New Perspectives in Theoretical and Applied Statistics

Edited by

Madan Lal Puri,
José Pérez Vilaplana &
Wolfgang Wertz

A Volume in the Wiley Series in Probability and Mathematical Statistics:
Vic Barnett, Ralph A. Bradley, J. Stuart Hunter, David G. Kendall,
Adrian F. M. Smith, Stephen M. Stigler, Geoffrey S. Watson—Advisory Editors

Efficiencies of Some Disjoint Spacings Tests Relative to a χ^2 Test

S. Rao Jammalamadaka
Department of Mathematics
University of California at Santa Barbara
Santa Barbara, California

Ram C. Tiwari
Department of Mathematics
Indian Institute of Technology
Bombay, India

SUMMARY

As is well known, the goodness-of-fit problem can be reduced, through a probability integral transformation, to testing randomness or uniformity on the interval [0, 1]. Among tests based symmetrically on the m-step spacings ($m \geq 1$), a generalized Greenwood statistic based on the sum of squares of the spacings is known to be locally most powerful. On the other hand, one may consider a χ^2 test statistic with an expected frequency of m in each cell, which is again locally most powerful among tests based symmetrically on the observed frequencies. This comparison is justified since, while the Greenwood test compares the observed and expected cell lengths, holding the observed number in each cell to m, the χ^2 test compares the observed and expected frequencies, holding the expected number in each cell to m. By considering a suitable sequence of alternatives, we compare the asymptotic relative efficiencies of these two tests as well as a third entropy-type

test based on the m-step spacings. These results generalize some earlier work of the authors (Jammalamadaka and Tiwari, 1985).

1. INTRODUCTION

Let $X_1, X_2, \ldots, X_{n-1}$ be independent random variables (r.v.) with a common continuous distribution function F. The goodness-of-fit problem of testing whether a specified distribution generated the observations can be reduced to testing if the observations have a uniform distribution, through the probability integral transformation. Thus we may (and shall) assume, without any loss of generality, that the support of F is $[0,1]$ and that the null hypothesis of interest is

$$H_0: f(x) = 1, \qquad x \in [0,1] \tag{1.1}$$

where f denotes the probability density function. One classical approach is to use the χ^2 procedure. Suppose we have N classes

$$\left(\frac{i-1}{N}, \frac{i}{N}\right], \qquad i = 1, \ldots, N$$

and O_i denotes the observed frequency in the ith class. The χ^2 statistic is given by

$$\begin{aligned} T_{n,N} &= \frac{N}{n} \sum_{i=1}^{N} \left(O_i - \frac{n}{N}\right)^2 \\ &= \frac{N}{n}\left(\sum_{i=1}^{N} O_i^2\right) - n \end{aligned} \tag{1.2}$$

For reasons that will become clear soon, we shall be interested in the case where $n, N \to \infty$ such that the expected frequency in each class $n/N \to m$, $0 < m < \infty$.

An alternative approach to testing (1.1) is to construct tests based on spacings. Let $\{X'_k\}$ denote the order statistics, with the notation $X'_0 = 0$ and $X'_k = 1 + X'_{k-n}$ for $k \geq n$, circularly for convenience. Then the non-overlapping (or disjoint) m-step spacings ($1 \leq m < n$) are defined by

$$D^{(m)}_{k \cdot m} = X'_{k \cdot m} - X'_{(k-1) \cdot m}, \qquad k = 1, \ldots, [n/m] \tag{1.3}$$

where $[x]$ denotes the integer part of x. Since we are concerned with

asymptotic results as $n \to \infty$ with m fixed, there is no loss of generality in assuming $[n/m]$ is an integer N. When $m = 1$, $\{D_k^{(1)}\}$ are simply $\{D_k\}$ are called one-step or simple spacings. The order statistics as well as the spacings should have an extra subscript n, which we suppress throughout this chapter for notational simplicity. Tests based on simple spacings, for the goodness-of-fit problem, have been considered in the literature. See, for instance, Pyke (1965), Rao and Sethuraman (1975), and references contained there. Del Pino (1979), following the approach taken by Rao and Sethuraman (1975), studies tests based on disjoint m spacings defined in (1.3). We consider two such tests, namely,

$$V_n^{(m)} = \frac{1}{N} \sum_{i=1}^{N} (nD_{i\cdot m})^2 \tag{1.4}$$

and an entropy-type statistic

$$E_n^{(m)} = \frac{1}{N} \sum_{i=1}^{N} \left(nD_{i\cdot m}^{(m)}\right) \log\left(nD_{i\cdot m}^{(m)}\right) \tag{1.5}$$

The statistic in (1.4) is a generalization of the so-called Greenwood statistic [see, e.g., Rao and Kuo (1984)], and the one in (1.5) for the special case $m = 1$ has been considered earlier by Gebert and Kale (1969) and more recently by Jammalamadaka and Tiwari (1985). The statistic $V_n^{(m)}$ may be seen as being equivalent to

$$\frac{1}{m} \sum_{i=1}^{N} \left[m - nD_{i\cdot m}^{(m)}\right]^2$$

which in this form may be thought of as the dual of the χ^2 statistic (1.2) with the observed frequencies in each cell (of length $D_{i\cdot m}^{(m)}$) being fixed at m.

To compare the asymptotic relative efficiencies of these three tests, we need to consider their distribution theory under the following sequence of alternatives:

$$A_n\colon f_n(x) = 1 + \frac{l(x)}{n^{1/4}}, \qquad 0 \le x \le 1 \tag{1.6}$$

where $l(\cdot)$ is square-integrable and continuously differentiable on $[0, 1]$. This sequence of alternatives has been considered before. See, for instance, Rao and Sethuraman (1975) and Del Pino (1979).

Under the sequence of alternatives (1.6), the asymptotic normality of the three statistics is established in Section 2 and comparison of asymptotic efficiencies made in Section 3.

2. ASYMPTOTIC NORMALITY OF $V_n^{(m)}$, $E_n^{(m)}$, AND $T_{n,N}$

To establish the asymptotic normality of $V_n^{(m)}$ and $E_n^{(m)}$, we use the following result of Del Pino (1979). See also Rao and Kuo (1984).

Theorem 2.1 (Del Pino, 1979). Under the alternatives (1.6), if $h(\cdot)$ is a function satisfying some regularity conditions (Del Pino, 1979) and S is a Gamma$(m, 1)$ random variable with density $s^{m-1}e^{-s}/\Gamma m$ for $s \geq 0$, then

$$N^{-1/2} \sum_{i=1}^{N} \left[h\left(nD_{i\cdot m}^{(m)}\right) - Eh(S) \right] \xrightarrow{d} N(\mu, \sigma^2)$$

where

$$\mu = \left(\int_0^1 l^2(t)\, dt \right) \mathrm{Cov}\left(h(S), (S - m - 1)^2\right)/2\sqrt{m} \tag{2.1}$$

and

$$\sigma^2 = \mathrm{Var}(h(S)) - \mathrm{Cov}^2(h(S), S)/m \tag{2.2}$$

The statistic $V_n^{(m)}$ as a special case of this with $h(x) = x^2$ has already been studied by Del Pino (1979) and Rao and Kuo (1984), and we state the needed result.

Theorem 2.2. Under the sequence of alternatives (1.6), the random variable

$$\sqrt{N}\left(V_n^{(m)} - m(m+1)\right) \xrightarrow{d} N\left(\mu_1, \sigma_1^2\right)$$

where

$$\mu_1 = \sqrt{m} \cdot (m+1)\left(\int_0^1 l^2(t)\, dt \right) \quad \text{and} \quad \sigma_1^2 = 2m(m+1)$$

Now we consider $E_n^{(m)}$ in (1.8) and establish the following

Theorem 2.3. Under the sequence of alternatives (1.6), the random variable

$$\sqrt{N}\left[E_n^{(m)} - m\left(1 + \frac{1}{2} + \cdots + \frac{1}{m} - \gamma\right)\right] \xrightarrow{d} N\left(\mu_2, \sigma_2^2\right)$$

where

$$\mu_2 = \frac{\sqrt{m}}{2}\left(\int_0^1 l^2(t)\, dt \right)$$

and

$$\sigma_2^2 = m(m+1)\left\{\frac{\pi^2}{6} - \sum_{j=1}^{m}\frac{1}{j^2}\right\} - m$$

(Here γ is the Eüler constant, 0.5772... .)

Proof. It can be checked that the function $h(x) = x \log x$ satisfies the simple set of sufficient conditions [see Eq. (3.3), p. 1061, of Del Pino (1979)]. Then to apply Theorem 2.1, we need to evaluate the centering constant $E(S \log S)$ and the asymptotic mean and variance in (2.1) and (2.2) for this case. Evaluation of the required integrals turns out to be quite complex but manageable. For instance

$$\begin{aligned} E(S \log S) &= \frac{1}{\Gamma m}\int_0^\infty e^{-x} \cdot x^m \cdot \log x \cdot dx \\ &= \frac{1}{\Gamma m}\left\{m\int_0^\infty e^{-x} \cdot x^{m-1} \cdot \log x \cdot dx + \Gamma m\right\} \\ &= m\left\{1 + \frac{1}{2} + \cdots + \frac{1}{m} - \gamma\right\} \end{aligned}$$

For getting μ, we need

$$\begin{aligned} &\operatorname{Cov}\left(S \log S, (S-m-1)^2\right) \\ &\quad = E\left\{S^3 \log S + (m+1)^2 S \log S - 2(m+1)S^2 \log S\right\} \\ &\qquad - E(S \log S)E\left(S^2 + (m+1)^2 - 2(m+1)S\right) \\ &\quad = \left\{(m+2)(m+1)m\left(1 + \frac{1}{2} + \cdots + \frac{1}{m+2} - \gamma\right)\right. \\ &\qquad + (m+1)^2 m\left(1 + \frac{1}{2} + \cdots + \frac{1}{m} - \gamma\right) \\ &\qquad \left. - 2(m+1)^2 m\left(1 + \frac{1}{2} + \cdots + \frac{1}{m+1} - \gamma\right)\right\} \\ &\qquad - m\left\{1 + \frac{1}{2} + \cdots + \frac{1}{m} - \gamma\right\}\left\{(m+1)m + (m+1)^2 - 2(m+1)m\right\} \\ &\quad = m \end{aligned}$$

and hence μ_2 follows from (2.1). Similarly, one can show that the expression (2.2) for σ^2 reduces to the one given in Theorem 2.3. ■

From these two theorems, 2.2 and 2.3, one obtains the asymptotic null distributions under (1.1) by setting $l(x) \equiv 0$ as well as the special case for simple spacings by taking $m = 1$.

Finally, we consider the asymptotic distribution of the χ^2 statistic $T_{n,N}$ in (1.2) under the alternatives (1.6). For this, we use Theorem 2.1 of Holst and Rao (1980, p. 25) on the asymptotic distribution of statistics based on multinomial frequencies. The proof of the following result is essentially similar to that of Theorem 2.7 of Jammalamadaka and Tiwari (1985) and is omitted.

Theorem 2.4. Under the alternatives (1.6), the random variable

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

has an asymptotic normal distribution with mean 0 and variance 2 as $n, N \to \infty$ such that $n/N \to m$, finite.

3. ASYMPTOTIC RELATIVE EFFICIENCIES OF $V_n^{(m)}$, $E_n^{(m)}$, AND $T_{n,N}$

The Pitman asymptotic relative efficiency (ARE) of a test relative to another 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 null. Under certain regularity conditions [see, for example, Fraser (1957)], which include a condition about the type of alternatives, asymptotic normality of the test statistic under a sequence of alternatives, etc., the "efficacy" of a test statistic is given by (μ_Δ^4/σ^4), where μ_Δ and σ 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.

From Theorem 2.2, the efficacy of the test statistic $V_n^{(m)}$ is

$$\mathrm{Eff}\left(V_n^{(m)}\right) = \frac{m^2(m+1)^4\left(\int_0^1 l^2(t)\,dt\right)^4}{4m^2(m+1)^2}$$

$$= \frac{(m+1)^2}{4} \cdot \left(\int_0^1 l^2(t)\,dt\right)^4 \tag{3.1}$$

Similarly from Theorem 2.3, the efficacy of $E_n^{(m)}$ is

$$\text{Eff}\left(E_n^{(m)}\right) = \frac{m\left(\int_0^1 l^2(t)\,dt\right)^4}{16\left\{m(m+1)\left[\frac{\pi^2}{6} - \Sigma_{j=1}^{m}\frac{1}{j^2}\right] - m\right\}^2}$$

$$= \frac{\left(\int_0^1 l^2(t)\,dt\right)^4}{16\left\{(m+1)\left[\frac{\pi^2}{6} - \Sigma_{j=1}^{m}\frac{1}{j^2}\right] - 1\right\}^2}. \tag{3.2}$$

And finally from Theorem 2.4, the efficacy of the χ^2 statistic $T_{n,N}$ is

$$\text{Eff}(T_{n,N}) = \frac{m^2}{4}\left(\int_0^1 l^2(t)\,dt\right)^4 \tag{3.3}$$

Since all these efficacies depend on the alternatives only through the multiplying constant $(\int_0^1 l^2(t)\,dt)^4$, it makes the comparison easy. It is clear that $V_n^{(m)}$ based on the sum of squares of disjoint m spacings is superior to the χ^2 statistic or the entropy-type statistic $E_n^{(m)}$. Also, the test based on $E_n^{(m)}$ is asymptotically more efficient than the χ^2 statistic with expected frequency of m in each cell. Thus spacings tests seem preferable to comparable χ^2 procedures. One can make a table of the relative efficiencies of these three tests using (3.1), (3.2), and (3.3) for various values of m. These conclusions agree with the relative ordering obtained earlier by the authors (Jammalamadaka and Tiwari, 1985) for the case $m = 1$.

REFERENCES

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.

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

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

Jammalamadaka, S. Rao and Tiwari, R. C. (1985). Asymptotic comparison of three tests for goodness of fit. *J. Statist. Plan. Inf.* **12**, 295–304.

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

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

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.