Statistics & Probability Letters 8 (1989) 451–456
North-Holland

October 1989

LIMIT DISTRIBUTION OF SPACINGS STATISTICS WHEN THE SAMPLE SIZE IS RANDOM

Girish ARAS, S. Rao JAMMALAMADAKA and X. ZHOU

Statistics and Applied Probability Program, University of California, Santa Barbara, CA 93106, USA

Received April 1988
Revised November 1988

Abstract: A random sample size version of the central limit theorem is obtained for a general class of symmetric statistics based on uniform spacings. An important application to goodness of fit test for a Poisson process is discussed.

Keywords: Doeblin–Anscombe theorem, spacings, goodness of fit tests.

1. Introduction

Let $X_1, \ldots, X_{n-1}$ be the order statistics from a uniform distribution on [0, 1] and $D_i = X_i - X_{i-1}$, $i = 1, 2, \ldots, n$, be the uniform spacings, where $X_0 = 0$ and $X_n = 1$. Statistics of the form $\sum_{i=1}^{n} h(nD_i)$ are of interest in goodness of fit problems, where h is a real valued function defined on $(0, \infty)$. A general method available for proving limit theorems for functions of spacings which uses a conditional approach, was introduced by LeCam (1958). See also Pyke (1965) and Rao and Sethuraman (1975).

Pyke (1972, p. 419) poses the following "striking" open question. Let $\{N(t): t > 0\}$ be a positive integer-valued process for which $N(t)/t \xrightarrow{P} 1$ as $t \to \infty$. If V_n, a spacings statistic, converges in distribution as n tends to infinity, does the same weak limit hold for $V_{N(t)}$ as $t \to \infty$? The main theorem in this paper solves the above problem for a large class of statistics V_n.

Random sample size versions of limit theorems have previously been obtained for example, for sums, maxima and empirical processes of independent, identically distributed random variables; but nothing has been obtained for similar functions based on uniform spacings which happen to be exchangeable random variables.

Section 2 discusses the main theorem while an important application, a goodness of fit test for a Poisson process is discussed in Section 3.

2. The main theorem

Let $Z_1, Z_2, \ldots,$ be an independent, identically distributed sequence of exponential random variables with mean one and $\bar{Z}_n = n^{-1}(Z_1 + Z_2 + \cdots + Z_n)$. Then it is well known (see Pyke, 1965) that

$$(nD_1, \ldots, nD_n) \sim (Z_1/\bar{Z}_n,\ Z_2/\bar{Z}_n, \ldots, Z_n/\bar{Z}_n) \tag{1}$$

where $\sim$ means that the quantities on either side have the same distribution. We use this crucial fact to establish the following theorem.

Theorem 1. *Assume that*

(A) (i) *h is differentiable in $(0, \infty)$ and either h' is bounded on any closed interval in $(0, \infty)$ and monotone in the neighborhood of 0 and ∞, or h' is bounded on $(0, \infty)$;*

(ii) $\int_0^\infty h^2(x)\, e^{-x}\, dx < \infty$;

(iii) *there exists an $\alpha < 1$ such that $\int_0^\infty (xh'(x))^2\, e^{-\alpha x}\, dx < \infty$;*

(B) *$\{N(t): t > 0\}$ is a positive integer-valued process such that $N(t)/t \xrightarrow{P} 1$ as $t \to \infty$.*

Let

$$V_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [h(nD_i) - Eh(Z_i)]. \tag{2}$$

If (A) *and* (B) *hold, then $V_{N(t)}$ converges in distribution to a normal random variable V with mean zero and variance $\sigma^2 = [\mathrm{Var}(h(Z_1)) - \mathrm{Cov}(h(Z_1), Z_1)]$.*

All the standard examples of $h(\cdot)$ including $h(x) = x^r$, $(r > -\frac{1}{2})$, $h(x) = \log x$, $h(x) = x \log x$ satisfy our assumption (A). The proof of the above theorem depends on the corresponding result for V_n (for non-random sample size n) and a theorem of Doeblin (1938) and Anscombe (1952). We now state the result for non-random sample size.

Theorem 2. *Under the assumption* (A) *of Theorem* 1, $V_n \xrightarrow{D} N(0, \sigma^2)$ *where* $\sigma^2 = \mathrm{Var}(h(Z_1)) - \mathrm{Cov}^2(h(Z_1), Z_1)$. □

A proof of this result may be obtained for instance, by specializing Theorem 3 of Sethuraman and Rao (1969) or Theorem 4.2 of Kuo and Rao (1981) which discusses tests based on higher-order spacings. For the Doeblin–Anscombe theorem, the crucial concept is that of uniform continuity in probability of a sequence of random variables which we now define.

Definition 1. Let $\{Y_{n,k}: k = 1, 2, \ldots, n;\ n = 1, 2, \ldots\}$ be a triangular array of random variables. It is said to be *uniformly negligible in probability* (u.n.i.p.) if and only if for every $\varepsilon > 0$ there exist a $\delta > 0$ such that

$$P\Big\{ \max_{0 \leqslant k \leqslant n\delta} |Y_{n,k}| > \varepsilon \Big\} < \varepsilon \quad \text{for all } n \geqslant 1.$$

Definition 2. A sequence $\{W_n: n \geqslant 1\}$ of random variables is said to be *uniformly continuous in probability* (u.c.i.p.) if and only if $\{Y_{n,k} = W_{n+k} - W_n: k = 1, 2, \ldots, n;\ n = 1, 2, \ldots\}$ is u.n.i.p.

Note that $\{W_n,\ n \geqslant 1\}$ is u.c.i.p. if it converges to a finite limit with probability one as $n \to \infty$.

Definition 3. Let $\{Y_{n,k}\}$ be as in Definition 1. It is said to be *uniformly bounded in probability* (u.b.i.p.) if and only if for every $\varepsilon > 0$ there exist $\delta > 0$ and $M > 0$ such that

$$P\Big(\max_{0 \leqslant k \leqslant n\delta} |Y_{n,k}| > M \Big) < \varepsilon \quad \text{for all } n \geqslant 1.$$

Theorem 3 (Doeblin–Anscombe). *Suppose that $\{Y_n,\ n \geqslant 1\}$ is u.c.i.p. Let $\{N(t),\ t > 0\}$, be as in condition* (B) *of Theorem* 1 *and let $M(t)$ be the integer part of t. Then $Y_{N(t)} - Y_{M(t)} \to 0$ in probability as $t \to \infty$. As a consequence if Y_n converges in distribution to a random variable Y, then $Y_{N(t)}$ does the same as $t \to \infty$.* □

Example 1. If $W_1, W_2, \ldots,$ are i.i.d. random variables with finite mean μ and finite variance σ^2, then $Y_n = (S_n - n\mu)/(\sigma\sqrt{n})$, $n \geqslant 1$, is u.c.i.p., where $S_n = \sum_{i=1}^n W_i$. Also if $\mu = 0$, then $\{Y_{n,k} = (1/\sqrt{n})(S_{n+k} - S_n)$: $n \geqslant 1\}$ is u.n.i.p.

The reader is referred to Woodroofe (1982, p. 10) for the proofs of Theorem 3 as well as the first part of the Example 1. The second part of the example follows from the Kolmogorov's inequality. The following two propositions are straightforward.

Proposition 1. *If* $\{Y_{n,k}\}$ *and* $\{Z_{n,k}\}$ *are u.n.i.p., then* $\{Y_{n,k}+Z_{n,k}\}$ *is u.n.i.p.* □

Proposition 2. *If* $\{Y_{n,k}\}$ *is u.n.i.p. and* $\{W_{n,k}\}$ *is u.b.i.p., then* $\{Y_{n,k}\cdot W_{n,k}\}$ *is u.n.i.p.* □

Now we are ready to prove our main theorem.

Proof of Theorem 1. Let

$$U_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{h(Z_i/\bar{Z}_n)-Eh(Z_1)\}. \tag{3}$$

Then $\{V_n,\ n\geqslant 1\}$ defined in (2) and $\{U_n,\ n\geqslant 1\}$ have the same distribution as a result of representation (1). Thus Theorems 2 and 3 would imply our Theorem 1 if $\{U_n\colon n\geqslant 1\}$ is u.c.i.p., which is what we establish below. Now,

$$\begin{aligned}U_{n+k}-U_n&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{h(Z_i/\bar{Z}_{n+k})-h(Z_i/\bar{Z}_n)\}+\frac{1}{\sqrt{n}}\sum_{i=n+1}^{n+k}\{h(Z_i/\bar{Z}_{n+k})-h(Z_i)\}\\&\quad+\frac{1}{\sqrt{n}}\sum_{i=n+1}^{n+k}\{h(Z_i)-Eh(Z_1)\}\\&=A_{n,k}+B_{n,k}+C_{n,k},\quad\text{say.}\end{aligned}$$

Since $\{C_{n,k}\}$ is u.n.i.p. by Example 1 and condition (A)(ii), we need only to prove that $\{A_{n,k}\}$ and $\{B_{n,k}\}$ are also u.n.i.p. because of Proposition 1. By assumption (A)(i) and the mean value theorem,

$$A_{n,k}=\sqrt{n}\,(\bar{Z}_n-\bar{Z}_{n+k})\frac{1}{n}\sum_{i=1}^{n}\frac{h'(\xi_{ink})Z_i}{\bar{Z}_{n+k}\bar{Z}_n}, \tag{4}$$

where ξ_{ink} lies between $Z_i/\bar{Z}_{n+k}$ and $Z_i/\bar{Z}_n$. It can be easily checked that

$$\sqrt{n}\,(\bar{Z}_n-\bar{Z}_{n+k})=\frac{k}{n+k}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i-1)-\frac{\sqrt{n}}{n+k}\sum_{i=n+1}^{n+k}(Z_i-1). \tag{5}$$

Since for $k\leqslant n\delta$, $k/(n+k)<\delta$ and $(1/\sqrt{n})\sum_{i=1}^{n}(Z_i-1)$ is stochastically bounded, the first term on the right hand side of (5) is u.n.i.p. Also since $\sqrt{n}/(n+k)\leqslant 1/\sqrt{n}$, the Kolmogorov inequality implies that the second term on the right hand side of (5) is u.n.i.p. From Proposition 1, $\sqrt{n}\,(\bar{Z}_n-\bar{Z}_{n+k})$ is u.n.i.p. Also from the remark after Definition 2, $(\bar{Z}_{n+k}\bar{Z}_n)^{-1}$ is u.b.i.p. Thus if we show that $(1/n)\sum_{i=1}^{n}h'(\xi_{ink})Z_i$ is u.b.i.p, then Propositions 1 and 2 would imply that $\{A_{n,k}\}$ is u.n.i.p. Note that if $\max_{0\leqslant k\leqslant n\delta}|\bar{Z}_{n+k}-1|<\theta$, then

$$Z_i/(1+\theta)\leqslant\xi_{ink}\leqslant Z_i/(1-\theta)\quad\text{for every } i=1,2,\ldots,n \text{ and } k\leqslant n\delta. \tag{6}$$

Assumption (A)(i) implies that there exist $0<a<b<\infty$ and $M>0$ such that either $|h'(x)|\leqslant M$ for $x\in[a,b]$ and $|h'(x)|$ is monotone in $(0,a)$ and (b,∞), or $|h'(x)|\leqslant M$ for every x in $(0,\infty)$. Thus (6) implies that

$$|h'(\xi_{ink})|\leqslant M+|h'(Z_i/(1+\theta))|+|h'(Z_i/(1-\theta))| \tag{7}$$

for every $i = 1, 2, \ldots, n$ and for all $k \leqslant n\delta$. Let α be as in (A)(iii). Then

$$E|h'(Z_1/\alpha)Z_1| = \alpha^2 \int_0^\infty |h'(y)|\, y\, e^{-\alpha y}\, dy < \infty$$

and similarly,

$$E|h'(Z_1/(2-\alpha))Z_1| = (2-\alpha)^2 \int_0^\infty |h'(y)|\, y\, e^{-(2-\alpha)y}\, dy < \infty.$$

since $2 - \alpha > \alpha$. Thus, on the event $\max_{0 \leqslant k \leqslant n\delta} |\bar{Z}_{n+k} - 1| \leqslant \theta = 1 - \alpha$,

$$\max_{0 \leqslant k \leqslant n\delta} \left| \frac{1}{n} \sum_{i=1}^{n} h'(\xi_{ink}) Z_i \right| \leqslant \frac{M}{n} \sum_{i=1}^{n} Z_i + \frac{1}{n} \sum_{i=1}^{n} |h'(Z_i/(2-\alpha))Z_i| + \frac{1}{n} \sum_{i=1}^{n} |h'(Z_i/\alpha)Z_i|. \quad (8)$$

Consequently, by the law of large numbers, for every $\varepsilon > 0$, $\exists M_1 > 0$ such that

$$P\left(\max_{0 \leqslant k \leqslant n\delta} \left| \frac{1}{n} \sum_{i=1}^{n} h'(\xi_{ink}) Z_i \right| > M_1,\ \max_{0 \leqslant k \leqslant n\delta} |\bar{Z}_{n+k} - 1| \leqslant 1 - \alpha \right) < \tfrac{1}{2}\varepsilon \quad (9)$$

for every n. On the other hand, an application of the Chebysev's inequality gives that for sufficiently small δ,

$$\begin{aligned} P\left(\max_{0 \leqslant k \leqslant n\delta} |\bar{Z}_{n+k} - 1| > 1 - \alpha \right) &\leqslant \sum_{k \leqslant n\delta} P\left(|\bar{Z}_{n+k} - 1| > 1 - \alpha \right) \leqslant \frac{1}{(1-\alpha)^2} \sum_{k \leqslant n\delta} E(\bar{Z}_{n+k} - 1)^2 \\ &= \frac{1}{(1-\alpha)^2} \sum_{k \leqslant n\delta} \frac{1}{n+k} \leqslant \frac{\delta}{(1-\alpha)^2} < \tfrac{1}{2}\varepsilon \end{aligned} \quad (10)$$

for every n. Thus (9) and (10) imply that $(1/n)\sum_{i=1}^{n} h'(\xi_{ink}) Z_i$ is u.b.i.p. and hence $\{A_{n,k}\}$ is u.n.i.p. as noted earlier.

Applying the mean value theorem once more, we obtain

$$B_{n,k} = \frac{1}{\sqrt{n}} \sum_{i=n+1}^{n+k} h'(\zeta_{ink}) \left(\frac{Z_i}{\bar{Z}_{n+k}} - Z_i \right)$$

where ζ_{ink} is between $Z_i/\bar{Z}_{n+k}$ and Z_i. Similar to the arguments leading to (8), we can show that if $\max_{0 \leqslant k \leqslant n\delta} |\bar{Z}_{n+k} - 1| \leqslant 1 - \alpha$, then

$$\begin{aligned} |B_{n,k}| \leqslant{}& \frac{1-\alpha}{\alpha} \frac{M}{\sqrt{n}} \sum_{i=n+1}^{n+k} |Z_i - 1| \\ &+ \frac{1-\alpha}{\alpha} \frac{1}{\sqrt{n}} \sum_{i=n+1}^{n+k} \left\{ |h'(Z_i/(2-\alpha))Z_i| - E\,|h'(Z_1/(2-\alpha))Z_1| \right\} \\ &+ \frac{1-\alpha}{\alpha} \frac{1}{\sqrt{n}} \sum_{i=n+1}^{n+k} \left\{ |h'(Z_i/\alpha)Z_i| - E\,|h'(Z_1/\alpha)Z_1| \right\} \\ &+ \frac{M}{\alpha} \frac{k}{\sqrt{n}} |\bar{Z}_{n+k} - 1| \left\{ M + E\,|h'(Z_1/(2-\alpha))Z_1| + E\,|h'(Z_1/\alpha)Z_1| \right\}. \end{aligned} \quad (11)$$

Note that the first three terms in the right hand side of (11) are u.n.i.p. by the Kolmogorov inequality and condition (A)(iii). Moreover,

$$\frac{k}{\sqrt{n}} |\bar{Z}_{n+k} - 1| \leqslant \frac{k}{n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - 1) \right| + \left| \frac{1}{\sqrt{n}} \sum_{i=n+1}^{n+k} (Z_i - 1) \right|. \quad (12)$$

The first term of the right hand side in (12) is obviously u.n.i.p. and so is the second term by the Kolmogorov inequality. Thus the left hand side of (12) is u.n.i.p., which implies that the last term in (11) is also u.n.i.p. It follows that $B_{n,k}$ is u.n.i.p. and the proof is complete. □

3. Application and extensions

Consider a renewal process, a sequence of independent non-negative random variables $\{X_i, i \geq 1\}$ with common distribution function F. For $t > 0$, let $N(t) = \max\{k \geq 0: X_1 + X_2 + \cdots + X_k \leq t\}$ denote the number of observations observed up to time t. Thus the sample takes the form $X_1, X_2, \ldots, X_{N(t)}$ and $t - X_1 - X_2 - \cdots - X_{N(t)}$. Of particular interest is the theory of spacings corresponding to the special case in which F is the exponential distribution function with mean 1, that is to say, under the null-hypothesis, the renewal process is a Poisson process with parameter 1. Consider the sample in the modified form $t^{-1}X_1, t^{-1}X_2, \ldots, t^{-1}X_{N(t)}, 1 - t^{-1}(X_1 + X_2 + \cdots + X_{N(t)})$, so that conditionally, given $N(t) = n$, the sample is equivalent to the uniform spacings from a sample of n independent uniform random variables. There are several popular goodness of fit tests based on uniform spacings (see Pyke, 1965). Theorem 1 permits us to use the asymptotic distribution theory for the non-random sample size case to the above described random sample size case.

An illustrative example. Suppose that a fire station received $N(t) = 20$ calls in a particular $t = 24$ hour period and we wish to test if these are uniformly distributed over the entire day corresponding to a Poisson process or they tend to cluster around some particular time of the day. Suppose the calls are received at the following times:

1.10, 4.30, 6.00, 6.10, 7.00, 8.00, 8.30, 8.45, 9.30, 10.05, 13.00, 14.10, 16.00, 17.50,

19.30, 21.15, 22.00, 22.15, 23.00, 23.30.

We compare $D_i = X_i/t$, $i = 1, \ldots, N(t)$ and $D_{N(t)+1} = 1 - t^{-1}(X_1 + \cdots + X_{N(t)})$ where the $\{X_i\}$ are the inter-arrival times. From the theory of spacings, we know that (cf. Sethuraman and Rao, 1970) $\sum_{i=1}^{n+1}((n+1)D_i)^2$ is approximately $N(2(n+1), 4n)$ for large enough n (non-random). But from Theorem 1, this asymptotic normal distribution can also be used to find the critical values for the random sample size case. In this example, Theorem 1 shows that the statistic

$$T_t = \frac{1}{2\sqrt{N(t)}}\left(\sum_{i=1}^{N(t)+1}\left[(N(t)+1)D_i\right]^2 - 2(N(t)+1)\right)$$

is approximately a standard normal random variable and we compute the observed $T_t \approx -1.120$ based on the above data. This value is not significant even at level 0.10. Thus we would not reject the hypothesis that the calls arrived uniformly in the day.

Theorem 1 easily extends to the asymptotic distribution theory developed for test statistics based on m-step spacings for any finite m. See for instance Kuo and Rao (1981). Whether the extension to random sample size, also holds when m depends on n and goes to ∞ (as in Hall, 1986; Jammalamadaka, Zhou and Tiwari, 1986), needs further investigation and possibly additional restrictions on the class of such statistics.

One might also consider sequential test procedures in which the test statistic is based on spacings. Although none have been developed to the knowledge of the authors, Theorem 1 allows one to use the large sample distribution theory for spacings, to be applied in such a situation.

References

Anscombe, F. (1952), Large sample theory of sequential estimation, *Proc. Cambridge Philos. Soc.* **48**, 600–607.

Doeblin, W. (1938), Sur deux problèmes de M. Kolmogorov concernant des chaines de nombrables, *Bull. Soc. Math. France* **66**, 210–220.

Hall, P. (1986), On powerful distributional tests based on sample spacings, *J. Multivariate Anal.* **19**, 201–224.

Jammalamadaka, S.R., X. Zhou and R. Tiwari (1986), Asymptotic efficiencies of spacings tests for goodness of fit, Tech. Rept. No. 4, Statistics Program, Univ. of California (Santa Barbara, CA).

Kuo, M. and J.S. Rao (1981), Limit theory and efficiencies for tests based on higher order spacings, *Proc. Indian Statist. Inst. Golden Jubilee Internat. Conf. Statist.: Application and New Directions*, pp. 333–352.

LeCam, L. (1958), Un théorème sur la division d'une intervalle par des points près au hasard, *Publ. Inst. Statist. Univ. Paris* **7**, 7–16.

Pyke, R. (1965), Spacings, *J. Roy. Statist. Soc. Ser. B* **7**, 395–449.

Pyke, R. (1972), Spacings revisited, *Sixth Berkeley Symp. Math. Statist. Probab., Vol. 1*, pp. 417–427.

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.

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 Univ. Press, London/New York) pp. 267–273.

Woodroofe, M. (1982), *Nonlinear Renewal Theory in Sequential Analysis* (Soc. Indust. Appl. Math., Philadelphia, PA).