Z. Wahrscheinlichkeitstheorie verw. Gebiete
51, 55–62 (1980)

Zeitschrift für
Wahrscheinlichkeitstheorie
und verwandte Gebiete

A Note on Limiting Distributions for Spacings Statistics

James A. Koziol

Department of Mathematics, University of California, San Diego,
La Jolla, California 92093, USA

Summary. Hájek's projection method is used to prove asymptotic normality for a class of spacings statistics.

1. Introduction

Let $X_1, X_2, \ldots, X_n$ be n independent random variables uniformly distributed on $[0,1]$, and let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the order statistics obtained by arranging the X_i's in increasing order. Set $X_{(0)}=0$ and $X_{(n+1)}=1$. Then the spacings of the sample are defined by

$$D_i = X_{(i)} - X_{(i-1)}, \qquad i=1,2,\ldots,n+1. \tag{1.1}$$

Let

$$S_n = \sum_{i=1}^{n+1} g(D_i), \tag{1.2}$$

where g is a square integrable function on the unit interval. From the exchangeability of the D_i,

$$E[S_n] = (n+1)\,E[g(D_1)] = (n+1)\int_0^1 g(t)\,n(1-t)^{n-1}\,dt, \tag{1.3}$$

$$\begin{aligned} E[S_n^2] &= (n+1)\,E[g^2(D_1)] + (n+1)\,nE[g(D_1)\,g(D_2)] \\ &= (n+1)\int_0^1 g^2(t)\,n(1-t)^{n-1}\,dt \\ &\quad + (n+1)n \iint\limits_{\substack{s,t\ge 0\\ 0\le s+t\le 1}} g(s)\,g(t)\,n(n-1)(1-s-t)^{n-2}\,ds\,dt. \end{aligned} \tag{1.4}$$

Interest in the statistics (1.2) arose from suggestions by Greenwood (1946), Kendall, Kimball (1947, 1950) and others that such statistics might be useful for

0044-3719/80/0051/0055/$01.60

assessing goodness of fit. An excellent survey of results concerning spacings tests was made by Pyke (1965). As Pyke noted, there are two general methods of deriving limit theorems for statistics of the form (1.2): Darling (1953) found a simple formula for the characteristic function of S_n, which led to its limiting distribution, and LeCam (1958) approached the problem by exploiting the well-known construction of uniform spacings as exponential random variables proportional to their sum. In this note it is shown that asymptotic distributional results can also be obtained using Hájek's projection method. Indeed, in view of asymptotic results concerning linear rank statistics [Hájek, 1968; Dupăc and Hájek, 1969] and linear combinations of order statistics [Stigler, 1969, 1974] derived upon using the projection method, it is not at all surprising that the method yields useful results with spacings statistics also. The projection approximation of S_n by a sum of independent random variables is given in the next section. In the concluding section, S_n and its projection are shown to be asymptotically a.s. equivalent in mean square, from whence the asymptotic normality of S_n follows directly.

2. Hájek's Projection Method

Hájek's projection method will be used to find a sum of independent random variables that well approximates S_n in mean square; it will be shown that this sum and S_n are asymptotically equivalent. For future reference, Hájek's (1968) projection lemma is now stated.

Lemma 1. *Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables, and let H be the Hilbert space of a.s. equivalence classes of square integrable statistics depending on $Y_1, Y_2, \ldots, Y_n$. Let L be the closed linear subspace of H consisting of statistics of the form $L=\sum_{i=1}^{n} l_i(Y_i)$, where the l_i are functions such that $El_i^2(Y_i)<\infty$. If*

$$S=S(Y_1, Y_2, \ldots, Y_n)\in H,$$

then the projection of S on L is given by

$$\hat{S}=\sum_{i=1}^{n} E(S\,|\,Y_i)-(n-1)ES.$$

Moreover, $E\hat{S}=ES$ and $E(S-\hat{S})^2=\sigma^2(S)-\sigma^2(\hat{S})$.

The projection $\hat{S}_n$ of S_n is given in the following lemma.

Lemma 2. *For any fixed n, let $Y_1 \leqq Y_2 \leqq \ldots \leqq Y_{n-1}$ denote the order statistics among $X_1, X_2, \ldots, X_{n-1}$; set $Y_0=1$. Let f_i denote the density function of Y_i, $1 \leqq i \leqq n-1$, let f_{ij} denote the joint density function of Y_i and Y_j, and let $dF(\underline{Y})$ be the joint density of all the Y_i. Then the projection $\hat{S}_n$ of S_n may be written*

$$\hat{S}_n = nES_{n-1} - (n-1)ES_n$$
$$+ \sum_{k=1}^{n} \int \dots \int \sum_{j=0}^{n-1} I[Y_j < X_k < Y_{j+1}][g(X_k - Y_j) + g(Y_{j+1} - X_k)$$
$$- g(Y_{j+1} - Y_j)] \, dF(\underline{Y}), \tag{2.1}$$

where I is the standard indicator function.

Proof. The proof is motivated by that of a proposition of Stigler (1974). From (1.1),

$$D_1 = X_{(1)} = \min[Y_1, X_n] = \min[Y_1 - Y_0, X_n - Y_0].$$

Hence

$$E[g(D_1)|X_n] = \int_0^1 \{I[Y_0 < X_n < Y_1] \, g(X_n - Y_0)$$
$$+ I[Y_1 < X_n] \, g(Y_1 - Y_0)\} f_1(Y_1) \, dY_1.$$

Similarly, since

$$D_{n+1} = 1 - X_{(n)} = 1 - \max[Y_{n-1}, X_n] = \min[Y_n - Y_{n-1}, Y_n - X_n],$$

$$E[g(D_{n+1})|X_n] = \int_0^1 \{I[X_n < Y_{n-1}] \, g(Y_n - Y_{n-1})$$
$$+ I[Y_{n-1} < X_n < Y_n] \, g(Y_n - X_n)\} f_{n-1}(Y_{n-1}) \, dY_{n-1}.$$

Next, write D_2 as

$$D_2 = X_{(2)} - X_{(1)} = \begin{cases} Y_1 - X_n & \text{if } Y_0 < X_n < Y_1 \\ X_n - Y_1 & \text{if } Y_1 < X_n < Y_2 \\ Y_2 - Y_1 & \text{if } Y_2 < X_n. \end{cases}$$

It follows that

$$E[g(D_2)|X_n] = \iint_{0 \leq Y_1 \leq Y_2 \leq 1} \{I[Y_0 < X_n < Y_1] \, g(Y_1 - X_n)$$
$$+ I[Y_1 < X_n < Y_2] \, g(X_n - Y_1)$$
$$+ I[Y_2 < X_n] \, g(Y_2 - Y_1)\} f_{1,2}(Y_1, Y_2) \, dY_1 \, dY_2.$$

Similarly,

$$E[g(D_n)|X_n] = \iint_{0 \leq Y_{n-2} \leq Y_{n-1} \leq 1} \{I[X_n < Y_{n-2}] \, g(Y_{n-1} - Y_{n-2})$$
$$+ I[Y_{n-2} < X_n < Y_{n-1}] \, g(Y_{n-1} - X_n)$$
$$+ I[Y_{n-1} < X_n < Y_n] \, g(X_n - Y_{n-1})\}$$
$$\cdot f_{n-2,n-1}(Y_{n-2}, Y_{n-1}) \, dY_{n-2} \, dY_{n-1}.$$

In general, for $3 \leq i \leq n-1$,

$$
\begin{aligned}
E[g(D_i)|X_n] = \int \dots \int \{ & I[X_n < Y_{i-2}]\, g(Y_{i-1} - Y_{i-2}) \\
& + I[Y_{i-2} < X_n < Y_{i-1}]\, g(Y_{i-1} - X_n) \\
& + I[Y_{i-1} < X_n < Y_i]\, g(X_n - Y_{i-1}) \\
& + I[Y_i < X_n]\, g(Y_i - Y_{i-1})\}\, dF(\underline{Y}).
\end{aligned}
$$

The X_i are independent, identically distributed; therefore, $E[g(D_i)|X_k]$ can be found from $E[g(D_i)|X_n]$ merely by replacing X_n by X_k in the appropriate formulas. Because

$$
\begin{aligned}
E[S_n|X_k] &= E\left[\sum_{i=1}^{n+1} g(D_i)|X_k\right] \\
&= \sum_{i=1}^{n+1} E[g(D_i)|X_k],
\end{aligned}
$$

suitable reduction of $\sum_{i=1}^{n+1} E[g(D_i)|X_k]$ must be found. Such a reduction is afforded by the identities

$$
I[X_k < Y_{i-2}] = \sum_{j=0}^{i-3} I[Y_j < X_k < Y_{j+1}], \qquad i \geqq 3,
$$

and

$$
I[Y_i < X_k] = \sum_{j=i}^{n-1} I[Y_j < X_k < Y_{j+1}], \qquad i \geqq 1.
$$

Then,

$$
\begin{aligned}
& \sum_{i=3}^{n+1} g(Y_{i-1} - Y_{i-2}) I[X_k < Y_{i-2}] \\
&\quad = \sum_{i=3}^{n+1} g(Y_{i-1} - Y_{i-2}) \sum_{j=0}^{i-3} I[Y_j < X_k < Y_{j+1}] \\
&\quad = \sum_{j=0}^{n-4} I[Y_j < X_k < Y_{j+1}] \sum_{i=3+j}^{n-1} g(Y_{i-1} - Y_{i-2}),
\end{aligned}
$$

and

$$
\begin{aligned}
& \sum_{i=1}^{n-1} g(Y_i - Y_{i-1}) I[Y_i < X_k] \\
&\quad = \sum_{i=1}^{n-1} g(Y_i - Y_{i-1}) \sum_{j=2}^{n-1} I[Y_j < X_k < Y_{j+1}] \\
&\quad = \sum_{j=1}^{n-1} I[Y_j < X_k < Y_{j+1}] \sum_{i=1}^{j} g(Y_i - Y_{i-1}).
\end{aligned}
$$

Also,

$$
\begin{aligned}
I[Y_0 < X_k < Y_1]\, g(X_k - Y_0) &+ I[Y_0 < X_k < Y_1] g(Y_1 - X_k) \\
&+ I[Y_1 < X_k < Y_2] g(X_k - Y_1)
\end{aligned}
$$

$$+\sum_{i=3}^{n-1}\{I[Y_{i-2}<X_k<Y_{i-1}]g(Y_{i-1}-X_k)+I[Y_{i-1}<X_k<Y_i]g(X_k-Y_{i-1})\}$$
$$+I[Y_{n-2}<X_k<Y_{n-1}]g(Y_{n-1}-X_k)+I[Y_{n-1}<X_k<Y_n]g(X_k-Y_{n-1})$$
$$+I[Y_{n-1}<X_k<Y_n]g(Y_n-X_k)$$
$$=\sum_{i=1}^{n}I[Y_{i-1}<X_k<Y_i]g(Y_i-X_k)+\sum_{i=0}^{n-1}I[Y_i<X_k<Y_{i+1}]g(X_k-Y_i)$$
$$=\sum_{i=0}^{n-1}I[Y_i<X_k<Y_{i+1}][g(X_k-Y_i)+g(Y_{i+1}-X_k)].$$

Collecting all terms,

$$\sum_{i=1}^{n+1}E[g(D_i)|X_k]$$
$$=\int\dots\int\Bigg\{I[Y_0<X_k<Y_1]\left[g(X_k-Y_0)+g(Y_1-X_k)+\sum_{j=1}^{n-1}g(Y_{j+1}-Y_j)\right]$$
$$+\sum_{l=1}^{n-2}I[Y_l<X_k<Y_{l+1}]\Bigg[\sum_{j=1}^{l}g(Y_j-Y_{j-1})+g(X_k-Y_l)$$
$$+g(Y_{l+1}-X_k)+\sum_{j=l+1}^{n-1}g(Y_{j+1}-Y_j)\Bigg]$$
$$+I[Y_{n-1}<X_k<Y_n]\left[\sum_{j=1}^{n-1}g(Y_j-Y_{j-1})+g(X_k-Y_{n-1})+g(Y_n-X_k)\right]\Bigg\}dF(\underline{Y}).$$

Observe that the integrand in (2.1) is

$$\sum_{l=0}^{n-1}I[Y_l<X_k<Y_{l+1}]$$
$$\cdot\left\{\sum_{j=0}^{n-1}g(Y_{j+1}-Y_j)+g(X_k-Y_l)+g(Y_{l+1}-X_k)-g(Y_{l+1}-Y_l)\right\}$$
$$=\sum_{j=0}^{n-1}g(Y_{j+1}-Y_j)+\sum_{j=0}^{n-1}I[Y_j<X_k<Y_{j+1}]$$
$$\cdot[g(X_k-Y_j)+g(Y_{j+1}-X_k)-g(Y_{j+1}-Y_j)],$$

because $Y_0<X_k<Y_n$. Therefore,

$$\sum_{i=1}^{n+1}E[g(D_i)|X_k]=ES_{n-1}+\int\dots\int\Bigg\{\sum_{j=0}^{n-1}I[Y_j<X_k<Y_{j+1}]$$
$$\cdot[g(X_k-Y_j)+g(Y_{j+1}-X_k)-g(Y_{j+1}-Y_j)]\Bigg\}dF(\underline{Y}).$$

Hence

$$\hat{S}_n=\sum_{k=1}^{n}E[S_n|X_k]-(n-1)ES_n$$
$$=\sum_{k=1}^{n}E\left[\sum_{i=1}^{n+1}g(D_i)|X_k\right]-(n-1)ES_n$$

$$=nES_{n-1}-(n-1)ES_n$$
$$+\sum_{k=1}^{n}\int\cdots\int\sum_{j=0}^{n-1}I[Y_j<X_k<Y_{j+1}]$$
$$\cdot\,[g(X_k-Y_j)+g(Y_{j+1}-X_k)-g(Y_{j+1}-Y_j)]\,dF(\underline{Y}),$$

as was to be shown.

3. The Main Result

For particular choices of the function g, it is possible to calculate explicitly the mean and variance of S_n. Attention will henceforth be restricted to one of these cases, namely, $g(x)=x^r$, where r is any positive constant [excluding the trivial case $r=1$]. From (1.3) and (1.4), it follows that

$$E[S_n]=\Gamma(r+1)/n^{r-1},$$
$$E[S_n^2]=(n+1)n\frac{\Gamma(r+1)\Gamma(n)}{\Gamma(r+n+1)}+(n-1)n^2(n+1)\frac{\Gamma^2(r+1)\Gamma(n-1)}{\Gamma(2r+n+1)}.$$

From the projection lemma, it is known that $E[\hat{S}_n]=E[S_n]$. Hence to establish that S_n and $\hat{S}_n$ are asymptotically a.s. equivalent in mean square, it is sufficient to show that

$$\frac{E[\hat{S}_n^2]}{E[S_n^2]}\sim 1 \quad \text{as } n\to\infty. \tag{3.1}$$

In order to demonstrate (3.1), note first that

$$E[S_n^2]=(n-1)n^2(n+1)\frac{\Gamma^2(r+1)\Gamma(n-1)}{\Gamma(2r+n+1)}+O(n^2). \tag{3.2}$$

Next, since

$$\frac{nES_{n-1}}{(n-1)ES_n}\sim 1 \quad \text{as } n\to\infty,$$

it follows that

$$E[\hat{S}_n^2]\sim E\left\{\sum_{k=1}^{n}\int\cdots\int\sum_{j=0}^{n-1}I[Y_j<X_k<Y_{j+1}]\right.$$
$$\left.\cdot[g(X_k-Y_j)+g(Y_{j+1}-X_k)-g(Y_{j+1}-Y_j)]\,dF(\underline{Y})\right\}^2$$
$$\geqq(n^2-n)\left\{E\int\cdots\int\sum_{j=0}^{n-1}I[Y_j<X_1<Y_{j+1}]\right.$$
$$\left.\cdot[g(X_1-Y_j)+g(Y_{j+1}-X_1)-g(Y_{j+1}-Y_j)]\,dF(\underline{Y})\right\}^2$$
$$=(n-1)n^3\frac{\Gamma^2(r+1)\Gamma^2(n)}{\Gamma^2(r+n+1)}. \tag{3.3}$$

[(3.3) is established upon noting that

$$E\int\dots\int\sum_{j=0}^{n-1} I[Y_j<X_1<Y_{j+1}]\,g(Y_{j+1}-Y_j)\,dF(\underline{Y})$$
$$=\int\dots\int\sum_{j=0}^{n-1}(Y_{j+1}-Y_j)\,g(Y_{j+1}-Y_j)\,dF(\underline{Y})$$
$$=n\int t\,g(t)(n-1)(1-t)^{n-2}\,dt;$$

similarly,

$$E\int\dots\int\sum_{j=0}^{n-1} I[Y_j<X_1<Y_{j+1}][g(X_1-Y_j)+g(Y_{j+1}-X_1)]\,dF(\underline{Y})$$
$$=\sum_{j=0}^{n-1}\int\dots\int\left[\int_{y_j}^{y_{j+1}} g(X-Y_j)+g(Y_{j+1}-X)\,dX\right]dF(\underline{Y})$$
$$=\sum_{j=0}^{n-1}2\iint\left[\int_0^{y_{j+1}-y_j} g(t)\,dt\right]f_{j,j+1}(Y_j,Y_{j+1})\,dY_j\,dY_{j+1}$$
$$=2n\int_0^1 g(t)(1-t)^{n-1}\,dt,$$

by changing variables and inverting the order of integration.]

By applying Stirling's formula in (3.2) and (3.3), it may be shown that (3.1) holds.

From the central limit theorem, $\hat{S}_n$ is asymptotically normal; together with the a.s. mean square equivalence of S_n and $\hat{S}_n$, this proves the following theorem.

Theorem. *For* $g(x)=x^r$, $r>0$, $r\neq 1$, *the random variable* S_n *of* (1.2) *has a limiting normal distribution, that is,*

$$\lim_{n\to\infty}\Pr\left(\frac{S_n-E[S_n]}{\sigma[S_n]}<x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt.$$

This theorem was explicitly proved by Darling (1953), using the technique described previously.

One may readily extend this result to a more general class of functions g. First, it is clear that the theorem still obtains if g is an arbitrary polynomial. The following corollary may thereby be proved.

Corollary. *Let* $g(x)=g_1(x)-g_2(x)$, $0<x<1$, *where the* g_i *are nondecreasing, square integrable, and absolutely continuous inside* (0, 1). *Then the random variable* S_n *has a limiting normal distribution.*

Proof. The proof is patterned after that of Theorem 2.3 of Hájek (1968), so few details are given. From Lemma 5.1 of Hájek's paper, g may be decomposed as

$$g(x)=\psi(x)+\varphi_1(x)-\varphi_2(x),\qquad 0<x<1,$$

where ψ is a polynomial, φ_1 and φ_2 are nondecreasing, and

$$\int_0^1\varphi_1^2(x)\,dx+\int_0^1\varphi_2^2(x)\,dx<\alpha\qquad\text{for arbitrary }\alpha>0.$$

Denote

$$S_\psi = \sum_{i=1}^{n+1} \psi(D_i) \quad \text{and} \quad S_j = \sum_{i=1}^{n+1} \varphi_j(D_i), \qquad j=1,2,$$

and observe that $S_n = S_\psi + S_1 - S_2$. For arbitrary $\varepsilon > 0$, it is possible to choose $\alpha = \alpha(\varepsilon)$ so that

$$\|S_n - S_\psi\| = \|S_1 - S_2\| \leqq \|S_1\| + \|S_2\| \leqq \varepsilon \|S_n\|^{-1},$$

where $\|.\|$ denotes the norm on the Hilbert space of a.s. equivalence classes of square integrable functions on $[0,1]$ relative to the probability measure induced by the spacings. Since the relative contributions of S_1 and S_2 to S_n are asymptotically negligible compared to S_ψ, and since L_2 convergence implies convergence in distribution, it follows that S_n and S_ψ have the same limiting distribution. But this establishes the corollary.

The author thanks Dr. John Petkau for useful discussions.

References

1. Darling, D.A.: On a class of problems relating to the random division of an interval. Ann. Math. Statist. **24**, 239–253 (1953)
2. Dupăc, V., Hâjek, J.: Asymptotic normality of simple linear rank statistics under alternatives II. Ann. Math. Statist. **40**, 1992–2017 (1969)
3. Greenwood, M.: The statistical study of infectious diseases. J. Roy. Statist. Soc. A, **109**, 85–110 (1946)
4. Hâjek, J.: Asymptotic normality of simple linear rank statistics under alternatives. Ann. Math. Statist. **39**, 325–346 (1968)
5. Kimball, B.F.: Some basic theorems for developing tests of fit for the case of the nonparametric probability distribution function, I. Ann. Math. Statist. **18**, 540–548 (1947)
6. Kimball, B.F.: On the asymptotic distribution of the sum of powers of unit frequency differences. Ann. Math. Statist. **21**, 263–271 (1950)
7. LeCam, L.: Une théorème sur la division d'une intervalle par des points près au hasard. Pub. Inst. Statist. Univ. Paris, **7**, 7–16 (1958)
8. Pyke, R.: Spacings. J. Roy. Statist. Soc. B, **27**, 395–499 (1965)
9. Stigler, S.: Linear functions of order statistics. Ann. Math. Statist. **40**, 770–788 (1969)
10. Stigler, S.: Linear functions of order statistics with smooth weight functions. Ann. Statist. **2**, 676–693 (1974)

Received October 26, 1978; in revised form April 17, 1979