Asymptotic Spacings Theory with Applications to the Two-Sample Problem
Author(s): Lars Holst and J. S. Rao
Source: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, Vol. 9, No. 1 (1981), pp. 79-89
Published by: Statistical Society of Canada
Stable URL: http://www.jstor.org/stable/3315298
Accessed: 03/05/2009 10:35

The Canadian Journal of Statistics
Vol. 9, No. 1, 1981, Pages 79–89
La Revue Canadienne de Statistique

Asymptotic spacings theory with applications to the two-sample problem

Lars HOLST* and J. S. RAO

University of Uppsala and *University of California, Santa Barbara*

Key words and phrases: Two-sample problems, nonparametric methods, sample spacings, spacing frequencies, limit distributions, asymptotic efficiencies.
AMS 1980 *subject classifications:* Primary 62E20, 62G10, 62G05; secondary 60F05.

ABSTRACT

The asymptotic distribution theory of test statistics which are functions of spacings is studied here. Distribution theory under appropriate close alternatives is also derived and used to find the locally most powerful spacing tests. For the two-sample problem, which is to test if two independent samples are from the same population, test statistics which are based on "spacing-frequencies" (i.e., the numbers of observations of one sample which fall in between the spacings made by the other sample) are utilized. The general asymptotic distribution theory of such statistics is studied both under the null hypothesis and under a sequence of close alternatives.

1. INTRODUCTION

Let $X_1, \dots, X_{n-1}$ be $n - 1$ independently and identically distributed continuous real-valued random variables with a common distribution function (d.f.) $F(\cdot)$ whose support is [0, 1]. If $0 \le X'_{1n} \le \dots \le X'_{n-1,n} \le 1$ denote the order statistics, then the sample spacings are defined to be

$$D_{kn} = X'_{kn} - X'_{k-1,n}, \qquad k = 1, \dots, n, \tag{1.1}$$

with the notation $X'_{0n} = 0$ and $X'_{nn} = 1$. In particular, if $F(\cdot)$ is the uniform distribution on [0, 1], we will use the special symbols $\{U_k, i = 1, \dots, n-1\}$ for the sample, $0 = U'_{0n} \le U'_{1n} \le \cdots \le U'_{n-1,n} \le U'_{nn} = 1$ for the order statistics and

$$T_{kn} = U'_{kn} - U'_{k-1,n}, \qquad k = 1, \dots, n, \tag{1.2}$$

for the uniform spacings.

Tests based on sample spacings are useful in many statistical contexts, e.g., for goodness of fit, tests on Poisson processes, monotone failure rate, etc. For an excellent review of such problems, refer to Pyke (1965). See also Rao (1976) for applications connected with circular data.

We are interested in the asymptotic distribution theory of spacing statistics of the type

$$V_n = \sum_{k=1}^{n} h_{kn}(D_{kn}), \tag{1.3}$$

* Part of the work of this author was completed while he was at the Mathematics Research Center, University of Wisconsin, Madison, under the sponsorship of the U.S. Army under Contract No. DAAG29-75-C-0024.

where $\{h_{kn}(\cdot),\ 1 \leq k \leq n,\ n \geq 1\}$ is a sequence of real-valued Borel-measurable functions. The case where $h_{kn}(\cdot) = h_n(\cdot)$ for all k, i.e., where one takes a symmetric function of $\{D_{kn}\}$, has already been considered in the literature. See, for instance, Le Cam (1958), Pyke (1965) or Rao and Sethuraman (1975). Pyke (1965) also discusses generalizations in several directions. Holst (1979b) considers functions of higher-order spacings from a uniform distribution. But the cases of interest to us here are not readily available in the literature. In Section 2 the asymptotic distribution of V_n is studied when the spacings are from a fixed alternative d.f. $F(\cdot)$. In Section 3, the asymptotic theory of statistics of the type V_n is considered for close alternatives to the uniform distribution. The problem of finding the asymptotically locally most powerful spacings test is also discussed.

Next consider two independent samples of size $n - 1$ and m respectively from two continuous distributions, and the problem of testing that these are identical. By applying the usual probability integral transformation on both samples, we can suppose that the first sample is from the uniform distribution on $[0, 1]$ and that the second comes from the distribution F (say) on $[0, 1]$. Let $Y_1, \ldots, Y_m$ denote the (transformed) second sample, and set

$$S_k = \text{number of } Y_j\text{'s in } [U'_{k-1}, U'_k), \qquad k = 1, \ldots, n. \tag{1.4}$$

In Section 4 we shall study the asymptotic distribution of test statistics of the form

$$Q_n = \sum_{k=1}^{n} h_{kn}(S_k), \tag{1.5}$$

for close alternatives $F = F_n$ to the uniform distribution. Here $m, n \to \infty$ such that

$$m/n \to \rho, \qquad 0 < \rho < \infty. \tag{1.6}$$

Tests based on such spacing-frequencies $\{S_k\}$ have been considered for two-sample problems, for instance, by Dixon (1940), Godambe (1961), Blumenthal (1963, 1967), and Weiss (1976).

A few words about notation. Many quantities like U'_i, D_i, as well as the functions $\{h_{kn}(\cdot)\}$, depend on n. For notational convenience we shall suppress this suffix, except where it is essential. Convergence in distribution will be denoted by $\overset{D}{\to}$. We write $X \sim Y$ to denote that X has the same distribution as Y. We will let Z denote an exp(1) random variable (r.v.) with density e^{-z} for $z \geq 0$, η a geometric r.v. with $P(\eta = j) = \rho^j/(1 + \rho)^{j+1}$, and Poi($\lambda$) a Poisson r.v. with mean λ. For a r.v. X_n, we write $X_n = O_p(g(n))$ if $X_n/g(n) \to 0$ in probability and write $X_n = O_p(g(n))$ if for each $\varepsilon > 0$, there is a $K_\varepsilon < \infty$ such that $P\{\,|X_n/g(n)| > K_\varepsilon\} < \varepsilon$ for n sufficiently large.

2. ASYMPTOTIC DISTRIBUTION OF V_n UNDER A FIXED ALTERNATIVE

In this section, we establish the asymptotic normality of V_n defined in (1.3), under the following conditions. Here, as in the following, let

$$\xi_k = \xi_{kn} = (k - 0.5)/n. \tag{2.1}$$

(C1) The function $G = F^{-1}$ on $[0, 1]$ has derivative $G'(x) = g(x) > 0$ for all x and $g''(x)$ is continuous on $[0, 1]$.

(C2) The function $h_{kn}(x)$ is of the form

$$h_{kn}(x) = h(x, \xi_{kn}), \qquad k = 1, \ldots, n \text{ and } 0 < x < \infty,$$

for some function $h(x, y)$ defined on $[0, \infty) \times [0, 1]$. For any fixed y,

$$h(x, y), \qquad h_x = \frac{\partial}{\partial x} h \quad \text{and} \quad h_{xx} = \frac{\partial^2}{\partial x^2} h$$

are continuous and bounded by $c_1(x^{c_2} + 1)$ for some nonnegative constants c_1 and c_2.

We then have the following basic

THEOREM 2.1. *As $n \to \infty$, under assumptions* C1 *and* C2,

$$n^{-1/2} \sum_{k=1}^{n} \{h(nD_k, \xi_k) - \mathscr{E}h(Zg(\xi_k), \xi_k)\}$$
$$\xrightarrow{D} \mathscr{N}\left(0, \int_0^1 \mathscr{V}ar(h(Zg(x), x))\, dx - \left(\int_0^1 \mathscr{C}ov(Z, h(Zg(x), x))\, dx\right)^2\right), \tag{2.2}$$

where $\xi_k = (k - 0.5)/n$ and Z is an exp(1) *r.v.*

Proof. Without any loss of generality we can and shall assume that

$$\mathscr{E}h(Zg(x), y) = 0 \qquad \text{for all} \quad 0 \le x, y \le 1. \tag{2.3}$$

Since $X_i' \sim G(U_i')$, it is clear that for $m \le n - 2$,

$$n^{-1/2} \sum_{k=1}^{m} h(nD_k, \xi_k) \sim n^{-1/2} \sum_{k=1}^{m} h\left(n\left[G\left(\sum_{j=1}^{k} T_j\right) - G\left(\sum_{j=1}^{k-1} T_j\right)\right], \xi_k\right), \tag{2.4}$$

where $\{T_j\}$ are the uniform spacings defined in (1.2). Let $Z_1, Z_2, \ldots$ be a sequence of independently and identically distributed exp(1) random variables, and define $W_k = \sum_{j=1}^{k} Z_j$, $k = 1, 2, \ldots$, for the partial sums with $W_0 = 0$. Then by a lemma of Holst (1979b, p. 1069),

$$\mathscr{E}\left(\exp\left(itn^{-1/2} \sum_{k=1}^{m} h(nD_k, \xi_k)\right)\right) = \frac{1 + O(1/n)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathscr{E}\left(\exp\left(\frac{iu}{\sqrt{n}} \sum_{m+1}^{n} (Z_k - 1)\right)\right)$$
$$\times\, \mathscr{E}\left(\exp\left(\frac{i}{\sqrt{n}} \sum_{1}^{m} [th(n[G(W_k/n) - G(W_{k-1}/n)], \xi_k) + u(Z_k - 1)]\right)\right). \tag{2.5}$$

From Lemma 2.4, if $m, n \to \infty$ in such a way that $m/n \to \gamma \le 1$, then

$$\mathscr{E}\left(\exp\left(\frac{i}{\sqrt{n}} \sum_{1}^{m} [th(n[G(W_k/n) - G(W_{k-1}/n)], \xi_k) + u(Z_k - 1)]\right)\right)$$
$$\to \exp\left(-\frac{1}{2} \int_0^{\gamma} \mathscr{V}ar(th(Zg(x), x) + uZ)\, dx\right). \tag{2.6}$$

Now on applying the extended Lebesgue dominated-convergence theorem in (2.5) (see e.g., C. R. Rao 1973, p. 136), we have for $\gamma < 1$,

$$\mathscr{E}\left(\exp\left(itn^{-1/2} \sum_{1}^{m} h(nD_k, \xi_k)\right)\right)$$
$$\to (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{-(u^2/2)(1-\gamma)} \exp\left(-\frac{1}{2}\left(t^2 \int_0^{\gamma} \mathscr{V}ar(h(Zg(x), x))\, dx\right.\right.$$

$$+ 2ut \int_0^\gamma \mathscr{C}ov\big(h(Zg(x), x), Z\big)\, dx + u^2\gamma\Big)\Big)\, du$$

$$\to \exp\left(-\frac{t^2}{2}\left[\int_0^1 \mathscr{V}ar\big(h(Zg(x), x)\big)\, dx - \left(\int_0^1 \mathscr{C}ov\big(h(Zg(x), x), Z\big)\, dx\right)^2\right]\right) \quad (2.7)$$

as $\gamma \to 1$. Similarly, $\sum_{m+1}^{n} h(nD_k, \xi_k)$ can be studied. By arguments similar to those of Le Cam (1958, p. 13), we obtain

$$\mathscr{E}\left(\exp\left(itn^{-1/2}\sum_1^n h(nD_k, \xi_k)\right)\right)$$

$$\to \exp\left(-\frac{1}{2}t^2\left[\int_0^1 \mathscr{V}ar(h(Zg(x), x))\, dx - \left(\int_0^1 \mathscr{C}ov\big(h(Zg(x), x), Z\big)\, dx\right)^2\right]\right) \quad (2.8)$$

which proves the assertion using the continuity theorem on characteristic functions. Q.E.D.

LEMMA 2.1. *Let G satisfy condition* C1, *and* $\{W_k\}$ *denote partial sums of independently and identically distributed* exp(1) *r.v.'s* $\{Z_k\}$. *Then as* $n \to \infty$

$$n[G(W_k)/n) - G(W_{k-1}/n)] = g(\xi_k)\, Z_k + g'(\xi_k)Z_k\frac{W_{k-1} - (k-1)}{n} + o_p(n^{-1/2}), \quad (2.9)$$

where $o_p(\cdot)$ is uniform in k.

Proof. A Taylor expansion gives

$$G(W_k/n) = G(\xi_k) + (W_k/n - \xi_k)g(\xi_k) + \tfrac{1}{2}(W_k/n - \xi_k)^2 g'(\xi_k)$$
$$+ \tfrac{1}{6}\,(W_k/n - \xi_k)^3 g''\,(\xi_k + \theta_{kn}(W_k/n - \xi_k)), \quad (2.10)$$

where $0 < \theta_{kn} < 1$. As $\max_k n^{-1/2}\,|\,W_k - k\,| = O_p(1)$, the continuity of g'' gives

$$g''(\xi_k + \theta_{kn}(W_k/n - \xi_k)) = g''(\xi_k) + o_p(1), \quad (2.11)$$

where $o_p(1)$ is uniform in k. In a similar way we get

$$n\big[G(W_k/n) - G(W_{k-1}/n)\big]$$
$$= Z_k\, g(\xi_k) + n^{-1}Z_k(W_{k-1} - (k-1))g'(\xi_k) + o_p(n^{-1/2}). \quad (2.12)$$

Q.E.D.

Using Taylor expansions as above, we obtain from Lemma 2.1

LEMMA 2.2. *Under assumptions* C1 *and* C2,

$$n^{-1/2}\sum_{k=1}^{m} h(n\big[G(W_k/n) - G(W_{k-1}/n)\big], \xi_k) =$$

$$n^{-\frac{1}{2}}\sum_1^m\left\{h(Z_k\, g(\xi_k), \xi_k) + h_x(Z_k\, g(\xi_k), \xi_k)Z_k\, g'(\xi_k)\left(W_{k-1} - \frac{k-1}{n}\right)\right\} + o_p(1). \quad (2.13)$$

LEMMA 2.3. *Under assumptions* C1, C2 *and* (2.3),

$$n^{-1/2}\sum_1^m h(n\big[G(W_k/n) - G(W_{k-1}/n)\big], \xi_k) = n^{-1/2}\sum_1^m h(Z_k\, g(\xi_k), \xi_k) + o_p(1). \quad (2.14)$$

Proof. First observe that (2.3) implies

$$0 = \frac{\partial}{\partial x}\,\mathscr{E}(h(Zg(x), y)) = \mathscr{E}\left(\frac{\partial}{\partial x} h(Zg(x), y)\right)$$
$$= \mathscr{E}\big(Zg'(x)h_x(Zg(x), y)\big), \tag{2.15}$$

so that the r.v.

$$v_k(Z_k) = h_x(Z_k\, g(\xi_k), \xi_k)Z_k\, g'(\xi_k) \tag{2.16}$$

has $\mathscr{E}v_k(Z_k) = 0$ and moments of all orders. Also

$$\mathscr{E}\big(v_k(Z_k)[W_{k-1} - (k-1)]\big) = \mathscr{E}v_k(Z_k)\,\mathscr{E}\big(W_{k-1} - (k-1)\big) = 0,$$

and therefore,

$$n^{-3}\,\mathscr{V}\!ar\left(\sum_1^m v_k(Z_k)[W_{k-1} - (k-1)]\right) = n^{-3}\sum_1^m \mathscr{E}v_k^2(Z_k)\cdot(k-1) = O_p(n^{-1}).$$

The assertion now follows from (2.13). Q.E.D.

LEMMA 2.4. *Under conditions* C1, C2 *and* (2.3), *as* $n \to \infty$,

$$n^{-1/2}\sum_1^m h(n[G(W_k/n) - G(W_{k-1}/n)], \xi_k) \xrightarrow{D} \mathscr{N}\left(0, \int_0^\gamma \mathscr{V}\!ar(h(Zg(x), x))\,dx\right). \tag{2.17}$$

Proof. By Lemma 2.3, it is sufficient to consider the sum of independent r.v.'s

$$n^{-1/2}\sum_1^m h(Z_k\, g(\xi_k), \xi_k).$$

Since $h(x, y)$ is bounded by $c_1(x^{c_2} + 1)$, it is easy to verify the Liapunov condition, which establishes the asymptotic normality. As $\mathscr{V}\!ar(h(Zg(x), x))$ is continuous in x, $0 \le x \le 1$, it follows that

$$\mathscr{V}\!ar\left(n^{-1/2}\sum_1^m h(Z_k\, g(\xi_k), \xi_k)\right) \to \int_0^\gamma \mathscr{V}\!ar(h(Zg(x), x))\,dx. \qquad \text{Q.E.D.}$$

For the uniform distribution on [0, 1] we get

COROLLARY 2.1. *As* $n \to \infty$, *under condition* C2 *and* (2.3),

$$n^{-1/2}\sum_1^n h(nT_k, \xi_k) \xrightarrow{D} \mathscr{N}\left(0, \int_0^1 \mathscr{V}\!ar(h(Z, y))\,dy - \left(\int_0^1 \mathscr{C}\!ov(h(Z, y), Z)\,dy\right)^2\right).$$

Remark 2.1. The regularity conditions C1 and C2 on $G = F^{-1}$ and h are somewhat stringent and can be relaxed. But the relaxation calls for more complex proofs, as in Holst and Rao (1980), and is not very useful for statistical purposes.

Remark 2.2. The special case $h(x, y) = a(y)x$ leads to sums of the form $\sum_{k=1}^m a_{kn}D_{kn}$ which are linear combinations of order statistics from the distribution function F. Necessary and sufficient conditions for the asymptotic normality for the particular case of linear combinations of uniform order statistics are given in Hecker (1976). See also Weiss (1962).

3. CLOSE ALTERNATIVES AND LOCALLY OPTIMAL TESTS

Let $\{D_k\}$ be spacings from the distribution

$$F_n(y) = y + L_n(y)/n^{1/2}, \qquad 0 \le y \le 1, \tag{3.1}$$

where $L_n(0) = L_n(1) = 0$. As in Rao and Sethuraman (1975), assume that $L_n(y)$ is uniformly close to a function $L(y)$ which is twice continuously differentiable with derivatives $L'(y) = l(y)$ and $L''(y) = l'(y)$. Under this type of smoothness conditions $\{D_k\}$ can be related to the uniform spacings $\{T_k\}$ by the relation (cf. (3.8) of Rao and Sethuraman 1975)

$$nD_k = n\left[F_n^{-1}(U'_k) - F_n^{-1}(U'_{k-1})\right] = nT_k\left(1 - \frac{l(\xi_k)}{\sqrt{n}}\right) + o_p(n^{-1/2}), \tag{3.2}$$

where $o_p(\cdot)$ is uniform in k. By partial integration it follows from the assumption (2.3) that $\mathscr{E}(Zh_x(Z, y)) = \mathscr{C}ov(h(Z, y), Z)$. By Taylor expansion one gets after some calculation:

Lemma 3.1. *Under condition* C2, *the difference*

$$n^{-1/2} \sum_1^n h(nD_k, \xi_k) - n^{-1/2} \sum_1^n h(nT_k, \xi_k) \to -\int_0^1 l(y)\, \mathscr{C}ov(Z, h(Z, y))\, dy \tag{3.3}$$

in probability as $n \to \infty$.

We may combine Corollary 2.1 and the above lemma to formulate the main result of this section.

Theorem 3.1. *If the condition* C2 *and the assumptions on* $L_n(x)$ *hold, then under the close alternatives in* (3.1),

$$n^{-1/2} \sum_1^n h(nD_k, \xi_k) \xrightarrow{D} \mathscr{N}\left(-\int_0^1 l(y)\, \mathscr{C}ov(Z, h(Z, y))\, dy,\ \int_0^1 \mathscr{V}ar(h(Z,y))\, dy - \left(\int_0^1 \mathscr{C}ov(h(Z, y), Z)\, dy\right)^2\right). \tag{3.4}$$

To find the asymptotically locally most powerful test against the specific alternative (3.1), we need to find the function $h(x, y)$ which maximizes

$$e_h = \frac{\int_0^1 l(y)\, \mathscr{C}ov(Z, h(Z, y))\, dy}{\sqrt{\int_0^1 \mathscr{V}ar(h(Z, y))\, dy - \left(\int_0^1 \mathscr{C}ov(h(Z,y), Z)\, dy\right)^2}}. \tag{3.5}$$

As in Holst (1972), it follows that $h(x, y) = l(y) \cdot x$ maximizes e_h. The optimal spacings test is thus to reject $\mathscr{H}_0$ if

$$\sum_{k=1}^n l(\xi_k) D_k < c. \tag{3.6}$$

This statistic is linear in the spacings and hence is a weighted linear combination of the order statistics.

For illustration consider testing the null hypothesis that a random sample $(X_1, \ldots, X_{n-1})$ is from a logistic distribution with $F(t) = 1/(1 + e^{-t})$ against translation alternatives. After transforming the data to $[0, 1]$ through $x = F(t)$ one finds that

$L_n(x)$ of (3.1) converges to $L(x)$ with derivative $l(x) = 2x - 1$, $0 < x < 1$. The test simplifies to a test based on the sample mean. Analogously, asymptotically optimal spacings tests can be derived for various other problems including for close "scale" alternatives.

For the symmetric case, where $h(x, y) = h(x)$, it is seen from Theorem 3.1 that no power is obtained for alternatives at a "distance" of $n^{-1/2}$, i.e., of the type (3.1). Power for alternatives at a distance of $n^{-1/4}$ can be obtained through an analysis similar to that above. The so called Greenwood statistic $\sum_1^n D_i^2$ is optimal. See Rao and Sethuraman (1975) or Holst and Rao (1980).

4. THE TWO-SAMPLE PROBLEM

In this section the test statistic $Q_n = \sum_1^n h_{kn}(S_k)$ is studied under close alternatives. First note that given the U-observations the probability of a Y-value falling in the interval $[U'_{k-1}, U'_k)$ is $T_k = U'_k - U'_{k-1}$ if the second sample comes from the same distribution as the first. More generally this probability is $D_k = F(U'_k) - F(U'_{k-1})$ under the alternative $F(\cdot)$. Thus the conditional distribution

$$\mathscr{L}((S_1, \ldots, S_n) \mid U_1, \ldots, U_{n-1}) = \text{Mult}(m, D_1, \ldots, D_n) \tag{4.1}$$

is a multinomial distribution under the alternative $F(\cdot)$. Consider a sequence of alternatives given by (3.1) and satisfying the conditions of Section 3. We have

$$D_k = F_n(U'_k) - F_n(U'_{k-1}) = T_k + \frac{L_n\left(\sum_1^k T_j\right) - L_n\left(\sum_1^{k-1} T_j\right)}{n^{1/2}}, \tag{4.2}$$

where $\{T_k\}$ are the uniform spacings as in (1.2).

THEOREM 4.1. *Let $m,n \to \infty$ in such a way that $m/n \to \rho$, $0 < \rho < \infty$. Let η denote a geometric random variable with $P(\eta = j) = \rho^j/(1 + \rho)^{j+1}$. Under the assumptions on $L_n(x)$ in connection with* (3.1),

$$n^{-1/2} \sum_{k=1}^{n} h(\xi_k, S_k) \xrightarrow{D} \mathcal{N}(\mu, \sigma^2) \tag{4.3}$$

where

$$\mu = -\frac{1}{1+\rho} \int_0^1 l(u)\, \mathscr{C}ov(\eta, h(u, \eta))\, du \tag{4.4}$$

and

$$\sigma^2 = \int_0^1 \mathscr{V}ar(h(u, \eta))\, du - \frac{1}{\mathscr{V}ar(\eta)}\left(\int_0^1 \mathscr{C}ov(\eta, h(u, \eta))\, du\right)^2. \tag{4.5}$$

Proof. Using (4.1), it follows from Holst (1979a) that for any $N \leq n$,

$$\mathscr{E}\left(\exp\left(itn^{-1/2} \sum_1^N h_k(S_k)\right) \middle| \mathbf{D}\right) = \left[2\pi P\left(\sum_1^n Y_k = m \middle| \mathbf{D}\right)\right]^{-1}$$
$$\times \int_{-\pi}^{\pi} \mathscr{E}\left(\exp\left(itn^{-1/2} \sum_1^N h_k(Y_k) + iu \sum_1^n (Y_k - mD_k)\right) \middle| \mathbf{D}\right) du, \tag{4.6}$$

where, conditional on $\mathbf{D} = (D_1, \ldots, D_n)$, the $\{Y_k\}$'s are independent with $Y_k \sim \mathrm{Poi}(mD_k)$, and $h_k(\cdot) = h(\xi_k, \cdot)$.

Furthermore, since $\mathscr{L}(\sum_1^n Y_k \mid \mathbf{D}) = \mathrm{Poi}(m)$ for any $\mathbf{D}$, it follows that

$$P\left(\sum_1^n Y_k = m \,\middle|\, \mathbf{D}\right) = \frac{m^m e^{-m}}{m!} = (2\pi m)^{-1/2} \exp(o(1)) \tag{4.7}$$

by Stirling's formula. From (4.2) and the assumptions, it can be seen as in Lemma 3.1 that for any real numbers t and u,

$$\frac{\mathscr{E}\left(\exp\left\{i \sum_1^N \left[\dfrac{th_k(Y_k) - \mathscr{E}(h_k(Y_k) \mid D_k)}{n^{1/2}} + \dfrac{u(Y_k - mD_k)}{m^{1/2}}\right]\right\} \,\middle|\, \mathbf{D}\right)}{\mathscr{E}\left(\exp\left\{i \sum_1^N \left[\dfrac{th_k(Y'_k) - \mathscr{E}(h_k(Y'_k) \mid T_k)}{n^{1/2}} + \dfrac{u(Y'_k - mT_k)}{m^{1/2}}\right]\right\} \,\middle|\, \mathbf{T}\right)} \to 1 \tag{4.8}$$

in probability as $m, n \to \infty$. Here the Y'_k's are independent $\mathrm{Poi}(mT_k)$ random variables. One can also prove that when n, $N \to \infty$ in such a way that $N/n \to \gamma$, $0 < \gamma < 1$,

$$n^{-1/2} \sum_1^N \mathscr{E}(h_k(Y_k) \mid D_k) - n^{-1/2} \sum_1^N \mathscr{E}(h_k(Y'_k) \mid T_k) = -A(\gamma) + o_p(1) \tag{4.9}$$

where

$$A(\gamma) = \int_0^\gamma l(u)\, \mathscr{C}ov(Z, \mathscr{E}(h(u, \eta) \mid Z))\, du \tag{4.10}$$

and Z is an $\exp(1)$ random variable, $\mathscr{L}(\eta \mid Z)$ is $\mathrm{Poi}(\rho Z)$ so that $P(\eta = j) = \rho^j/(1 + \rho)^{j+1}$. Using conditional expectations, we get

$$\begin{aligned}
\mathscr{E}&\left\{\exp\left(itn^{-1/2} \sum_1^N h_k(S_k)\right)\right\} = \mathscr{E}\left(\mathscr{E}\left\{\exp\left(itn^{-1/2} \sum_1^N h_k(S_k)\right) \,\middle|\, \mathbf{D}\right\}\right) \\
&= (2\pi)^{-1/2} \exp(o(1)) \int_{-\pi\sqrt{m}}^{\pi\sqrt{m}} \mathscr{E}\left[\exp\left(it \sum_1^N \mathscr{E}(h_k(Y_k) \mid D_k)\right)\right. \\
&\times \mathscr{E}\left\{\exp\left(\sum_1^N i\,\{tn^{-1/2}[h_k(Y_k) - \mathscr{E}(h_k(Y_k) \mid D_k)] + um^{-1/2}(Y_k - mD_k)\}\right) \,\middle|\, \mathbf{D}\right\} \\
&\times \left.\mathscr{E}\left\{\exp\left(ium^{-1/2} \sum_{N+1}^n (Y_k - mD_k)\right) \,\middle|\, \mathbf{D}\right\}\right] du.
\end{aligned} \tag{4.10}$$

The integrand in (4.10) is dominated by

$$\begin{aligned}
f_\nu(u) &= \mathscr{E}\left|\, \mathscr{E}\left\{\exp\left(ium^{-1/2} \sum_{N+1}^n (Y_k - mD_k)\right) \,\middle|\, \mathbf{D}\right\}\right| \\
&= \mathscr{E}\left|\, \exp\left\{m\left[1 - \sum_1^N T_j - \frac{1}{n^{1/2}} L_n\left(\sum_1^N T_j\right)\right](e^{m^{-1/2}iu} - 1 - m^{-1/2}iu)\right\}\right| \\
&\to f(u) = e^{-(1-\gamma)u^2/2}
\end{aligned}$$

as $\nu \to \infty$. Also as $\nu \to \infty$

$$\int_{-\pi\sqrt{m}}^{\pi\sqrt{m}} f_\nu(u)\, du \to \int_{-\infty}^{\infty} f(u)\, du. \tag{4.12}$$

Thus by the extended Lebesgue dominated-convergence theorem (see for instance C.R. Rao 1973, p. 136), it follows by combining the results above that

$$\lim \mathscr{E}(\exp(itn^{-1/2}\sum_1^N h_k(S_k))$$

$$= \int_{-\infty}^{\infty} \exp(itA(\gamma)) \lim \mathscr{E}\left[\mathscr{E}\left\{\exp\left(itn^{-1/2}\sum_1^N h_k(Y'_k) + ium^{-1/2}\sum_1^N (Y'_k - mT_k)\right)\middle|\mathbf{T}\right\}\right.$$

$$\left.\times (2\pi)^{-1/2}\mathscr{E}\left\{exp\left(ium^{-1/2}\sum_{N+1}^{n}(Y'_k - mT_k\right)\middle|\mathbf{T}\right\}\right] du$$

$$= \exp(itA(\gamma)) \lim \mathscr{E}\left\{\exp\left(itn^{-1/2}\sum_1^N h_k(S'_k\right)\right\}, \tag{4.13}$$

where

$$\mathscr{L}(S'_1, \ldots, S'_n \mid \mathbf{T}) = \text{Mult}\,(m; T_1, \ldots, T_n), \tag{4.14}$$

with the unconditional distribution

$$P(\mathbf{S}' = \mathbf{s}') = \binom{n+m-1}{n}^{-1}. \tag{4.15}$$

From the results of Holst (1979a), the asymptotic behaviour of a random variable of the type $\sum_1^N h(S'_k)$ can be deduced. In an analogous way

$$n^{-1/2}\sum_1^N h_k(S'_k) \xrightarrow{D} \mathscr{N}\left(0, \int_0^\gamma \mathscr{V}ar(h(u,\eta))\, du - \frac{1}{\mathscr{V}ar(\eta)}\left(\int_0^\gamma \mathscr{C}ov(\eta, h(u,\eta))\, du\right)^2\right)$$

Similarly one can study $\sum_{N+1}^n h_k(S'_k)$. Using an argument of Le Cam (1958, p. 13), we obtain

$$n^{-1/2}\sum_1^n h_k(S_k) \xrightarrow{D} \mathscr{N}\left(-A(1), \int_0^1 \mathscr{V}ar(h(u,\eta))\, du - \frac{1}{\mathscr{V}ar(\eta)}\left(\int_0^1 \mathscr{C}ov(\eta, h(u,\eta))\, du\right)^2\right).$$

By an elementary calculation, one finds $\mathscr{C}ov(Z, \mathscr{E}(h(u,\eta)|\, Z)) = \mathscr{C}ov(\eta, h(u,\eta))/(\rho + 1)$. Q.E.D.

As in Section 3, $h(u, j) = h(u)j$ gives an asymptotically optimal test statistic.

These results are proved here under rather restrictive assumptions which can be considerably weakened. That, however, makes the proofs technically more involved without providing much further insight into the statistical problem. For a proof under weaker conditions, see Holst and Rao (1980). Once the optimality of tests linear in $\{S_k\}$ is established, it is possible to study conditions on $\{a_k\}$ under which $\sum_1^n a_kS_k$ is asymptotically normal. For instance, by Theorem 3 of Holst (1979a), we have

THEOREM 4.2 (Holst). *If* $\sum_1^n a_k = 0$, $\sum_1^n a_k^2/n \to 1$, *and* $\max_{1\le k\le n} a_k^2/n \to 0$, *then* $\mathscr{L}(n^{-1/2} \sum_1^n a_k S_k') \to \mathcal{N}(0, \mathcal{V}ar(\eta))$, *where* η *is a geometric random variable with* $P(\eta = j) = \rho^j/(1+\rho)^{j+1}$.

If $h(u, j) = h(j)$, then the test statistics will be symmetric in the S_k's. From Theorem 4.1, it is seen that such statistics will give no power against alternatives of the type (3.1), which converges to the null at a rate of $n^{-1/2}$. But by arguments similar to those used here, it is possible to establish that such tests have power against alternatives converging to the null at a slower rate of $n^{-1/4}$ and that among such symmetric classes of statistics, $\sum_1^n S_k^2$ suggested by Dixon (1940) is asymptotically optimal, irrespective of $L_n(u)$. See Holst and Rao (1980) for details. It may be remarked here that the "run test" is of this symmetric type.

To illustrate the above results consider the two-sample problem with location alternatives, i.e., we have m observations from the distribution function $G(x - n^{-1/2}\theta)$ and $n - 1$ observations from $G(x)$ on $\mathbb{R}^1$. If G is sufficiently smooth, then it follows that

$$l(u) = -\frac{\theta G''(G^{-1}(u))}{G'(G^{-1}(u))} = -\theta a(u), \tag{4.18}$$

say. The asymptotically optimal test statistic is $\sum_1^n a(\xi_k)S_k$. For example, the logistic distribution gives the Wilcoxon-Mann-Whitney test.

5. SUMMARY

In this paper the asymptotic theory of spacing statistics of the form $V_n = \sum_{k=1}^n h_{kn}(D_{kn})$ is considered, where $\{D_{kn}, 1 \le k \le n, n \ge 1\}$ are the sample spacings from any fixed distribution function $F(x)$ on [0, 1], and $\{h_{kn}(\cdot), 1 \le k \le n, n \ge 1\}$ are real measurable functions satisfying regularity conditions. Also such statistics are studied for "alternatives" close to the uniform on [0, 1], and locally optimal tests derived. Next, let $U_1, \ldots, U_{n-1}$ and $V_1, \ldots, V_m$ be independent random samples from two continuous distribution functions. The problem considered is to test the null hypothesis that these two parent populations are identical. Let $U_1' \le \cdots \le U_{n-1}'$ be the ordered U-observations. Denote by S_k the number of V-observations falling in the interval $[U_{k-1}', U_k')$. For $\{h_{kn}(\cdot)\}$ given functions, asymptotic theory under the null and close alternatives is studied for test statistics of the form $\sum_{k=1}^n h_{kn}(S_k)$, using the results obtained for spacings.

RÉSUMÉ

On étudie la distribution asymptotique des tests basés sur l'espacement des données. Une théorie est également développée pour la distribution sous des alternatives voisines de l'hypothèse nulle; cette théorie est utilisée pour trouver, parmi les tests basés sur l'espacement des données, celui qui est localement le plus puissant. Dans le cas de deux échantillons indépendants, où l'hypothèse à tester est que les échantillons sont issus d'une même population, le test est basé sur la distribution des données de l'un des échantillons dans les intervalles entre les données de l'autre. Une théorie générale de la distribution asymptotique de telles statistiques est étudiée sous l'hypothèse nulle ainsi que sous une suite d'alternatives voisines.

REFERENCES

Blumenthal, S. (1963). The asymptotic normality of two test statistics associated with the two-sample problem. *Ann. Math Statist.*, 34, 1513–1523.

Blumenthal, S. (1967). Limit theorems for functions of shortest two-sample spacings and a related test. *Ann. Math. Statist.*, 38, 108–116.

Dixon, W.J. (1940). A criterion for testing the hypothesis that two samples are from the same population. *Ann. Math. Statist.*, 32, 199–204.

Godambe, V.P. (1961). On the two-sample problem: A heuristic method for constructing tests. *Ann. Math. Statist.*, 32, 1091–1107.

Hecker, H. (1976). A characterization of the asymptotic normality of linear combinations of order statistics from the uniform distribution. *Ann. Statist.*, 4, 1244–1246.

Holst, L., (1972). Asymptotic normality and efficiency for certain goodness of fit tests. *Biometrika*, 59, 137–145.

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

Holst, L. (1979b). Asymptotic normality of sum-functions of spacings, *Ann. Probab.*, 7, 1066–1072.

Holst, L., and Rao, J.S. (1980). Asymptotic theory for some families of two-sample nonparametric statistics. *Sankhyā Ser. A*, 42, 1–28.

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

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* Second Edition. Wiley, New York.

Rao, J.S. (1976). Some tests based on arc-lengths for the circle. *Sankhyā Ser. B*, 38, 329–338.

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.

Weiss, L. (1962). On the distribution of linear functions of spacings from a uniform distribution. *Math. Scand.*, 11, 149–150.

Received 23 July 1979
Revised 1 July 1980

Department of Mathematics
University of California
Santa Barbara, California 93106, U.S.A.