Journal of Statistical Planning and Inference 12 (1985) 295-304
North-Holland

ASYMPTOTIC COMPARISON OF THREE TESTS FOR GOODNESS OF FIT

S. Rao JAMMALAMADAKA

Department of Mathematics, University of California, Santa Barbara, Santa Barbara, CA 93106, USA

Ram C. TIWARI

Department of Mathematics, Indian Institute of Technology, Bombay, Powai, Bombay 400 076, India

Received 14 January 1985
Recommended by M.L. Puri

Abstract: The so-called Greenwood statistic, based on the sum of squares of the sample spacings, is known to be locally most powerful (LMP) among all tests based symmetrically on the sample spacings. On the other hand, the χ^2 criterion with the number of cells equal to the number of observations, is also known to be LMP among tests based symmetrically on the observed frequencies. While the latter compares the observed and expected frequencies holding the expected number in each cell to one, the former compares the expected and observed cell-lengths holding the observed number in each cell to one. We compare here these two test statistics with still another spacings test, $\sum D_i \log D_i$, on the basis of their asymptotic relative efficiency and conclude that the Greenwood statistic is superior.

AMS Subject Classification: Primary 62G20; Secondary 62E20.

Key words: Goodness of fit; Sample spacings; Greenwood statistic; χ^2-test; Asymptotic efficiency.

1. Introduction

Let $X_1, \ldots, X_{n-1}$ be independent random variables with a common distribution F. The goodness-of-fit problem is to test the hypothesis that F is equal to a specified distribution. A simple probability integral transformation on the random variables permits us to equate the specified distribution 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].

Let $0 \le X_1' \le \cdots \le X_{n-1}' \le 1$ be the order statistics. The sample spacings $(D_1, \ldots, D_n)$ are defined by

$$D_i = X_i' - X_{i-1}', \quad i = 1, \ldots, n, \tag{1.1}$$

where we put $X_0'=0$ and $X_n'=1$. Clearly the support of the distribution must be [0, 1] in order that this definition of the sample spacings is meaningful. Tests of the goodness-of-fit problem based on the normalized spacings $\{nD_i;\ i=1,\ldots,n\}$ have been proposed by several authors. See, for instance, Pyke (1965), Kale (1969), Sethuraman and Rao (1970), and Rao and Sethuraman (1975). More common among these are tests based symmetrically on spacings, namely, of the form

$$T_n=\frac{1}{n}\sum_{i=1}^{n} h(nD_i), \tag{1.2}$$

where, for instance, we may take $h(x)=x^r$ $(r>-\frac{1}{2})$, $\frac{1}{2}|x-1|$, and $\log x$. To compute the Pitman asymptotic relative efficiencies (ARE's) of various tests of the form (1.2), Sethuraman and Rao (1970) consider a sequence of alternative distributions with densities [see also Cibisov (1961), Weiss (1965)]

$$f_n(x)=1+\frac{l(x)}{n^{\delta}},\quad 0\le x\le 1,\ \delta>\tfrac{1}{4}, \tag{1.3}$$

converging to the density of the uniform distribution on [0, 1],

$$\mathrm{H}_0:\quad f(x)=1,\quad 0\le x\le 1, \tag{1.4}$$

under the hypothesis. Here $l(\cdot)$ is assumed to be square integrable and continuously differentiable on [0, 1]. Sethuraman and Rao (1970) demonstrate that the symmetric spacings tests cannot discriminate alternatives (1.3) if $\delta>\frac{1}{4}$ so that comparison of the ARE's must be made for a sequence of alternatives

$$\mathrm{A}_n:\quad f_n(x)=1+\frac{l(x)}{n^{1/4}},\quad 0\le x\le 1, \tag{1.5}$$

converging to the hypothesis (1.4) at the rate of $n^{-1/4}$. They also demonstrate that among a wide class of such tests, the Greenwood test statistic, namely,

$$V_2(n)=\frac{1}{n}\sum_{i=1}^{n}(nD_i)^2, \tag{1.6}$$

has maximum efficacy [see also Kuo and Rao (1984)]. In this paper, we also consider the spacings test based on the entropy

$$E_n=\frac{1}{n}\sum_{i=1}^{n}(nD_i)\log(nD_i) \tag{1.7}$$

and establish its asymptotic normality both under the hypothesis (1.4) and the alternatives (1.5) using the results of Sethuraman and Rao (1970). The asymptotic local power of E_n under the alternatives (1.5) has also been discussed by Gebert and Kale (1969) who use results of Weiss (1965). But the technique used by Weiss (1965), namely, substituting the spacings under the alternatives by uniform spacings with the scaling factor $f_n[F_n^{-1}(i/n+1)]$, is questionable in view of Pyke (1965, p. 417).

Thus, our results for E_n, while may be derived from those of Gebert and Kale (1969), use a different and a simpler approach and are on firm ground.

The third statistic we discuss is the usual χ^2-statistic with number of cells equal to the number of observations, i.e., with cell expectations of one each. This χ^2-statistic is thus

$$S_n = \sum_{i=1}^{n-1} (O_i - 1)^2, \tag{1.8}$$

where O_i is the observed frequency in the i-th cell, viz., $[(i-1)/(n-1), i/(n-1))$, $i = 1, \ldots, n-1$. It is of interest to compare this with the Greenwood test $V_2(n)$ in (1.6) since the latter, written in the equivalent form

$$n\left[\sum_{i=1}^{n} \left(D_i - \frac{1}{n}\right)^2\right], \tag{1.9}$$

may be thought of being analogous to (1.8). While the chi-square criterion S_n in (1.8) compares the observed and expected frequencies holding the expected frequencies in each cell to one, the Greenwood test in the form (1.9) compares the expected and observed cell-lengths holding the observed frequency in each cell to one. Under the alternatives (1.5), the asymptotic normality of these statistics is proved in Section 2 and the comparisons of asymptotic efficiencies are made in Section 3.

2. Asymptotic normality of E_n, $V_2(n)$ and S_n

The limiting distribution of various spacings statistics of the form (1.2) have been derived by Rao and Sethuraman (1975) both under the hypothesis and the alternatives using the ideas of the weak convergence of the empirical process of the normalized spacings. We state two results from this paper that we use to derive the limiting distribution of E_n.

Define the empirical distribution of the normalized spacings $\{nD_i; i = 1, \ldots, n\}$ by

$$H_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(nD_i; x), \quad x \geq 0, \tag{2.1}$$

where $I(z;x)$ is 1 if $z \leq x$ and 0 if $z > x$.

Let

$$G_n(x) = \begin{cases} (1 - e^{-x}) & \text{if } \delta > \frac{1}{4}, \\ (1 - e^{-x}) + \left(\frac{1}{\sqrt{n}} \int_0^1 l^2(t)\, dt\right) e^{-x}(x - \frac{1}{2}x^2) & \text{if } \delta = \frac{1}{4}. \end{cases} \tag{2.2}$$

Then we have:

Theorem 2.1 (Sethuraman and Rao (1970), Rao and Sethuraman (1975)). *Under the alternatives* (1.3), *the sequence of stochastic processes* $\{\varrho_n(x) = \sqrt{n}(H_n(x) - G_n(x)),$

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

$$K(x, y) = e^{-y}(1 - e^{-x} - xye^{-x}), \quad 0 \leq x \leq y \leq \infty. \tag{2.3}$$

This theorem on the empirical distribution of the normalized spacings forms the basic result for deriving the asymptotic distributions of test statistics of the form (1.2).

If $h(\cdot)$ is a function such that for $y \in D[0, \infty]$, the mapping $y \to \int_{\infty}^{0} h(x)\, dy(x)$ is continuous with probability one under ϱ, then

$$T_n = \frac{1}{n} \sum_{i=1}^{n} h(nD_i) = \int_0^{\infty} h(x)\, dH_n(x) \tag{2.4}$$

can be considered as a continuous functional of the proces ϱ_n, and we have, as a consequence of the invariance principle:

Theorem 2.2 (Sethuraman and Rao (1970), Rao and Sethuraman (1975)). *Under the sequence of alternatives* (1.3), *the random variable*

$$T_n^* = \sqrt{n}\left(T_n - \int_0^{\infty} h(x)\, dG_n(x)\right) \tag{2.5}$$

has a limiting normal distribution with mean zero and variance

$$\sigma^2 = \int_0^{\infty} \int_0^{\infty} h'(x)h'(y)K(x, y)\, dx\, dy. \tag{2.6}$$

As a consequence of this theorem, the following two results on the distribution of E_n and $V_2(n)$ follow:

Theorem 2.3. *Under the sequence of alternatives* (1.5) *the random variable*

$$E_n^* = \sqrt{n}(E_n - \mu_n) \tag{2.7}$$

has a limiting normal distribution with mean zero and variance $(\frac{1}{3}\pi^2 - 3)$ *where* μ_n *is given by*

$$\mu_n = (1 - \gamma) + \frac{1}{2\sqrt{n}} \int_0^1 l^2(t)\, dt \tag{2.8}$$

and γ *is Euler's constant.*

Proof. Observe that E_n may be written as

$$E_n = \int_0^{\infty} g(x)\, dH_n(x), \tag{2.9}$$

where $g(x) = x \log x$. The sufficient conditions (1) through (18) of Sethuraman and

Rao (1970, p. 410) are easily verified for the function $g(x)=x\log x$. Thus the map $y(x)\to\int_0^\infty g(x)\,dy(x)$ for $y(x)\in D[0,\infty]$ is continuous with probability one under $\{\varrho(x), 0\le x\le\infty\}$, and Theorem 2.2 applies. Further, μ_n is given by

$$\begin{aligned}\mu_n &= \int_0^\infty g(x)\,dG_n(x)\\ &= g(0)+\int_0^\infty g'(x)[1-G_n(x)]\,dx\\ &= \int_0^\infty (1+\log x)\left\{e^{-x}-\left(\frac{1}{\sqrt{n}}\int_0^1 l^2(t)\,dt\right)e^{-x}(x-\tfrac{1}{2}x^2)\right\}dx\\ &= (1-\gamma)+\frac{1}{2\sqrt{n}}\int_0^1 l^2(t)\,dt.\end{aligned}$$

Also, σ^2 is given by

$$\begin{aligned}\sigma^2 &= \iint\limits_{0\le x\le y\le\infty}(1+\log x)(1+\log y)e^{-y}(1-e^{-x}-xye^{-x})\,dx\,dy\\ &\quad+\iint\limits_{0\le y\le x\le\infty}(1+\log x)(1+\log y)e^{-x}(1-e^{-y}-xye^{-y})\,dx\,dy\\ &= \int_0^\infty(1+\log x)\left\{\int_x^\infty(1+\log y)e^{-y}\,dy\right\}dx\\ &\quad-\int_0^\infty(1+\log x)\left\{\int_x^\infty(1+\log y)e^{-y}\,dy\right\}dx\\ &\quad+\int_0^\infty(1+\log x)e^{-x}\left\{\int_0^x(1+\log x)\,dy\right\}dx\\ &\quad-\int_0^\infty(1+\log x)e^{-x}\left\{\int_0^x(1+\log y)e^{-y}\,dy\right\}dx\\ &= -3-2\gamma^2+2\int_0^\infty e^{-x}\log^2 x\,dx\\ &= \tfrac{1}{3}\pi^2-3,\end{aligned}$$

since (see, for example, Ryshik and Gradstein (1957), p. 197)

$$\int_0^\infty e^{-x}\log^2 x\,dx=\gamma^2+\tfrac{1}{6}\pi^2.$$

As a corollary we have:

Corollary 2.4 (cf. Gebert and Kale (1969)). *Under the hypothesis* (1.4), *the random*

variable $\sqrt{n}\{E_n-(1-\gamma)\}$ *has a limiting normal distribution with mean zero and variance* $\frac{1}{3}\pi^2-3$.

A similar result about $V_2(n)$, corresponding to $h(x)=x^2$ in (1.2), was discussed by Sethuraman and Rao (1970) (see also Kuo and Rao (1984)) and we state:

Theorem 2.5 (Sethuraman and Rao (1970), Kuo and Rao (1984)). *Under the sequence of alternatives* (1.5), *the random variable*

$$\sqrt{n}\left(V_2(n)-\left\{2+\frac{2}{\sqrt{n}}\int_0^1 l^2(t)\,dt\right\}\right)$$

has a limiting normal distribution with mean zero and variance 4.

Finally, we consider the asymptotic distribution of the chi-square test S_n in (1.8) under the alternatives (1.5). For this, we make use of Theorem 2.1 of Holst and Rao (1980, p. 25) on the asymptotic distribution of statistics based on multinomial frequencies. We state this result for completeness, in the present notations. This theorem is not a special case of Theorem 2 of Holst (1979) as stated there, but rather a straightforward extension to the non-identically distributed case.

Let $(O_{1n},\dots,O_{nn})$ be Mult$(n;p_{1n},\dots,p_{nn})$. We are interested in the asymptotic distribution of

$$W_n=\sum_{k=1}^{n} h_k(O_{kn}) \quad \text{as } n\to\infty. \tag{2.10}$$

where $\{h_k;\, k=1,\dots,n\}$ is a sequence of real-valued Borel-measurable functions. Let $\{\xi_{1n},\dots,\xi_{nn}\}$, $n\geq 1$, be a triangular array of independent Poisson random variables where ξ_{kn} is Pois(np_{kn}). Define

$$\lambda_n=\sum_{k=1}^{n} h_k(\xi_{kn}) \tag{2.11}$$

and let

$$\nu_n=E(\lambda_n) \quad \text{and} \quad \sigma_n^{*2}=\mathrm{Var}(\lambda_n). \tag{2.12}$$

For $0<q<1$, let $N=[nq]$, the integer part of (nq), and

$$\lambda_{nq}=\sum_{k=1}^{N} h_k(\xi_{kn}). \tag{2.13}$$

Theorem 2.6 (Theorem 2.1 of Holst and Rao (1980)). *Let* $\lambda_n, \nu_n, \sigma_n^*$ *and* λ_{nq} *be as defined in* (2.11), (2.12) *and* (2.13). *Assume that there exists a* $q_0<1$ *such that for* $q\geq q_0$, $\sum_{k=1}^{N} p_{kn}\to P_q$, $0<P_q<1$, *and suppose*

$$\begin{pmatrix} n^{-1/2}(\lambda_{nq}-E\lambda_{nq}) \\ n^{-1/2}\sum_{k=1}^{N}(\xi_{kn}-np_{kn}) \end{pmatrix} \xrightarrow{\ \mathrm{d}\ } \mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} A_q & B_q \\ B_q & P_q \end{pmatrix}\right)$$

where $A_q \to A_1$, $B_q \to B_1$ *and* $P_q \to 1$ *as* $q \to 1-0$. *Then as* $n \to \infty$,

$$n^{-1/2}(W_n - v_n) \longrightarrow N(0, A_1 - B_1^2).$$

Since $n \to \infty$, we shall write n instead of $n-1$ and the χ^2-statistic (1.8) is

$$S_n = \sum_{i=1}^{n} (O_{in} - 1)^2 = \sum_{i=1}^{n} O_{in}^2 - n. \tag{2.14}$$

In our case,

$$p_{kn} = \text{Probability of the } k\text{-th cell, namely, } \left[\frac{k-1}{n}, \frac{k}{n}\right)$$

$$= \int_{k-1/n}^{k/n} f_n(x)\,dx \simeq \frac{1}{n}\left[1 + \frac{l(k/n)}{n^{1/4}}\right] \tag{2.15}$$

under the alternatives (1.5).

We define

$$\lambda_n = \sum_{k=1}^{n} \xi_{kn}^2 - n \tag{2.16}$$

where $(\xi_{1n}, \ldots, \xi_{kn})$ are independent Pois(np_{kn}) as stated earlier. Now, it can be verified using the appropriate moments of the Poisson random variables that

$$v_n = E(\lambda_n) = \sum_{k=1}^{n} \left[1 + \frac{l(k/n)}{n^{1/4}}\right]^2$$

$$\simeq n\left(1 + \frac{1}{\sqrt{n}}\left(\int_0^1 l^2(t)\,dt\right)\right), \tag{2.17}$$

and

$$A_q = \lim_{n\to\infty} \frac{1}{N} \operatorname{Var}\left(\sum_{k=1}^{N} \xi_{kn}^2\right) = 11\,q \tag{2.18}$$

and

$$B_q = \lim_{n\to\infty} \frac{1}{N} \operatorname{Cov}\left(\sum_{k=1}^{N} \xi_{kn}^2, \sum_{k=1}^{N} \xi_{kn}\right) = 3\,q. \tag{2.19}$$

The joint asymptotic normality required in Theorem 2.6 is established if we verify for any real a, that the triangular sequence

$$\{Y_{kn} = a\xi_{kn}^2 + \xi_{kn};\ k = 1, \ldots, n\} \tag{2.20}$$

satisfies, for instance, Liapounov's condition (see Chung (1968), p. 200).

We need to show that

$$\sum_{k=1}^{N} E|Y_{kn}|^3 \Big/ \left[\operatorname{Var}\left(\sum_{k=1}^{N} Y_{kn}\right)\right]^{3/2} \tag{2.21}$$

goes to zero as $N \to \infty$. Since $n^{-1}\mathrm{Var}(\sum_{k=1}^{N} Y_{kn})$ has finite, non-zero limit, it follows that $\mathrm{Var}(\sum_{k=1}^{N} Y_{kn})$ is $O(\sqrt{N})$. It is easily checked that the numerator in (2.21) is of order N so that the ratio in (2.21) goes to zero as $N \to \infty$.

Thus from Theorem 2.6 we have:

Theorem 2.7. *Under the alternatives* (1.5), *the random variable* $n^{-1/2}(S_n - \nu_n)$ *has a normal distribution with mean zero and variance* $(A_1 - B_1^2) = 11 - 3^2 = 2$.

Under the hypothesis (1.4) the asymptotic distribution of S_n in (2.15) is immediately given by the following:

Corollary 2.8. *Under the hypothesis* (1.4), *the random variable* $n^{-1/2}(S_n - n)$ *has a limiting normal distribution with mean zero and variance* 2.

3. Pitman asymptotic relative efficiency of E_n, $V_2(n)$ and S_n

The Pitman asymptotic relative efficiency (ARE) of a test relative to another test is defined to be the limit of the inverse ratio of sample sizes required to obtain the same limiting power at a sequence of alternatives converging to the hypothesis. The limiting power should be a value between the limiting test size, α, and the maximum power, 1. If the limiting power of a test at a sequence of alternatives is α, then its ARE with respect to any other test with the same test size and with limiting power greater than α, is zero. On the other hand, if the limiting power of a test at a sequence of alternatives converges to a number in the open interval $(\alpha, 1)$, then a measure of rate of convergence, called efficacy, can be computed. Under certain standard regularity assumptions (see, for example, Fraser (1957)), which include a condition about the nature of the alternative, asymptotic normal distribution of the test statistic under the sequence of alternatives, etc., this efficacy is given by

$$\text{efficacy} = \frac{\mu^4}{\sigma^4}. \tag{3.1}$$

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

The efficacy of the test statistic E_n from Theorem 2.3 is

$$\mathrm{eff}(E_n) = \left(\int_0^1 l^2(t)\,\mathrm{d}t\right)^4 \Big/ 16(\tfrac{1}{3}\pi^2 - 3)^2. \tag{3.2}$$

The efficacy of the test statistic $V_2(n)$ can be computed from Theorem 2.5 and we obtain

$$\text{eff}(V_2(n)) = \left(\int_0^1 l^2(t)\,dt\right)^4 \Big/ 16. \tag{3.3}$$

Finally, the efficacy of the test statistic S_n from Theorem 2.7 is

$$\text{eff}(S_n) = \left(\int_0^1 l^2(t)\,dt\right)^4 \Big/ 4. \tag{3.4}$$

Since the efficacies in (3.2), (3.3) and (3.4) depend on the alternatives only through the multiplying constant, $(\int_0^1 l^2(t)\,dt)^4$, we define the 'modified efficacies' of E_n, $V_2(n)$ and S_n by the ratio of their efficacies to $(\int_0^1 l^2(t)\,dt)^4$. Thus from (3.2), (3.3) and (3.4), it follows that the test statistic E_n is 75% as efficient as $V_2(n)$, and the test statistic S_n is 25% as efficient as $V_2(n)$.

Remark 3.1. We can compare the ARE's of the test statistics E_n and S_n with the test statistic

$$V_r(n) = \frac{1}{n}\sum_{i=1}^{n} (nD_i)^r, \quad r \geq 0. \tag{3.5}$$

The efficacy of $V_r(n)$ can be computed from the expression (see Sethuraman and Rao (1970), p. 411, eqn. 21)

$$\text{eff}(T_n) = \frac{(\int_0^1 l^2(t)\,dt)^4 (\int_0^\infty h'(x)e^{-x}(x-\frac{1}{2}x^2)\,dx)^4}{\{\int_0^\infty \int_0^\infty h'(x)h'(y)K(x,y)\,dx\,dy\}^2} \tag{3.6}$$

by substituting $h(x) = x^r$, $r \geq 0$, where T_n is given by (1.2). After simplification, we have

$$\text{eff}(V_r(n)) = \frac{\{r(1-r)\}^4 (\int_0^1 l^2(t)\,dt)^4}{16\{\Gamma(2r+1)/\Gamma^2(r+1) - (1+r^2)\}^2}, \quad r \geq 0. \tag{3.7}$$

The Pitman ARE of E_n with respect to $V_r(n)$ is

$$\text{ARE}(E_n, V_r(n)) = \frac{\{\Gamma(2r+1)/\Gamma^2(r+1) - (1+r^2)\}^2}{\{r(1-r)\}^4 (\frac{1}{3}\pi^2 - 3)^2}, \quad r \geq 0. \tag{3.8}$$

Note that the ARE in (3.8) is independent of the alternatives. Similarly, the ARE of S_n with respect to $V_r(n)$ is

$$\text{ARE}(S_n, V_r(n)) = \frac{\{\Gamma(2r+1)/\Gamma^2(r+1) - (1+r^2)\}^2}{4\{r(1-r)\}^4}. \tag{3.9}$$

On the basis of some preliminary computations not only $V_2(n)$ but $V_r(n)$ for r in the range [1.1, 3.19] seems to be preferable to E_n.

References

Chung, K.L. (1968). *A Course in Probability Theory.* Academic Press, New York.

Cibisov, D.M. (1961). On the tests of fit based on sample spacings. *Theor. Probab. Appl.* 6, 325–329.

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

Gebert, J.B. and B.K. Kale (1969). Goodness of fit tests based on discriminatory information. *Statistische Hefte* 10, 192–200.

Holst, L. (1979). Two conditional limit theorems with applications. *Ann. Statist.* 7, 551–557.

Holst, L. and J.S. Rao (1980). Asymptotic theory for some families of two-sample nonparametric statistics, *Sankhya Ser. A* 42, 19–52.

Kale, B.K. (1969). Unified derivation of tests of goodness of fit based on spacings. *Sankhya Ser. A* 31, 43–48.

Kuo, M. and J.S. Rao (1984). Asymptotic results on the Greenwood statistic and some of its generalizations. *J. Roy Statist. Soc. Ser. B* 46, 228–237.

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

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

Ryshik, I.M. and I.S. Gradstein (1957). *Tables of Series, Products and Integrals.* Veb Deutsches Verlag, Berlin.

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

Weiss, L. (1965). On asymptotic sampling theory for distributions approaching the uniform distribution. *Z. Wahrsch. Verw. Geb.* 4, 217–221.