finally in Section 6 we discuss the statistic R_n^*, which has the maximum limiting efficiency in the class of tests $R_n(n\delta_n)$. We also provide a table that can be used to obtain critical values of R_n^* and illustrate, by means of a numerical example, how simple it is to use this R_n^* - test.

2. The statistic R_n and its exact null distribution.

Choose and fix a $\delta_n > 0$. We shall call a sample spacing 'small' if it is less than δ_n in length. The test criterion is to reject H_o, the hypothesis of uniformity when we observe too many 'small' spacings, since this clearly indicates clustering of the observations. At this stage we will leave open the choice of δ_n though a suitable value might be to take for instance $\delta_n = \frac{1}{n}$, the expected length of any spacing under uniformity. Since T_i's are of order (1/n) under H_o, we consider the so-called "normalized" spacings $\{nT_i\}$ and define

$$R_n = R_n(n\delta_n) = \{\text{number of } (nT_i)\text{'s} \leq n\delta_n\},$$

$$= \text{number of spacings } T_i \text{ smaller than } \delta_n.$$

and reject H_o if R_n is too large. The exact distribution of R_n is given by the following.

Theorem 2.1 Under the hypothesis of uniformity, the probability function of R_n is given by

$$P(R_n = k) = \binom{n}{k} \sum_{j=0}^{k} (-1)^j \binom{k}{j} \langle 1-(n-k+j)\,\delta_n \rangle^{n-1}$$

(2.1) $$\text{for } k = 0, 1, \ldots, n-1$$

$$= 0 \quad \text{otherwise}$$

with the notation $\langle x \rangle = x$ if $x > 0$ and $= 0$ if $x \le 0$.

Proof:

Let E_i denote the event that i^{th} spacing T_i exceeds δ_n, $i = 1, \ldots, n$ and let P_m denote the probability that a specified set of m arcs exceed δ_n. Clearly we have to have $m \le \left[\frac{1}{\delta_n}\right]$, the largest integer contained in $\frac{1}{\delta_n}$.

Since the spacings are exchangeable,

$$P_m = P(E_{i_1} \cap \ldots \cap E_{i_m})$$

$$= P(E_1 \cap \ldots \cap E_m)$$

$$= P(T_1 > \delta_n, \ldots, T_m > \delta_n)$$

$$= \int_{\delta_n}^{1} \cdots \int_{\delta_n}^{1-\sum_{i=1}^{m-2} t_i} \int_{\delta_n}^{1-\sum_{i=1}^{m-1} t_i} g(t_1, \ldots t_m)\, dt_m \ldots dt_1$$

$$= \begin{cases} (1 - m\,\delta_n)^{n-1} & \text{for } 0 < m \le \left[\frac{1}{\delta_n}\right] \\ 0 & \text{otherwise} \end{cases}$$

$$= \langle 1 - m\,\delta_n \rangle^{n-1}$$

with the notation $\langle\ \rangle$ used in (2.1). Now if S_m denotes the probability that any subset m of these n events take place, then because of exchangeability,

$$S_m = \binom{n}{m} P_m \tag{2.2}$$

$$= \binom{n}{m} \left\langle 1 - m\delta_n \right\rangle^{n-1} .$$

Further if Π_m denotes the probability that exactly m of these n events take place, then we have (see e.g. Feller I, p. 106)

$$\Pi_m = \sum_{j=m}^{n} (-1)^{j-m} \binom{j}{m} S_j$$

which on substituting (2.2) gives

$$= \sum_{j=m}^{n} (-1)^{j-m} \binom{n}{j}\binom{j}{m} \left\langle 1 - j\delta_n \right\rangle^{n-1}$$

$$= \binom{n}{m} \sum_{j=m}^{n} (-1)^{j-m} \binom{n-m}{n-j} \left\langle 1 - j\delta_n \right\rangle^{n-1}$$

using again the notation $\langle\ \rangle$ of (2.1). Finally since $R_n = k$ if and only if exactly $(n-k)$ of the spacings exceed δ_n (hence exactly k arcs are smaller than δ_n), we have

$$P(R_n = k) = \Pi_{n-k}$$

$$= \binom{n}{k} \sum_{j=n-k}^{n} (-1)^{j-(n-k)} \binom{k}{n-j} \left\langle 1 - j\delta_n \right\rangle^{n-1}$$

$$= \binom{n}{k} \sum_{j=0}^{k} (-1)^{j} \binom{k}{j} \left\langle 1 - (n-k+j)\delta_n \right\rangle^{n-1}$$

q. e. d.

Remark 1.

The distribution in (2.1) can also be derived by using the results of Darling (1953) who gives the characteristic function of $N_n(\alpha, \beta)$, the number of spacings with values between α and β. It is given by

$$E\left(e^{i\xi N_n(\alpha,\beta)}\right) = \frac{(n-1)!}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^z z^{-n}\left\{1+\left(e^{i\xi}-1\right)\left(e^{-z\alpha}-e^{-z\beta}\right)\right\}^n dz.$$

Since our $R_n = N_n(0, \delta_n)$, the characteristic function of R_n is obtained by letting $\alpha = 0$ and $\beta = \delta_n$, i.e.

$$E\left(e^{i\xi R_n}\right) = \frac{(n-1)!}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^z z^{-n}\left\{1+(e^{i\xi}-1)(1-e^{-z\delta_n})\right\}^n dz .$$

If we expand the factor in braces and select the coefficient of $e^{i\xi k}$ for any fixed $k = 0, 1, \ldots, n$,

$$P(R_n = k) = \frac{(n-1)!}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^z z^{-n}\left\{e^{-(n-k)\delta_n z}\left(1-e^{-\delta_n z}\right)^k\right\} dz$$

$$= \binom{n}{k}\sum_{j=0}^{k}\binom{k}{j}(-1)^j \frac{(n-1)!}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{z(1-(n-k+j)\delta_n)} z^{-n}\, dz$$

$$= \binom{n}{k}\sum_{j=0}^{k}\binom{k}{j}(-1)^j \left\langle 1-(n-k+j)\delta_n\right\rangle^{n-1} .$$

The last equality follows from the Residue Theorem.

Remark 2.

Another spacings statistic of interest is $U_n = \frac{1}{2}\sum_{i=1}^{n} |T_i - \frac{1}{n}|$ discussed in detail by Rao (1969) in connection with testing uniformity of circular distributions. Its density function was investigated by Darling (1953), Sherman (1950) and Rao (1969). We show below that this statistic U_n is closely related to $R_n(1)$ with $\delta_n = 1/n$. Let $K = n - R_n(1)$ denote the (random) number of spacings with lengths larger than $1/n$ and

$$S_K = T_{(n-k+1)} + T_{(n-k+2)} + \ldots + T_{(n)}$$

$$= T_{R_n(1)+1} + \cdots + T_{(n)}$$

where $T_{(1)} \leq \ldots \leq T_{(n)}$ are the ordered spacings. Thus S_K denotes the sum of those K largest spacings which exceed $1/n$ in length. Notice that

$$U_n = \frac{1}{2}\sum_{i=1}^{n} |T_i - \frac{1}{n}|$$

$$= \sum_{\{i: T_i > \frac{1}{n}\}} (T_i - \frac{1}{n}) \tag{2.3}$$

$$= S_K - \frac{K}{n}.$$

Mauldon (1951) derived the distribution of S_k, the sum of the k largest spacings for any fixed k. Treating this as the conditional density of S_K given K=k and using (2.1), we can write the joint density of (S_K, K) and hence obtain the density of U_n through the relation (2.3). The resulting expression for the density of U_n is very complex and attempts to show that this is identical to the density given for instance in Darling (1953), have not been successful.

3. Asymptotic null distribution of R_n.

In this section, we establish the asymptotic normality of R_n under the hypothesis of uniformity as well as under a suitable sequence of alternatives. Notice that for computing the Pitman Asymptotic Relative Efficiency (ARE) of R_n, which will be considered in the next section, it is enough to obtain the limiting distributions under a sequence of alternatives which converge to the hypothesis (see for instance Rao and Sethuraman (1975)). Hence we will specify the alternative hypotheses by a sequence of distribution functions $A_n(x)$ depending on n and converging to the uniform distribution, which corresponds to the null hypothesis. Under the alternative hypothesis, we specify the distribution function to be

$$A_n(x) = x + L_n(x) / n^{\frac{1}{4}}, \qquad 0 \leq x \leq 1 \tag{3.1}$$

where $L_n(0) = L_n(1) = 0$. We further assume that $L_n(x)$ is twice differentiable on $[0,1]$ and there is a function $L(x)$ which is twice continuously differentiable and such that $L(0) = L(1) = 0$, $n^{1/4} \sup_{0 \leq x \leq 1} | L_n''(x) - \ell'(x) | = o(1)$ where $\ell(x)$ and $\ell'(x)$ are the first and second derivatives of $L(x)$. This sequence of alternatives is smooth in a certain sense and has been considered before. See for instance Rao and Sethuraman (1975).

We define the empirical distribution function of the "normalized" spacings $\{nT_i, \ i = 1, \ldots n\}$ by

$$H_n(x) = \sum_{i=1}^{n} I(nT_i;x)/n \qquad \text{for } x \geq 0 \tag{3.2}$$

where
$$I(z;x) = \begin{cases} 1 & \text{if } z \leq x \\ 0 & \text{otherwise} \end{cases} .$$

Let

$$(3.3)\quad G_n(x) = 1 - e^{-x} + e^{-x}(x - x^2/2) \cdot \left(\int_0^1 \ell^2(p)\,dp\right)/\sqrt{n} \qquad \text{for } x \geq 0$$

and

$$\{\zeta_n(x) = \sqrt{n}\,(H_n(x) - G_n(x)), \quad x \geq 0\}.$$

$\zeta_n(\cdot)$ can be considered as a stochastic process with values in $D[0,\infty]$. See Rao and Sethuraman (1975) from which we have the following

Theorem 3.1 (Rao and Sethuraman (1975))

Under the alternatives (3.1), the sequence of stochastic processes $\{\zeta_n(x) = \sqrt{n}(H_n(x) - G_n(x)), x \geq 0\}$ converges weakly to the Gaussian process $\{\zeta(x), x \geq 0\}$ in $D[0,\infty]$ with mean function zero and covariance kernel

$$K(s,t) = e^{-t}(1 - e^{-s} - ste^{-s}) \quad \text{for} \quad 0 \leq s \leq t \leq \infty.$$

Moreover if $g(\cdot)$ is a real-valued measurable function on $D[0,\infty]$ which is a.e. continuous with respect to the probability measure induced by the Gaussian process $\{\zeta(x), x \geq 0\}$, then the distribution of the real-valued random variable $g(\zeta_n)$ converges weakly to that of $g(\zeta)$ as $n \to \infty$.

At this stage we will assume that δ_n is of the form $\delta_n = \delta/n$ for some $\delta > 0$. Since the individual T_i's are of order $1/n$ in probability under the hypothesis, for asymptotic purposes this would be the correct normalization. When $\delta_n = \delta/n$, we have the following theorem on $R_n = R_n(n\delta_n) = R_n(\delta)$.

Theorem 3.2.

Under the sequence of alternatives (3.1), $\sqrt{n}\left(\frac{R_n(\delta)}{n} - G_n(\delta)\right)$ where $G_n(x)$ is defined in (3.3), has a limiting $N(0,\sigma^2)$ distribution with $\sigma^2 = e^{-\delta}(1 - e^{-\delta} - \delta^2 e^{-\delta})$.

Proof: Note $R_n(\delta)$ = number of $(n T_i) \leq (n \delta_n) = \delta$

$$= n\, H_n(\delta)$$

where $H_n(x)$ is the empirical distribution of the normalised spacings and is defined in (3. 2). Thus

$$\sqrt{n}\,\left(\frac{R_n(\delta)}{n} - G_n(\delta)\right) = \sqrt{n}\,[H_n(\delta) - G_n(\delta)]$$

$$= \zeta_n(\delta) .$$

Therefore the stated result follows from Theorem 3. 1. q. e. d.

Corollary 3. 3.

Under the null hypothesis of uniformity

$$\sqrt{n}\,\left(\frac{R_n(\delta)}{n} - (1 - e^{-\delta})\right) / \{ e^{-\delta}(1 - e^{-\delta} - \delta^2 e^{-\delta})\}^{\frac{1}{2}}$$

has a limiting $N(0, 1)$ distribution.

This Corollary 3. 3 may also be obtained alternately using Theorem 9. 1 of Darling (1953). But unfortunately the expression for the limiting variance given there, is incorrect. We now state the correct version without proof. This result may also be obtained as a corollary from Theorem 3. 1 of Rao and Sethuraman (1975).

Theorem 3. 4.

Denote by $N_n(\frac{a}{n}, \frac{b}{n})$ the number of spacings whose length lies between $\frac{a}{n}$ and $\frac{b}{n}$. Then the random variable $N_n(\frac{a}{n}, \frac{b}{n})$ is asymptotically normally distributed with an asymptotic mean and variance given by

$$\mu_n = n(e^{-a} - e^{-b})$$

$$\sigma_n^2 = n[(e^{-a} - e^{-b}) - (e^{-a} - e^{-b})^2 - (a e^{-a} - b e^{-b})^2] .$$

4. The ARE of R_n relative to U_n.

For a definition of ARE, see Fraser (1957). The ARE of a test relative to another may be defined as the limit of the inverse ratio of sample sizes required to obtain the same limiting power at a sequence of alternatives converging to the null hypothesis. This limiting power should be a value in between the limiting size α and the maximum power 1, in order that it can give an insight into the power behaviour of the test. If this converges to a number in the interval $(\alpha, 1)$, then a measure of the rate of this convergence, called 'efficacy' can be computed. Under certain standard regularity assumptions (see e.g. Fraser (1957)) which includes a condition about the nature of alternatives and the asymptotic normality of the test statistic under the alternatives, which are satisfied here, the 'efficacy' is given by

$$\text{efficacy} = \left(\frac{\mu}{\sigma}\right)^4 \tag{4.1}$$

in this case. Here μ and σ^2 are the mean and variance of the limiting normal distribution under the sequence of alternatives (3.1) when the test-statistic has been normalized to have a limiting normal distribution with mean zero and finite variance under the hypothesis. In such a situation, the ARE of one test with respect to another is simply the ratio of their efficacies.

From Corollary 3.3, $\sqrt{n}\left(\frac{R_n}{n} - (1 - e^{-\delta})\right)$ has a limiting normal distribution with mean zero and variance $e^{-\delta}(1-e^{-\delta}-\delta^2 e^{-\delta})$ under H_0. On the other hand, from Theorem 3.2, under the sequence of alternatives (3.1) the same statistic has a limiting normal distribution with mean $\left(\int_0^1 \ell^2(p)\,dp\right) e^{-\delta}\left(\delta - \frac{\delta^2}{2}\right)$ and the same variance. Hence the efficacy of

$R_n(\delta)$ is given by

$$(4.2) \qquad \frac{\left(\int_0^1 \ell^2(p)\,dp\right)^4 \left(\delta - \frac{\delta^2}{2}\right)^4}{\left(e^\delta - 1 - \delta^2\right)^2} .$$

Sethuraman and Rao (1970) show that the Pitman efficacy of U_n in this situation is given by

$$\frac{\left(\int_0^1 \ell^2(p)\,dp\right)^4}{\left(4(2e-5)\right)^2} .$$

Hence the ARE of R_n with respect to U_n is given by

$$(4.3) \qquad \frac{16(2e-5)^2\left(\delta - \frac{\delta^2}{2}\right)^4}{\left(e^\delta - 1 - \delta^2\right)^2} .$$

From the results of Rao and Sethuraman (1970) who compute the efficacies of many other spacings tests, one can compare the ARE of R_n with any of those tests. We will return to the expression (4.2) again in Section 6.

5. Limiting Bahadur Efficiency of R_n relative to U_n:

We refer the reader to Bahadur (1960) for the concepts of Bahadur Approximate slope (BAS) and Bahadur Approximate Efficiency (BAE). We use the same notations as in Bahadur (1960). We consider the class of alternative densities

$$(5.1) \qquad g_k(x) = 1 + k\,\ell(x), \qquad 0 \le x \le 1$$

where k is a real number and $\ell(x)$ is any square integrable function on $[0, 1]$ with $\int_0^1 \ell(x)\, dx = 0$. For instance in connection with the circle, taking $\ell(x) = \cos 2\pi x$ yields the so called cardioid curve. Here k is a scale parameter and since uniformity corresponds to $k = 0$, the null hypothesis formulates H_0: $k = 0$. These alternatives are very similar to those formulated earlier in (3.1). We now take as the standard sequence

$$(5.2) \qquad T_n^{(1)} = \left(\frac{R_n(\delta)}{n} - (1 - e^{-\delta})\right) / \{ e^{-\delta}(1-e^{-\delta} - \delta^2 e^{-\delta}) \}^{\frac{1}{2}} .$$

Since $T_n^{(1)}$ has a $N(0, 1)$ distribution asymptotically from Corollary 3.3, this sequence of test statistics satisfies conditions (1), (2) and (3) on p. 276 of Bahadur (1960) with $a = 1$. To find the probability limit of $T_n^{(1)}/\sqrt{n}$, we state a result from Rao (1969).

<u>Theorem 5.1</u> (Rao (1969)).

Under the alternative distribution $G(x)$ on $[0, 1]$ with continuous density $g(\cdot)$, the statistic $H_n(a)$ defined in (3.2) converges in probability to $1 - \int_0^1 \exp(-a\, g(u)\, dG(u)$.

Thus under the alternative (5.1)

$$(5.3) \qquad \frac{R_n}{n} = H_n(\delta) \xrightarrow{Pr} 1 - \int_0^1 \exp(-\delta g_k(u))\, g_k(u)\, du$$

$$= 1 - e^{-\delta} \int_0^1 e^{-\delta k \ell(u)} (1 + k\ell(u))\, du .$$

As in Rao (1972), the comparison of the limiting efficiencies is made easier by considering approximations to the slopes when k is small, since in any case we let $k \to 0$ for obtaining the limiting efficiencies. Thus for k small, by expanding the exponential function in a power series and noting that $\int_0^1 \ell(x)\,dx = 0$, the probability limit in (5.3) can be shown to be

$$1 - e^{-\delta}[1 + k^2(\int_0^1 \ell^2(x)\,dx)(\frac{\delta^2}{2} - \delta) + o(k^2)] .$$

Hence the BAS of $T_n^{(1)}$ is given by

(5.4) $$C_1(k) = (\frac{\delta^2}{2} - \delta)^2 \cdot k^4 \cdot (\int_0^1 \ell^2(x)\,dx)^2 / 4(e^{\delta}-1-\delta^2) .$$

on the otherhand, similar calculations yield the BAS of the standardized U_n to be

$$C_2(k) = k^4(\int_0^1 \ell^2(x)\,dx)^2 / 8(2e-5) .$$

Thus the limiting Bahadur efficiency of R_n relative to U_n is

(5.5) $$\lim_{k \to 0} \frac{C_1(k)}{C_2(k)} = 4(\delta - \frac{\delta^2}{2})^2 (2e-5)/(e^{\delta} - 1 - \delta^2)$$

This value, it may be noted, is the square root of the ARE derived in (4.3).

6. The statistic R_n^* and a table of significance points.

In this section, we consider the class of tests $\{R_n(\delta)\}$ for varying δ and select the one with maximum efficacy. This amounts to finding out the value of δ for which the expression (4.2) (or equivalently (5.4)) is a maximum. The mathematical problem of finding the maximum does not appear simple but using a computer, it may be checked that the maximum efficiency is attained close to a value of $\delta = 0.7379$. For example, it may be seen that the efficiency of $R_n(1)$ relative to $R_n(0.7379)$ is close to 86% . Thus if one were to restrict consideration to this class of tests, then it is clearly best to take $\delta = 0.7379$. But from a practical point of view, we suggest using a more reasonable fraction like $\delta = 0.75$. Since the loss of efficiency in doing this is insignificant, we advocate the use of the statistic

$$(6.1) \qquad R_n^* = R_n(0.75) = \{\text{number of } T_i\text{'s} \le \frac{3}{4n}\}$$

as the best among this class. From Theorem 2.1, Corollary 3.3 and equation (4.2), we have the following result regarding the exact, asymptotic distributions of R_n^* as well as its efficacy .

Corollary 6.1:

The following results hold for the statistic R_n^* defined in (6.1):

(a) Exact null distribution:

$$(6.2) \qquad P(R_n^* = k) = \binom{n}{k} \sum_{j=0}^{k} (-1)^j \binom{k}{j} \left\langle 0.25 + (0.75)\left(\frac{k-j}{n}\right)\right\rangle^{n-1} \quad \text{for } k = 0, 1, \ldots, (n-1).$$

$$= 0 \qquad \text{otherwise}.$$

with the notation $\langle x \rangle = x$ if $x > 0$ and $= 0$ if $x \le 0$.

(b) Asymptotic Null Distribution:

$\sqrt{n}\left(\frac{R_n^*}{n} - 0.5276\right) / \{0.3517\}$ has a limiting $N(0,1)$ distribution.

(c) Pitman efficiency:

The Pitman efficacy of R_n^* against the alternatives (3.1) is given by $(0.1570)\left(\int_0^1 \ell^2(p)\,dp\right)^4$.

Using the exact null distribution of R_n^* given in (6.2), the following table of cumulative probability function $F(k) = P(R_n^* \leq k)$ in the upper tail area has been constructed for sample sizes $n = 3(1)\,10(5)\,100$. if the observed value k of R_n^* is such that $F(k)$ (from Table 6.1) exceeds $(1-\alpha)$, then we reject the hypothesis of uniformity H_0 at that level α.

Table 6.1. Distribution Function of the Statistic R_n^* in the range of 0.90 to 1.00.

n	k	F(k)	F(k+1)	F(k+2)	F(k+3)	F(k+4)	F(k+5)
3	1	.6350	1.0000				
4	2	.8418	1.0000				
5	2	.5545	.9392	1.0000			
6	3	.7583	.9780	1.0000			
7	4	.8818	.9923				
8	4	.7030	.9465	.9974			
9	5	.8334	.9772	.9991			
10	5	.6621	.9134	.9907			
15	8	.7280	.9142	.9841	.9985		
20	11	.7752	.9207	.9810	.9971		
25	14	.8112	.9286	.9801	.9960		
30	17	.8399	.9364	.9804	.9954		
35	20	.8632	.9436	.9813	.9951		
40	23	.8825	.9501	.9825	.9950		
45	26	.8986	.9559	.9838	.9950		
50	28	.8281	.9121	.9611	.9852	.9952	
55	31	.8514	.9237	.9657	.9865	.9954	
60	34	.8710	.9336	.9697	.9878	.9957	
65	37	.8878	.9421	.9733	.9890	.9960	
70	39	.8293	.9023	.9494	.9764	.9901	
75	42	.8505	.9147	.9557	.9791	.9911	
80	45	.8689	.9254	.9611	.9816	.9920	
85	48	.8849	.9346	.9659	.9837	.9929	
90	51	.8988	.9427	.9700	.9856	.9936	
95	53	.8538	.9110	.9497	.9736	.9872	.9943
100	56	.8706	.9216	.9558	.9768	.9887	.9949

It may be remarked here that the data need not be scaled to the interval (0, 1) in order to calculate R_n^*. We now illustrate by means of a numerical example, the extreme simplicity in using the statistic R_n^* for testing uniformity. It may be remarked here that the simplicity in using R_n^* in our view, more than compensates for the lower asymptotic efficiency.

Example

Consider a fire station which received 20 calls on a particular day. We want to know if these calls are randomly distributed over the entire day or if they tend to cluster around some particular time of the day. Suppose that the calls are received at 1:00, 4:30, 6:00, 6:10, 7:00, 8:00, 8:30, 8:45, 9:30, 10:05 a.m. and 1:00, 2:10, 4:00, 5:50, 7:30, 9:15, 10:00, 10:15, 11:00, 11:30 p.m. Since $\delta_n = (0.75)\ 24/20 = 0.9$ hrs. = 54 mts., R_n^* is the number of spacings less than 54 minutes. We see easily that $R_n^* = 10$. This R_n^* value of 10, when n = 20, is not significant even at $\alpha = 10\%$ as can be seen from Table 6.1. Hence we have no reason to reject the hypothesis that these calls are randomly distributed throughout the day. We may remark here that for the purpose of this test the data could very well be accumulated over several cycles (days) instead of just one.

BIBLIOGRAPHY

Bahadur, R.R. (1960): Stochastic Comparison of tests, Ann. Math. Statist., 31, 276-95.

Darling, D.A. (1953): On a class of problems related to the random division of an interval, Ann. Math. Statist., 24, 239-253.

Feller, W. (1966): An introduction to the probability theory and its applications, Vol. I, John Wiley, New York.

Fraser, D.A.S. (1957): Nonparametric methods in statistics, John Wiley, New York.

Mardia, K.V. (1972): Statistics of directional data, Academic Press.

Mauldon, J.G. (1951): Random division of an interval, Proc. Camb. Phil. Soc., 47, 331-336.

Pyke, R. (1965): Spacings, J. Roy. Stat. Soc. (Ser. B) 27, 395-449.

Rao, J.S. (1969): Some contributions to the analysis of circular data, Unpublished Ph.D. thesis, Indian Statistical Institute, Calcutta.

Rao, J.S. (1972): Bahadur efficiencies of some tests for uniformity on the circle, Ann. Math. Statist., 43, 468-479.

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: In Nonparametric Techniques in Statistical Inference, Ed. M.L. Puri, Cambridge Univ. Press, 405-415.

Sherman, B. (1950): A random variable related to the spacing of sample values, Ann. Math. Statist., 21, 339-51.

Wilks, S.S. (1962): Mathematical Statistics, John Wiley, New York.