The Annals of Statistics
1979, Vol. 7, No. 5, 1058–1065

ON THE ASYMPTOTIC DISTRIBUTION OF k-SPACINGS WITH APPLICATIONS TO GOODNESS-OF-FIT TESTS[1]

BY GUIDO E. DEL PINO

Universidad de Chile

Let $X_1, \cdots, X_n$ be an ordered sample from a distribution A_n on [0, 1]. The k-spacings $D_1(N, k), \cdots, D_N(N, k)$ are defined and the weak convergence of their empirical distribution function under a sequence of alternatives A_n approaching the uniform distribution is established. This is then applied to find the limiting distribution of $W_n(g, k) = N^{-\frac{1}{2}}\Sigma_{i=1}^N(g(NkD_i(N, k)) - a)$ where g is a smooth function and k is fixed. The statistics $W_n(g, k)$ can be used to test the hypothesis that the observations are uniformly distributed in [0, 1]. The asymptotic relative efficiency of $W_n(g, k)$ with respect to $W_n(g, 1)$ is shown to increase without limit for several functions g. The test with $g(x) = x^2$ is shown to be asymptotically optimal within the class $W_n(g, k)$ for any fixed k. The paper extends results of Rao and Sethuraman.

1. Introduction. Let $X_1, X_2, \cdots, X_n$ be an ordered sample from a distribution A_n on [0, 1]. For any fixed k, the k-spacings are defined by

$$\begin{aligned} D_1(N, k) &= X_k \\ D_i(N, k) &= X_{ik} - X_{(i-1)k} \qquad\qquad i = 2, \cdots, N-1 \\ D_N(N, k) &= 1 - X_{(N-1)k} \end{aligned} \tag{1.1}$$

where N is the smallest integer greater than or equal to $(n + 1)/k$. For notational simplicity, the arguments N and k will not be indicated explicitly. Also, since we will only be concerned with asymptotic properties it will be assumed, without loss of generality, that $n + 1 = Nk$. When $k = 1$, the k-spacings reduce to the usual spacings considered in the literature. They will be called simple spacings.

Rao and Sethuraman (1975) study the asymptotic behavior of the empirical distribution function F_N of the normalized simple spacings $(n + 1)D_i$, under a sequence of smooth alternatives A_n converging to the uniform distribution at the rate $n^{-\delta}$. Statistics symmetric in the simple spacings can be viewed as functionals of F_N and be used to test the null hypothesis that the observations are uniformly distributed in [0, 1]. Rao and Sethuraman show that tests based symmetrically on the simple spacings are asymptotically unable to detect alternatives approaching the uniform at a faster rate than $n^{-\frac{1}{4}}$.

The main motivation behind this paper is to explain the reasons for the poor performance of these tests and to suggest modifications to improve them. The basic idea is that the problem arises because symmetric functions of the simple spacings

Received February 1975; revised March 1977.

[1] This research was done while the author was at the University of Wisconsin, Madison, and was supported by the Air Force Office of Scientific Research under Grant No. AFOSR 72-2363B.

AMS 1970 *subject classifications.* Primary 62G30; secondary 62E20.

Key words and phrases. k-spacings, asymptotic distributions, goodness of fit.

do not take into account the smoothness of the probability density function. The situation is somewhat analogous to that of using a histogram with as many cells as observations. On the other hand, symmetric functions are useful in that they typically provide consistent tests against wide classes of alternatives. A compromise is achieved by using symmetric functions on the k-spacings (they will not be symmetric in the simple spacings if $k > 1$, unless they are constant). Although on intuitive grounds one should let k increase with the sample size n, in the present paper we restrict ourselves to the case k fixed. In the last section we discuss the general case briefly.

In Section 2 we study the weak convergence in the Skorokhod topology of the empirical process based on the k-spacings. This is done by an application of a theorem in Rao and Sethuraman (1975). In Section 3 we obtain the asymptotic normal distribution of

$$W_n(g, k) = N^{-\frac{1}{2}}\Sigma_{i=1}^N(g(NkD_i) - a)$$

under a sequence of alternatives approaching the uniform distribution at the rate $n^{-\frac{1}{4}}$. Statistics of this type have often been proposed in the literature for $k = 1$ (see e.g., Greenwood (1946), Moran (1947), Sherman (1950), Kimball (1950), Darling (1953), Pyke (1965). The asymptotic results in Sections 2 and 3 for the null case are of independent interest since they can be applied to several problems arising in connection with the Dirichlet distribution. In Section 4 we find the asymptotic relative efficiency (ARE) between different tests of the type W_n. It is shown for several functions g that the ARE of $W_n(g, k)$ with respect to $W_n(g, 1)$ increases without limit as k tends to infinity. In Section 5 we prove that $g(x) = x^2$ gives an asymptotically optimal test among tests of the form $W_n(g, k)$. It is argued that this optimality property is likely to hold among *all* tests depending symmetrically on the k-spacings.

2. Weak convergence of the empirical process of the k-spacings. Let $Y_1, \cdots, Y_N$ be independent and identically distributed random variables with probability density function

$$h_k(y) = \frac{1}{\Gamma(k)}y^{k-1}e^{-y}$$

and corresponding distribution function H_k. It is well known that under the uniform distribution

$$(NkD_i, i = 1, \cdots, N) =_d (Y_i/T_n, i = 1, \cdots, N)$$

where

$$T_n = \frac{1}{Nk}\Sigma_{i=1}^N Y_i$$

and $=_d$ stands for "has the same distribution as". Consider the sequence of distributions

$$A_n(t) = t + L(t)m(n) \qquad 0 \leqslant t \leqslant 1 \tag{2.1}$$

where L is twice differentiable, $L(0) = L(1) = 0$, and $m(n) = 0(n^{-\frac{1}{4}})$. The derivatives of A_n and L are denoted by a_n and l respectively. Let

$$J_n(x) = \frac{1}{N}\Sigma_{i=1}^N H_k(e_{ni}x) \tag{2.2}$$

where

$$e_{ni} = 1 + l\left(\frac{i}{N}\right)m(n) - L\left(\frac{i}{N}\right)l'\left(\frac{i}{N}\right)m^2(n) + o(m^2(n)).$$

Let F_n be the empirical distribution function of the normalized spacings NkD_i. Define the empirical process $\tilde{\eta}_n$ as

$$\tilde{\eta}_n(x) = N^{\frac{1}{2}}(F_n(x) - J_n(x)) \quad \text{and} \quad \xi_n = N^{\frac{1}{2}}(T_n - 1) \quad 0 \leqslant x \leqslant \infty. \tag{2.3}$$

It can be easily checked that all the conditions for the application of Corollary 2.7 of Rao and Sethuraman (1975), page 309, are satisfied except for a trivial modification to allow for the asymptotic variance of ξ_n not being equal to one. Thus we get the following.

THEOREM 1. *The processes* $(\tilde{\eta}_n(x), 0 \leqslant x \leqslant \infty)$ *converge weakly in* $D[0, \infty]$ *to a Gaussian process* $(\tilde{\eta}(x); 0 \leqslant x \leqslant \infty)$ *with mean zero and covariance function*

$$K(x, y) = \min(H_k(x), H_k(y)) - H_k(x)H_k(y) - \frac{1}{k}xyh_k(x)h_k(y). \tag{2.4}$$

Weak convergence in $D[0, \infty]$ is discussed in Rao and Sethuraman (1975). Let

$$\eta_n(x) = N^{\frac{1}{2}}(F_n(x) - H_k(x)) \tag{2.5}$$

and let

$$V_n(x) = N^{\frac{1}{2}}(J_n(x) - H_k(x)). \tag{2.6}$$

Then

$$\eta_n(x) = \tilde{\eta}_n(x) + V_n(x).$$

The asymptotic distribution of $\tilde{\eta}_n$ is the same under the null distribution and under the sequence A_n, its asymptotic distribution being given in Theorem 1. A Taylor expansion of $J_n\,(x)$ gives

$$V_n(x) = \left(N^{\frac{1}{2}}m^2(n)\right)\left(xh_k(x) + \frac{x^2}{2}h_k'(x)\right)\textstyle\int_0^1 l^2(t)\,dt \tag{2.7}$$

$$+ o\left(N^{\frac{1}{2}}m^2(n)\right) \qquad \text{uniformly in } x.$$

It is seen from (2.7) that if $m(n)$ tends to zero faster than $n^{-\frac{1}{4}}$ then V_n converges to zero uniformly. Hence the proper choice of $m(n)$ is

$$m(n) = O(n^{-\frac{1}{4}}).$$

Since any constant factor can be assimilated into $L(x)$ we can take, without loss of generality, $m(n) = n^{-\frac{1}{4}}$ and

$$a_n(t) = 1 + l(t)n^{-\frac{1}{4}} \qquad 0 \leqslant t \leqslant 1. \tag{2.8}$$

We then have

THEOREM 2. *Under the sequence of alternatives* (2.8), *the processes* $(\eta_n(x), 0 \leqslant x \leqslant \infty)$ *converge weakly in* $D[0, \infty]$ *to a Gaussian process* $(\eta(x), 0 \leqslant x \leqslant \infty)$ *with mean function*

$$v(x) = \tfrac{1}{2}((k+1)x - x^2)h_k(x)\textstyle\int_0^1 l^2(t)\, dt \tag{2.9}$$

and covariance function (2.4).

It can be checked (cf. del Pino (1976), pages 163–166) that the probability measures induced by the process η and the process η_0 corresponding to $l = 0$, are mutually absolutely continuous and thus they cannot be discriminated with probability one.

3. Asymptotic distribution of W_n.

Consider now the statistic

$$W_n = N^{-\frac{1}{2}}\textstyle\sum_{i=1}^{N}(g(NkD_i) - a) \tag{3.1}$$

where

$$a = Eg(Y)$$

with Y having distribution H_k. The statistic W_n can be rewritten as

$$W_n = \textstyle\int_0^\infty g(x)\, d\eta_n(x). \tag{3.2}$$

By imposing smoothness conditions on the function g so that (3.2) can be integrated by parts and by applying the law of the iterated logarithm it is possible to obtain, in the same way as Rao and Sethuraman (1970), rather complicated sufficient conditions for the continuity of W_n as a functional of η_n in the Skorokhod topology. For practical purposes it is better to have an easy-to-check set of sufficient conditions implying these. A convenient set is

(3.3) (i) g is absolutely continuous in $(0, \infty)$ and g' is bounded on any closed interval in $(0, \infty)$.

(ii) Either g is monotone in the neighborhood of 0 and ∞, or g' is bounded on $[0, \infty]$.

(iii) $\lim_{x\to\infty} e^{-\alpha x}g^2(x) = 0$ for some $\alpha < 1$.

(iv) $\lim_{x\to 0} x^{\beta}g^2(x) = 0$ for some $\beta < k$.

We assume in what follows that (3.3) holds. Under these conditions, W_n converges in distribution to

$$W = -\textstyle\int_0^\infty g'(x)\eta(x)\, dx.$$

Let μ and σ^2 be the mean and variance of W. Then

$$\mu = \frac{(k+1)k^{\frac{1}{2}}}{2} \int_0^1 l^2(t)\, dt \int_0^\infty g'(x)(h_{k+2}(x) - h_{k+1}(x))\, dx \tag{3.4}$$

$$= (4k)^{-\frac{1}{2}} \int_0^1 l^2(t)\, dt \int_0^\infty g(x)(x^2 - 2(k+1)x + k(k+1))h_k(x)\, dx \tag{3.5}$$

$$= (4k)^{-\frac{1}{2}} \int_0^1 l^2(t)\, dt \int_0^\infty x^2 g''(x) h_k(x)\, dx, \tag{3.6}$$

this last expression being true only when g'' exists. Also

$$\sigma^2 = \int_0^\infty \int_0^\infty g'(x)g'(y)K(x, y)\, dx\, dy \tag{3.7}$$

$$\begin{aligned} &= \int_0^\infty g^2(x)h_k(x)\, dx - \left(\int_0^\infty g(x)h_k(x)\, dx\right)^2 \\ &\quad - \frac{1}{k}\left(\int_0^\infty g(x)(x-k)h_k(x)\, dx\right)^2. \end{aligned} \tag{3.8}$$

Note that expressions (3.5) and (3.8) do not involve the derivative of g. The differentiability conditions seem to be unnatural to the problem but are inherent in the method of proof. Under the null hypothesis, the results can be obtained using only the assumption that $g(Y)$ has a finite variance when Y has distribution H_k. This can be done by extending a theorem of LeCam (1958) along the lines of Pyke (1965) or del Pino (1976).

4. Asymptotic relative efficiency. We turn now to the question of computing the asymptotic relative efficiency (ARE(1, 2)) of two tests $W_n(g_1, k_1)$ and $W_n(g_2, k_2)$ corresponding to different g and k. Let $\mu(g_i, k_i)$, $\sigma^2(g_i, k_i)$ denote the asymptotic mean and variance of $W_n(g_i, k_i)$ under the sequence of alternatives (2.8). Then (see Fraser (1957), page 273)

$$\text{ARE}(1, 2) = \left(\frac{\mu(g_1, k_1)}{\sigma(g_2, k_2)}\right)^4 \Big/ \left(\frac{\mu(g_2, k_2)}{\sigma(g_2, k_2)}\right)^4.$$

Rao and Sethuraman use the wrong exponent 2 instead of 4 in the above expression. We will compute

$$e(g, k) = \mu^2(g, k) / \left(\sigma^2(g, k)\left(\int_0^1 l^2(t)\, dt\right)^2\right) \tag{4.1}$$

for several g and k. Then ARE(1, 2) is obtained by

$$\text{ARE}(1, 2) = e^2(g_1, k_1)/e^2(g_2, k_2).$$

EXAMPLES.

(1) $g(x) = x^\alpha$, $2\alpha > -k$ $\alpha(\alpha - 1) \neq 0$. From (3.6) and (3.8) we get

$$e(g, k) = \frac{1}{4k}\left(\frac{\alpha^2(\alpha-1)^2\Gamma^2(\alpha+k)}{\Gamma(2\alpha+k)\Gamma(k) - \Gamma^2(\alpha+k)(1+\alpha^2/k)}\right).$$

In particular for $\alpha = 2$

$$e(g, k) = \frac{k+1}{2}.$$

In general it can be proved that

$$\lim_{k\to\infty} \frac{2e(k)}{k} = 1 \quad \text{for } \alpha(\alpha - 1) \neq 0.$$

(2) $g(x) = \log x$. From (3.6) and (3.8)

$$e(g, k) = \frac{1}{4k\left(\Sigma_k^\infty \frac{1}{j^2} - \frac{1}{k}\right)}.$$

Again

$$\lim_{k\to\infty} \frac{2e(g, k)}{k} = 1$$

(3) $g(x) = |x - k|$. From (3.4) and (3.8)

$$e(g, k) = \frac{(k+1)^2 h_{k+2}^2(k)}{1 - 4kh_{k+1}^2(k) - \left[2(1 - h_{k+2}(k)) - \frac{2k}{k+1} h_{k+1}(k) - 1\right]^2}.$$

In particular

$$e(g, 1) = \frac{1}{8e - 20} \simeq 0.5726.$$

In this case

$$\lim_{k\to\infty} \frac{2e(g, k)}{k} = \frac{1}{\pi - 2} < 1.$$

5. Asymptotically most efficient test. Taking $g(x) = ag_1(x) + bg_2(x)$ and computing the asymptotic variance of $W_n(g, k)$, one obtains

$$\begin{aligned}\operatorname{Cov}(W(g_1, k), W(g_2, k)) &= \textstyle\int_0^\infty g_1(x) g_2(x) h_k(x)\, dx \\ &\quad - \left(\textstyle\int_0^\infty g_1(x) h_k(x)\, dx\right)\left(\textstyle\int_0^\infty g_2(x) h_k(x)\, dx\right) \\ &\quad - \frac{1}{k} \textstyle\int_0^\infty g_1(x)(x - k) h_k(x)\, dx \int_0^\infty g_2(x)(x - k) h_k(x)\, dx.\end{aligned}$$

Let $g_2(x) = x^2$. Then

$$\begin{aligned}\operatorname{Cov}(W(g_1, k), W(g_2, k)) &= \textstyle\int_0^\infty g_1(x)(x^2 - 2(k+1)x + k(k+1)) h_k(x)\, dx \\ &= (4k)^{\frac{1}{2}} \mu(g_1, k) / \left(\textstyle\int_0^1 l^2(t)\, dt\right).\end{aligned}$$

Thus $e(g, k)$ can be rewritten as

$$e(g, k) = \frac{\mathrm{Cov}^2(W(g, k), W(x^2, k))}{\mathrm{Var}(W(g, k))4k} \tag{5.1}$$

$$= \frac{k+1}{2}\rho^2(W(g, k), W(x^2, k)).$$

Expression (5.1) is maximized by choosing $g(x)$ as a multiple of x^2. The ARE of $W_n(g_1, k)$ with respect to $W_n(g_2, k)$ becomes

$$\mathrm{ARE}(1, 2) = \rho_1^4/\rho_2^4 \tag{5.2}$$

where ρ_i is the asymptotic correlation coefficient of $W_n(g_i, k)$ and $W_n(g, k)$ where $g(x) = x^2$. The optimality property of the statistic

$$T_{Nk} = N^{-\frac{1}{2}}\Sigma_{i=1}^{N}\left[(NkD_i)^2 - k(k+1)\right] \tag{5.3}$$

was proved in the case $k = 1$ by Rao and Sethuraman (1975). Also in this case Weiss (1957) considers the family of densities

$$f_\delta(x) = 1 + \delta\left(x - \tfrac{1}{2}\right) \qquad -\infty < \delta < \infty$$

and proves that the test rejecting the hypothesis of uniformity for large values of T_{Nk} is locally best unbiased within the class of all tests depending symmetrically on the simple spacings. The author has shown (del Pino (1976)) that the likelihood ratio statistic of the asymptotic Gaussian processes corresponding to the null and the sequence of alternative distributions, is a linear transformation of the stochastic integral

$$T = \int_0^\infty x^2\, d\eta(x).$$

This suggests that

$$T_{Nk} = \int_0^\infty x^2\, d\eta_n(x)$$

is asymptotically most efficient within the class of statistics based symmetrically on the k-spacings, for any fixed k.

6. Discussion. The examples in Section 4 indicate that k should be allowed to increase with the sample size n. Some purely formal manipulations suggest that choosing k proportional to $n^{4\delta-1}$ will produce tests with nontrivial asymptotic power under a sequence of alternatives approaching the uniform at the rate $n^{-\delta}$, $\frac{1}{4} \leqslant \delta \leqslant \frac{1}{2}$. A rigorous treatment of this case is given in the author's Ph.D. thesis and will be the subject of a separate paper. The rate $n^{-\frac{1}{2}}$ can only be achieved for a fixed number of spacings. For $g(x) = x^2$ and $g(x) = 1/x$ the resulting tests are related to chi-square tests with cells determined by sample quantiles. At first sight it seems as if one were throwing away too much information by using k-spacings with large k. That this is not the case asymptotically, follows from the results of Weiss (1974) on the asymptotic sufficiency of an increasing number of sample quantiles.

Simulation studies strongly confirm the increase in power obtained by taking k bigger than one. The behavior of the tests for $g(x) = x^2$ and $g(x) = \log x$ is very similar to that of chi-square tests, and they are not uniformly dominated by those based on the empirical distribution function. From a different point of view, the results in this paper may be useful in predicting the behavior of other tests, such as the chi-square and likelihood ratio tests for the multinomial distribution with equally likely cells. In particular, the test suggested by Kempthorne (1967) is likely to behave very poorly against smooth alternatives, at least for large sample sizes.

Acknowledgment. I would like to express my gratitude to Professor Grace Wahba for suggesting the problem and her encouragement in the course of this work. I would also like to thank Professors J. Sethuraman and J. S. Rao for making available some of their unpublished material.

REFERENCES

[1] Darling, D. A. (1953). On a class of problems relating to the random division of an interval. *Ann. Math. Statist.* **24** 239–253.

[2] del Pino, G. E. (1976). Spacings. Unpublished Ph.D. thesis, Univ. Wisconsin-Madison.

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

[4] Greenwood, M. (1946). The statistical study of infectious diseases. *J. Roy. Statist. Soc. Ser. A* **109** 85–110.

[5] Kempthorne, O. (1967). The classical problem of inference-goodness of fit. *Proc. Fifth Berkeley Symp. Math. Statist. Probability* **1** 235–249.

[6] Kimball, B. (1950). On the asymptotic distribution of the sum of powers of unit frequency differences. *Ann. Math. Statist.* **21** 263–271.

[7] LeCam, L. (1958). Un theoreme sur la division d'un intervalle par des points pris au hasard. *Publ. Inst. Statist. Univ. Paris* **7** 7–16.

[8] Moran, P. A. P. (1947). The random division of an interval. *J. Roy. Statist. Soc. Ser. B* **9** 92–98.

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

[10] Rao, J. S. and Sethuraman, J. (1970). Pitman effeciencies of tests based on spacings. In *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.), 267–273. Cambridge Univ. Press.

[11] 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.

[12] Sherman, B. (1950). A random variable related to the spacing of sample values. *Ann. Math. Statist.* **21** 339–361.

[13] Weiss, L. (1957). The asymptotic power of certain tests of fit based on sample spacings. *Ann. Math. Statist.* **28** 783–786.

[14] Weiss, L. (1974). The asymptotic sufficiency of a relatively small number of order statistics in goodness of fit. *Ann. Statist.* **2** 795–802.

Departamento de Matemáticas
Universidad de Chile
Casilla 2777
Santiago, Chile