Metrika, Volume 35, 1988, page 223–232

A Test of Goodness-of-Fit Based on Extreme Spacings with some Efficiency Comparisons

By S. Rao Jammalamadaka[1] and M. T. Wells[2]

Abstract: Tests for the goodness-of-fit problem based on sample spacings, i.e., observed distances between successive order statistics, have been used in the literature. We propose a new test based on the number of "small" and "large" spacings. The asymptotic theory under close alternative sequences is also given thus enabling one to calculate the asymptotic relative efficiencies of such tests. A comparison of the new test and other spacings tests is given.

1 Introduction

Let $X_1, \ldots, X_{n-1}$ be independently and identically distributed random variables with common distribution function (d.f.). The goodness of fit problem is to test if this d.f. is equal to a specified one. A probability integral transformation on the sample would permit us to test whether the data is uniformly distributed 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 $[0, 1]$.

Let $0 \leqslant X_{(1)} \leqslant X_{(2)} \ldots \leqslant X_{(n-1)} \leqslant 1$ be the order statistics. Define the sample spacings by

$$T_i = X_{(i)} - X_{(i-1)} \qquad i = 1, \ldots, n$$

where $X_{(0)} = 0$ and $X_{(n)} = 1$. Tests for uniformity based on spacings, have been proposed by several authors, see Pyke (1965) or Rao and Sethuraman (1975) and the references contained therein.

[1] S. Rao Jammalamadaka, Statistics and Applied Probability Program, University of California, Santa Barbara, CA 93106, USA.

[2] Martin T. Wells, Department of Economics and Social Statistics, Cornell University, Ithaca, NY 14850, USA.

0026–1335/88/3–4/223–232 $2.50 © 1988 Physica-Verlag, Heidelberg

In analyzing 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 an important problem. This is a necessary preliminary step before making inferences on the mean direction. For purposes of inference on the circle, one requires a statistic that is invariant under cyclical permutations of its arguments. Spacings form a maximal invariant for this problem. For instance, functions symmetric in all the arguments may be considered though they are not asymptotically efficient. See Sethuraman and Rao (1970). Thus spacings play an important role in testing goodness-of-fit on the circle.

This paper is an extension of the results of Puri, Rao and Yoon (1979). They discussed tests based on small spacings, while we derive tests which use both the small and the large spacings, i.e., the number of "extreme" spacings. As one would expect, it turns out that the later tests are much more efficient. In Section 2 we discuss the exact distribution of our statistics under the hypothesis of uniformity. Section 3 deals with the asymptotic distribution theory under a sequence of close alternatives. In Section 4 the asymptotic relative efficiency (ARE) of the proposed statistic is computed and comparisons made with other spacings statistics.

2 The Statistic $N(\alpha_n, \beta_n)$ and its Exact Distribution

Choose and fix $0 < \alpha_n < \beta_n < 1$. Define

$$N(\alpha_n, \beta_n) = \sum_{i=1}^{n} \{I(T_i \leqslant \alpha_n) + I(T_i \geqslant \beta_n)\} \tag{2.1}$$

where $I(A)$ is the indicator function of the event A. The test criterion is to reject H_0 if too many of the spacings fall outside the interval (α_n, β_n). At this stage we will leave the choice of α_n and β_n open. If $\alpha_n = \alpha$ and $\beta_n = \beta$, then the exact distribution of $N(\alpha, \beta)$ is given by the following

Theorem 2.1: Under the hypothesis of uniformity, the probability mass function of $N(\alpha, \beta)$ is given by

$$P(N(\alpha, \beta) = k) = \binom{n}{n-k} \sum_{j=0}^{k} \sum_{l=0}^{n-k+j} \binom{k}{j} \binom{n-j}{l} (-1)^{k-j+l}$$

$$< 1 + \alpha(n-j-l) + \beta l >^{n-1} \tag{2.2}$$

for $k = 0, \ldots, n$ with the notation $\langle x \rangle = x$ if $x > 0$ and $= 0$ if $x \leqslant 0$. In (2.2) we use the convention that $\binom{k}{n} = 0$ if $k < n$.

Proof: The characteristic function of $n - N(\alpha, \beta) = \bar{N}(\alpha, \beta)$ as given by Darling (1953) is

$$E(e^{it\bar{N}(\alpha, \beta)}) = \frac{(n-1)!}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{z} z^{-n} \{1 + (e^{it} - 1)(e^{-z\alpha} - e^{-z\beta})\}^{n} dz \tag{2.3}$$

where Re $z = c > 0$ and the path of integration is the straightline Re $z = c$. The term in the braces is equal to

$$\{e^{it}(e^{-z\alpha} - e^{-z\beta}) + (1 - (e^{-z\alpha} - e^{-z\beta})]\}^{n}$$

$$= \sum_{j=0}^{n} \binom{n}{j} [e^{it}(e^{-z\alpha} - e^{-z\beta})]^{j} \cdot [1 - (e^{-z\alpha} - e^{-z\beta})]^{n-j}.$$

Then for any fixed $k = 0, 1, \ldots, n$ the coefficient of e^{itk} is

$$\binom{n}{k} [e^{-z\alpha} - e^{-z\beta}]^{k} [1 - (e^{-z\alpha} - e^{-z\beta})]^{n-k}$$

$$= \binom{n}{k} [e^{-z\alpha} - e^{-z\beta}]^{k} \sum_{j=0}^{n-k} \binom{n-k}{j} (-1)^{n-j-k} (e^{-z\alpha} - e^{-z\beta})^{n-k-j}$$

$$= \binom{n}{k} \sum_{j=0}^{n-k} \binom{n-k}{j} (-1)^{n-j-k} (e^{-z\alpha} - e^{-z\beta})^{n-j}$$

$$= \binom{n}{k} \sum_{j=0}^{n-k} \sum_{l=0}^{n-j} \binom{n-k}{j} \binom{n-j}{j} (-1)^{n-j-k+l} e^{-z[\alpha(n-j-l) + \beta l]}.$$

Then

$$P(\bar{N}(\alpha, \beta) = k)$$

$$= \frac{(n-1)!}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^z z^{-n} \left\{ \binom{n}{k} \sum_{j=0}^{n-k} \sum_{l=0}^{n-j} \binom{n-k}{j} \binom{n-j}{l} \right.$$

$$\left. (-1)^{n-j-k+l} e^{-z[\alpha(n-j-l)+\beta l]} \right\} dz$$

$$= \binom{n}{k} \sum_{j=0}^{n-k} \sum_{l=0}^{n-j} \binom{n-k}{j} \binom{n-j}{l} (-1)^{n-j-k+l} \frac{(n-1)!}{2\pi i}$$

$$\int_{c-i\infty}^{c+i\infty} z^{-n} e^{z\{1-\alpha(n-j-l)+\beta l\}} dz$$

$$= \binom{n}{k} \sum_{j=0}^{n-k} \sum_{l=0}^{n-j} \binom{n-k}{j} \binom{n-j}{l} (-1)^{n-j-k+l} < 1 - \alpha(n-j-l) + \beta l >^{n-1}.$$

The last equality follows from the Residue Theorem. Now replace k by $n - k$ and the result follows.

3 Asymptotic Distribution of $N(\alpha_n, \beta_n)$

In this section, we establish the asymptotic normality of $N(\alpha_n, \beta_n)$ under the hypothesis of uniformity as well as under a suitable sequence of alternatives. To compute the Pitman Asymptotic Relative Efficiency (ARE) of $N(\alpha_n, \beta_n)$, in the next section, it will be sufficient to obtain the limiting distributions under a sequence of alternatives which converge to uniformity (see Rao and Sethuraman 1975). Under the alternative hypothesis, we specify the distribution to be

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

where $L_n(0) = L_n(1) = 0$. Further assume that $L_n(x)$ is twice differentiable on $[0, 1]$ and there is a function $L(x)$ which is twice differentiable with $L(0) = L(1) = 0$,

$n^{1/4} \sup_{0 \leqslant x \leqslant 1} |L_n''(x) - l'(x)| = 0(1)$, where $l(x)$ and $l'(x)$ are the first and second derivatives of $L(x)$. This type of alternatives have been considered, for instance, in Rao and Sethuraman (1975).

Define the empirical distribution function of the "normalized" spacings $\{nT_i$: $i = 1, 2, \ldots, n\}$ by

$$H_n(x) = \sum_{i=1}^{n} I(nT_i \leqslant x)/n, \quad x \geqslant 0. \tag{3.2}$$

Also, let

$$G_n(x) = 1 - e^{-x} + e^{-x}\left(x - \frac{x^2}{2}\right) \cdot \int_0^1 l^2(p)dp/\sqrt{n}, \quad x \geqslant 0 \tag{3.3}$$

and

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

Then, $\zeta_n(\cdot)$ can be considered as a stochastic process with values in $D[0, \infty)$.

Theorem 3.1 (Rao and Sethuraman 1975): Under the sequence of alternatives (3.1), the sequence of stochastic processes

$$\{\zeta_n(x) = \sqrt{n}(H_n(x) - G_n(x)); x \geqslant 0\}$$

converges weakly to the Gaussian process $\{\zeta(x); x \geqslant 0\}$ in $D[0, \infty)$ with mean function zero and covariance kernel

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

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 $\{\zeta(x)$: $x \geqslant 0\}$, then by the Invariance Principle, the distribution of $g(\zeta_n(x))$ converges weakly to that of $g(\zeta(x))$ as $n \to \infty$.

At this point we assume α_n and β_n are of the form $\alpha_n = \frac{a}{n}$ and $\beta_n = \frac{b}{n}$ for some $a, b > 0$.

Theorem 3.2: Under the sequence of alternatives (3.1),

$$\sqrt{n}\left\{\frac{1}{n}N\left(\frac{a}{n},\frac{b}{n}\right)-(1-G_n(b)+G_n(a))\right\},$$

where $G_n(\cdot)$ is defined in (3.3), has a limiting $N(0,\sigma^2)$ distribution with $\sigma^2=(e^{-a}-e^{-b})-(e^{-a}-e^{-b})^2-(ae^{-a}-be^{-b})^2$.

Proof: Note that

$$N\left(\frac{a}{n},\frac{b}{n}\right)=\sum_{i=1}^{n}\{I(nT_i\leqslant a\}+I(nT_i\geqslant b)\}=n(1-H_n(b)+H_n(a)).$$

Thus

$$\sqrt{n}\left\{\frac{1}{n}N\left(\frac{a}{n},\frac{b}{n}\right)-(1-G_n(b)+G_n(a))\right\}$$

$$=\sqrt{n}\left\{1-H_n(b)+H_n(a)-(1-G_n(b)+G_n(a))\right\}$$

$$=\sqrt{n}\,\{H_n(a)-G_n(a)\}-\sqrt{n}\,\{H_n(b)-G_n(b)\}=\{\zeta_n(a)-\zeta_n(b)\}.$$

Therefore, by Theorem 3.1

$$\{\zeta_n(a)-\zeta_n(b)\}\overset{D}{\to}N(0,\sigma^2)$$

where

$$\sigma^2=k(a,a)+k(b,b)-2k(a,b)$$

$$=e^{-a}(1-e^{-a}-a^2e^{-a})+e^{-b}(1-e^{-b}-b^2e^{-b})-2e^{-b}(1-e^{-a}-abe^{-a})$$

$$=(e^{-a}-e^{-b})-(e^{-a}-e^{-b})^2-(ae^{-a}-be^{-b})^2.$$

Note that

$$(1 - G_n(b) + G_n(a))$$

$$= 1 - e^{-a} + e^{-b} + \left[e^{-a}\left(a - \frac{a^2}{2}\right) - e^{-b}\left(b - \frac{b^2}{2}\right) \right] \int_0^1 l^2(p)dp/\sqrt{n}.$$

Corollary 3.3: Under the null hypothesis of uniformity $\sqrt{n}\left\{ \dfrac{N\left(\frac{a}{n}, \frac{b}{n}\right)}{n} - (1 - e^{-a} + e^{-b}) \right\}$ has limiting $N(0, \sigma^2)$ distribution with

$$\sigma^2 = (e^{-a} - e^{-b}) - (e^{-a} - e^{-b})^2 - (ae^{-a} - be^{-b})^2.$$

This corollary is stated in Puri, Rao and Yoon (1979). This is also Theorem 8.1 of Darling (1953) where unfortunately the expression for the limiting variance is incorrect. See Darling (1962) for the correction.

As special cases of Theorem 3.2, we can get the results of Puri, Rao, and Yoon (1979) by letting $a = 0$ and $b = \delta$. There the interest was in a test of uniformity based on the number of "small" spacings. We will call this statistic $R_n(\delta)$. Alternatively, we could look only at the number of "large" spacings by letting $b = \infty$. A special case of interest to us is $N_n\left(\frac{1-c}{n}, \frac{1+c}{n}\right)$. Its complement, i.e., $n - N_n\left(\frac{1-c}{n}, \frac{1+c}{n}\right)$, can be expressed as

$$S_n(c) = \sum_{i=1}^{n} I\left(\left|T_i - \frac{1}{n}\right| < \frac{c}{n}\right), \tag{3.4}$$

which counts the number of spacings which are within $\frac{c}{n}$ of the average (expected) length of a spacing, namely $\frac{1}{n}$.

4 The ARE of $N\left(\frac{a}{n}, \frac{b}{n}\right)$

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 power at a sequence of alternatives converging to the null hypothesis. The limiting power should be a value between the limiting size α and the maximum power 1, in order that it can give information about the power of the test. When this converges to a number in $(\alpha, 1)$, then a measure of the rate of this convergence, called the "efficacy" can be computed. Under certain standard regularity assumptions (see Fraser 1957), which are satisfied here, the efficacy is given by

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

Here μ and σ are the mean and standard deviation of the limiting distribution under the sequence of alternatives (3.1) when the test statistic has been normalized to have a limiting $N(0, 1)$ under the hypothesis. Define the ARE of two statistics T_1 with respect to T_2 as

$$\text{ARE}\,(T_1, T_2) = \frac{\text{eff}\,(T_1)}{\text{eff}\,(T_2)}. \tag{4.2}$$

From Corollary 3.3, $\sqrt{n}\left\{\frac{N\left(\frac{a}{n}, \frac{b}{n}\right)}{n} - (1 - e^{-a} + e^{-b})\right\} \xrightarrow{d} N(0, \sigma^2)$ where $\sigma^2 = (e^{-a} - e^{-b}) - (e^{-a} - e^{-b})^2 - (ae^{-a} - be^{-b})^2$ under H_0. However from Theorem 3.2, under the sequence of alternatives (3.1) the same statistic has a limiting normal distribution with asymptotic mean

$$\left[e^{-a}\left(a - \frac{a^2}{2}\right) - e^{-b}\left(b - \frac{b^2}{2}\right)\right] \int_0^1 l^2(p)dp$$

and the same variance. Hence the efficacy of $N\left(\frac{a}{n},\frac{b}{n}\right)$ is given by

$$\frac{\left[e^{-a}\left(a-\frac{a^2}{2}\right)-e^{-b}\left(b-\frac{b^2}{2}\right)\right]^4\left(\int_0^1 l^2(p)dp\right)^4}{[(e^{-a}-e^{-b})-(e^{-a}-e^{-b})^2-(ae^{-a}-be^{-b})^2]^2}.$$

As a special case, if we let $a = 0$ and $b = \delta$ (Puri, Rao, and Yoon 1979), the efficacy of the statistic $R_n(\delta)$ is

$$\operatorname{eff}(R_n(\delta)) = \frac{\left(\delta-\frac{\delta^2}{2}\right)^4\left(\int_0^1 l^2(p)dp\right)^4}{[e^{\delta}-1-\delta^2]^2}. \tag{4.4}$$

Also, if we set $a = 1 - c$ and $b = 1 + c$, the efficacy of the statistic $S_n(c)$ is given by

$$\operatorname{eff}(S_n(c)) = \frac{\left(\frac{c^2-1}{2}\right)^4\left(\int_0^1 l^2(p)dp\right)^4}{\left\{\frac{c}{2}\,c\operatorname{sch} c-1-[1-c\coth c]^2\right\}^2}. \tag{4.5}$$

Sethuraman and Rao (1970) show that the asymptotically most powerful test for the alternatives of the form (3.1) based on symmetric functions of spacings is (called the Greenwood statistic)

$$V_1 = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(nT_j)^2. \tag{4.6}$$

They show that $\operatorname{eff}(V_1) = \left(\int_0^1 l^2(p)dp\right)^4$. Hence the AREs of $N\left(\frac{a}{n},\frac{b}{n}\right)$, $R_n(\delta)$, and $S_n(c)$ can be found by dividing the efficacies of these statistics by $\left(\int_0^1 l^2(p)dp\right)^4$.

We now consider the class of tests $N\left(\frac{a}{n}, \frac{b}{n}\right)$ for varying a and b and select values which maximize (4.3). This is not an easy problem since the expression is quite complicated. It may be checked by computer that the maximum of (4.3) is attained if $a = 0.7355$ and $b = 4.3205$. Puri, Rao, and Yoon (1979) show that $\delta = 0.7379$ maximizes (4.4). To maximize (4.5) we choose $c = 0.6254$. The asymptotic relative efficiencies of $N_n\left(\frac{0.7355}{n}, \frac{4.3205}{n}\right)$, $S_n(0.6254)$ and $R_n(0.7379)$ with respect to the Greenwood statistic V_1 are 0.7941, 0.2511, and 0.1570 respectively. In spite of the simplicity of $N_n\left(\frac{0.7355}{n}, \frac{4.3205}{n}\right)$, it is worth noting that it has an efficiency of close to 80% with respect to V_1, which is known to be the most efficient among the class of symmetric tests.

References

Darling DA (1953) On a class of problems related to the random division of an interval. Ann Math Stat 24:239–253

Darling DA (1962) Correction to "On a class of problems related to the random division of an interval. Ann Math Stat 33:812

Fraser DAS (1957) Nonparametric methods in statistics. John Wiley, New York

Puri ML, Rao JS, Yoon Y (1979) A simple test for goodness of fit based on spacings with some efficiency comparisons. In: Jurecková J (ed) Contributions to statistics. Academia, Prague, pp 197–209

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

Rao JS, Sethuraman J (1975) Weak convergence of the empirical distribution function of random variables subject to perturbations and scale factors. Ann Stat 3:299–313

Sethuraman J, Rao JS (1970) Pitman efficiencies of tests based on spacings. In: Puri ML (ed) Nonparametric techniques in statistical inference. Cambridge University Press, pp 405–415

Received September 30, 1987