

F. CZEKAŁA (Wrocław)

NORMALIZING CONSTANTS FOR A STATISTIC
BASED ON LOGARITHMS OF DISJOINT m -SPACINGS

Abstract. The paper is concerned with the asymptotic normality of a certain statistic based on the logarithms of disjoint m -spacings. The exact and asymptotic mean and variance are computed in the case of uniform distribution on the interval $[0, 1]$. This result is generalized to the case when the sample is drawn from a distribution with positive step density on $[0, 1]$.

1. Introduction. Let X_1, \dots, X_N be a sample from the uniform distribution on the interval $[0, 1]$. Let us denote by $X_{(1)}, \dots, X_{(N)}$ the order statistics derived from this sample and define $X_{(0)} = 0$, $X_{(N+1)} = 1$. Let Y_0, Y_1, \dots, Y_N be i.i.d. exponential random variables with unit mean. It is known that

$$(1) \quad X_{(i)} \stackrel{d}{=} \exp\left(-\frac{Y_i}{i} - \dots - \frac{Y_N}{N}\right), \quad i = 1, \dots, N.$$

We will use the notation

$$(2) \quad X_{k:l} = \exp\left(-\sum_{i=k}^l \frac{Y_i}{i}\right), \quad 1 \leq k \leq l \leq N.$$

We also have

$$(3) \quad X_{(i)} \stackrel{d}{=} \frac{Y_1 + \dots + Y_i}{Y_0 + \dots + Y_N}.$$

We define the m -spacings from the sample X_1, \dots, X_N as the differences $D_{n,N}^{(m)} = X_{(n+m)} - X_{(n)}$, $n = 0, \dots, N+1-m$, $m \geq 1$. Let us introduce the

1991 *Mathematics Subject Classification*: 62E15, 62E20, 62G10, 62G30.

Key words and phrases: higher-order spacings; asymptotic normality; step density function.

Research supported by KBN (Committee for Scientific Research, Poland) Grant 2 1033 91 01.

following notation:

$$(4) \quad G_n^{(m)} = D_{n,N}^{(m)} \log D_{n,N}^{(m)}.$$

We are interested in the statistic

$$(5) \quad G_N = \sum_{0 \leq im \leq N+1-m} G_{im}^{(m)}$$

based on disjoint m -spacings. It can be proved as in Cressie (1976) or deduced from Del Pino (1979) that G_N is asymptotically normal. The general results of Del Pino (1979) give the asymptotic means and variances of statistics based on disjoint m -spacings in the form of means and variances of some other random variables. In any particular case these quantities have to be computed separately to obtain the values of normalizing constants. This method was used by Jammalamadaka & Tiwari (1986). In this paper we will find the exact values of $E(G_N)$ and $\text{Var}(G_N)$ in a closed and simple form. We will also establish the asymptotic normality of G_N for alternatives with step densities.

2. Main results

THEOREM 1. *Let $k \geq 1$ and $N+1 = km$. Then the random variable G_N has the mean and variance given by*

$$(6) \quad E(G_N) = - \sum_{i=m+1}^{N+1} \frac{1}{i},$$

$$(7) \quad \text{Var}(G_N) = \frac{1}{N+2} [F(m+1) - F(N+2)],$$

where

$$(8) \quad F(n) = n \sum_{i=n}^{\infty} \frac{1}{i^2}.$$

The method of the proof will be similar to that of Cressie (1976). To prove the main result we will need some lemmas.

LEMMA 1. *For $\alpha > -1$ and $n \geq 1$ we have*

$$(a) \quad E(X_{n:N}^\alpha) = \prod_{i=n}^N \frac{i}{i+\alpha},$$

$$(b) \quad E(X_{n:N}^\alpha \log X_{n:N}) = - \prod_{i=n}^N \frac{i}{i+\alpha} \sum_{j=n}^N \frac{1}{j+\alpha},$$

$$(c) \quad E(X_{n:N}^\alpha \log^2 X_{n:N}) = \prod_{i=n}^N \frac{i}{i+\alpha} \left[\sum_{j=n}^N \frac{1}{(j+\alpha)^2} + \left(\sum_{j=n}^N \frac{1}{j+\alpha} \right)^2 \right].$$

Proof. The proof is straightforward using the definition of $X_{n:N}$ and integrating with respect to the density of the exponential distribution.

LEMMA 2. For $n, m \geq 1$ we have

$$E(\log(1 - X_{n:n+m-1})) = - \sum_{k=1}^{\infty} \frac{1}{k} \prod_{i=n}^{n+m-1} \frac{i}{i+k} = - \sum_{i=m}^{n+m-1} \frac{1}{i}.$$

Proof. Expanding the logarithm into a power series we get

$$\begin{aligned} E(\log(1 - X_{n:n+m-1})) &= E\left(- \sum_{k=1}^{\infty} \frac{1}{k} X_{n:n+m-1}^k\right) \\ &= - \sum_{k=1}^{\infty} \frac{1}{k} E(X_{n:n+m-1}^k) = - \sum_{k=1}^{\infty} \frac{1}{k} \prod_{i=n}^{n+m-1} \frac{i}{i+k}. \end{aligned}$$

On the other hand, $X_{n:n+m-1}$ has the same distribution as the n th order statistic from a sample of size $n + m - 1$. Applying the transformation $x \rightarrow 1 - x$ we get $1 - X_{n:n+m-1} \stackrel{d}{=} X_{m:n+m-1}$ and hence

$$E(\log(1 - X_{n:n+m-1})) = E(\log X_{m:n+m-1}) = - \sum_{i=m}^{n+m-1} \frac{1}{i}.$$

LEMMA 3. For $n, m \geq 1$ we have

$$\sum_{k=1}^{\infty} \prod_{i=n}^{n+m} \frac{i}{i+k} = \frac{n}{m}.$$

Proof. Using Lemma 1 and the transformation $x \rightarrow 1 - x$ as in the proof of Lemma 2, we get

$$\begin{aligned} \sum_{k=1}^{\infty} \prod_{i=n}^{n+m} \frac{i}{i+k} &= \sum_{k=1}^{\infty} E(X_{n:n+m}^k) = E\left(\sum_{k=1}^{\infty} X_{n:n+m}^k\right) = E\left(\frac{X_{n:n+m}}{1 - X_{n:n+m}}\right) \\ &= E\left(\frac{1 - X_{m+1:n+m}}{X_{m+1:n+m}}\right) = E\left(\frac{1}{X_{m+1:n+m}}\right) - 1 \\ &= E\left(\exp\left(\sum_{i=m+1}^{n+m} \frac{Y_i}{i}\right)\right) - 1 = \prod_{i=m+1}^{n+m} \frac{i}{i-1} - 1 = \frac{n}{m}. \end{aligned}$$

LEMMA 4. For $n, m \geq 1$ we have

$$\sum_{k=1}^{\infty} \prod_{i=n}^{n+m-1} \frac{i}{i+k} \sum_{i=n}^{n+m-1} \frac{1}{i(i+k)} = \sum_{k=n+m}^{\infty} \frac{1}{k^2}.$$

Proof. This can be proved by induction. See Cressie (1976).

Proof of Theorem 1. We calculate $E(G_N)$ first. It follows from (3) that all m -spacings have the same distribution. Hence from Lemma 1(b) we

get

$$\begin{aligned} E(G_n^{(m)}) &= E(G_0^{(m)}) = E(X_{(m)} \log X_{(m)}) \\ &= E(X_{m:N} \log X_{m:N}) = -\frac{m}{N+1} \sum_{i=m}^N \frac{1}{i+1}. \end{aligned}$$

Thus

$$\begin{aligned} E(G_N) &= E\left(\sum_{i=0}^{k-1} G_{im}^{(m)}\right) = kE(G_0^{(m)}) \\ &= \frac{N+1}{m} \cdot \frac{-m}{N+1} \sum_{i=m}^N \frac{1}{i+1} = -\sum_{i=m}^N \frac{1}{i+1}. \end{aligned}$$

The variance of G_N is more complicated to obtain. We have

$$\text{Var}(G_N) = k \text{Var}(G_0^{(m)}) + k(k-1) \text{Cov}(G_0^{(m)}, G_m^{(m)}).$$

From Lemma 1 we get

$$\begin{aligned} \text{Var}(G_0^{(m)}) &= E(X_{m:N}^2 \log^2 X_{m:N}) - E^2(X_{m:N} \log X_{m:N}) \\ &= \frac{m(m+1)}{(N+1)(N+2)} \left[\sum_{i=m+2}^{N+2} \frac{1}{i^2} + \left(\sum_{i=m+2}^{N+2} \frac{1}{i} \right)^2 \right] \\ &\quad - \frac{m^2}{(N+1)^2} \left(\sum_{i=m+1}^{N+1} \frac{1}{i} \right)^2. \end{aligned}$$

To calculate $\text{Cov}(G_0^{(m)}, G_n^{(m)})$, $n \geq m$, let us note that

$$\begin{aligned} \log(X_{n+m:N} - X_{n:N}) &= \log X_{n+m:N} + \log \left(1 - \frac{X_{n:N}}{X_{n+m:N}} \right) \\ &= \log X_{n+m:N} + \log(1 - X_{n:n+m-1}). \end{aligned}$$

Hence

$$\begin{aligned} E(G_0^{(m)} G_n^{(m)}) &= E(X_{m:N} \log X_{m:N} (X_{n+m:N} - X_{n:N}) \log X_{n+m:N}) \\ &\quad + E(X_{m:N} \log X_{m:N} (X_{n+m:N} - X_{n:N}) \log(1 - X_{n:n+m-1})). \end{aligned}$$

The expected values above can be calculated using the identities

$$X_{n:N} = X_{n:n+m-1} X_{n+m:N}, \quad X_{m:N} = X_{m:n-1} X_{n:N}$$

and the fact that $X_{m:n-1}$, $X_{n:n+m-1}$, $X_{n+m:N}$ are independent. After elementary but lengthy calculations we get

$$\begin{aligned} (9) \quad E(X_{m:N} \log X_{m:N} (X_{n+m:N} - X_{n:N}) \log X_{n+m:N}) \\ = \frac{m^2}{(N+1)(N+2)} \left(\sum_{i=m+n+2}^{N+2} \frac{1}{i} \sum_{i=m+1}^{N+2} \frac{1}{i} + \sum_{i=m+n+2}^{N+2} \frac{1}{i^2} \right). \end{aligned}$$

Similarly, using the expansion $\log(1 - X_{n:m+n-1}) = -\sum_{k=1}^{\infty} k^{-1} X_{n:m+n-1}^k$ and Lemmas 1-4 we obtain

$$(10) \quad E(X_{m:N} \log X_{m:N} (X_{n+m:N} - X_{n:N}) \log(1 - X_{n:n+m-1})) \\ = \frac{m^2}{(N+1)(N+2)} \left(\sum_{i=m+1}^{m+n+1} \frac{1}{i} \sum_{i=m+1}^{N+2} \frac{1}{i} - \sum_{i=m+n+2}^{\infty} \frac{1}{i^2} \right).$$

Combining (9) and (10) we have for $n \geq m$,

$$E(G_0^{(m)} G_n^{(m)}) = \frac{m^2}{(N+1)(N+2)} \left[\left(\sum_{i=m+1}^{N+2} \frac{1}{i} \right)^2 - \sum_{i=N+3}^{\infty} \frac{1}{i^2} \right]$$

and

$$\text{Cov}(G_0^{(m)}, G_n^{(m)}) = E(G_0^{(m)} G_n^{(m)}) - E(G_0^{(m)}) E(G_n^{(m)}) \\ = \frac{m^2}{N+1} \left[\frac{1}{N+2} \left(\sum_{i=m+1}^{N+2} \frac{1}{i} \right)^2 - \frac{1}{N+1} \left(\sum_{i=m+1}^{N+1} \frac{1}{i} \right)^2 - \frac{1}{N+2} \sum_{i=N+3}^{\infty} \frac{1}{i^2} \right].$$

Now we can calculate $\text{Var}(G_N)$:

$$\text{Var}(G_N) = k \text{Var}(G_0^{(m)}) + k(k-1) \text{Cov}(G_0^{(m)}, G_m^{(m)}) \\ = \frac{m+1}{N+2} \left[\sum_{i=m+2}^{N+2} \frac{1}{i^2} + \left(\sum_{i=m+2}^{N+2} \frac{1}{i} \right)^2 \right] - \frac{m}{N+1} \left(\sum_{i=m+1}^{N+1} \frac{1}{i} \right)^2 \\ + (k-1)m \left[\frac{1}{N+2} \left(\sum_{i=m+1}^{N+2} \frac{1}{i} \right)^2 - \frac{1}{N+1} \left(\sum_{i=m+1}^{N+1} \frac{1}{i} \right)^2 - \frac{1}{N+2} \sum_{i=N+3}^{\infty} \frac{1}{i^2} \right] \\ = \frac{m+1}{N+2} \sum_{i=m+2}^{N+2} \frac{1}{i^2} - \frac{(k-1)m}{N+2} \sum_{i=N+3}^{\infty} \frac{1}{i^2} + \left(\sum_{i=m+1}^{N+2} \frac{1}{i} \right)^2 \\ - \left(\sum_{i=m+1}^{N+1} \frac{1}{i} \right)^2 + \frac{m+1}{N+2} \left[\left(\sum_{i=m+2}^{N+2} \frac{1}{i} \right)^2 - \left(\sum_{i=m+1}^{N+2} \frac{1}{i} \right)^2 \right] \\ = \frac{1}{N+2} \left[(m+1) \sum_{i=m+1}^{\infty} \frac{1}{i^2} - (N+2) \sum_{i=N+2}^{\infty} \frac{1}{i^2} \right].$$

In the next lemma we give the expected value and the variance of G_N in the case when $N+1$ is not divisible by m .

LEMMA 5. Assume that $N + 1$ is not divisible by m and let $n < N$ be the largest integer such that $n + 1$ is divisible by m , i.e. $n + 1 = km$. Then

$$(11) \quad E(G_N) = -\frac{n+1}{N+1} \sum_{i=m+1}^{N+1} \frac{1}{i},$$

$$(12) \quad \text{Var}(G_N) = \frac{(n+1)(n+2)}{(N+1)(N+2)} \\ \times \left[\sum_{i=n+3}^{N+2} \frac{1}{i^2} + \left(\sum_{i=n+3}^{N+2} \frac{1}{i} + \sum_{i=m+1}^{n+1} \frac{1}{i} \right)^2 + \text{Var}(G_n) \right] \\ - \frac{(n+1)^2}{(N+1)^2} \left(\sum_{i=m+1}^{N+1} \frac{1}{i} \right)^2$$

where $\text{Var}(G_n)$ is given by (7).

Proof. We have $G_N = \sum_{i=0}^{k-1} D_{im,N}^{(m)} \log D_{im,N}^{(m)}$ and

$$D_{im,N}^{(m)} = (X_{im+m:N} - X_{im:N}) \\ = X_{n+1:N}(X_{im+m:n} - X_{im:n}) = X_{n+1:N} D_{im,n}^{(m)}.$$

In the equation above $X_{n+1:N}$ is independent of $D_{im,n}^{(m)}$, thus

$$G_N = \sum_{i=0}^{k-1} X_{n+1:N} D_{im,n}^{(m)} \log (X_{n+1:N} D_{im,n}^{(m)}) \\ = \sum_{i=0}^{k-1} X_{n+1:N} \log X_{n+1:N} D_{im,n}^{(m)} + \sum_{i=0}^{k-1} X_{n+1:N} D_{im,n}^{(m)} \log D_{im,n}^{(m)} \\ = X_{n+1:N} \log X_{n+1:N} + X_{n+1:N} G_n$$

and $X_{n+1:N}$ is independent of G_n . Using the last identity and applying Lemma 1 we get easily the statement of Lemma 5.

THEOREM 2. $E(G_N)$ and $\text{Var}(G_N)$ given by (6), (11) and (7), (12) respectively are asymptotically equivalent to

$$(13) \quad e_N = \sum_{i=1}^m \frac{1}{i} - \log(N+2) - \gamma$$

and

$$(14) \quad \sigma_N^2 = \frac{1}{N+2} [F(m+1) - 1],$$

where $\gamma = 0.577\dots$ is Euler's constant.

Proof. We want to show that $\text{Var}(G_N)/\sigma_N^2 \rightarrow 1$ and $(E(G_N) - e_N)/\sigma_N \rightarrow 0$. Since $F(n) \rightarrow 1$ it is easy to show that (7) and (12) are equivalent to σ_N^2 . It remains to prove that $\sqrt{N+2}(E(G_N) - e_N) \rightarrow 0$. We can replace (11) by (6) because

$$\sqrt{N+2} \left| -\frac{n+1}{N+1} \sum_{i=m+1}^{N+1} \frac{1}{i} + \sum_{i=m+1}^{N+1} \frac{1}{i} \right| \leq \frac{m\sqrt{N+2}}{N+1} \sum_{i=m+1}^{N+1} \frac{1}{i} \rightarrow 0.$$

Let us define $\gamma_N = \sum_{i=1}^N i^{-1} - \log(N+1)$. Then

$$\gamma_N = \sum_{i=1}^N \left[\frac{1}{i} - \log\left(\frac{i+1}{i}\right) \right]$$

and $\gamma_N \rightarrow \gamma$. Applying the inequalities

$$0 < \frac{1}{i} - \log\left(\frac{i+1}{i}\right) < \frac{1}{i} - \frac{1}{i+1} < \frac{1}{i^2},$$

we get

$$\begin{aligned} & \sqrt{N+2} \left(-\sum_{i=m+1}^{N+1} \frac{1}{i} - e_N \right) \\ &= \sqrt{N+2} (\gamma - \gamma_{N+1}) = \sqrt{N+2} \sum_{i=N+2}^{\infty} \left(\frac{1}{i} - \log\left(\frac{i+1}{i}\right) \right) \\ &\leq \sqrt{N+2} \sum_{i=N+2}^{\infty} \frac{1}{i^2} \leq \sum_{i=N+2}^{\infty} \frac{1}{i^{3/2}} \rightarrow 0. \end{aligned}$$

It is easy to see that the expressions for e_N and σ_N^2 given by (13) and (14) agree with the results of Gebert and Kale (1969) for $m = 1$ and Jammalamadaka & Tiwari (1986) for $m \geq 1$.

3. The case of step densities. In this section we will find the asymptotic distribution of G_n in the case when the underlying distribution has a positive step density on $[0, 1]$. For $m = 1$ it was found by Czekala (1993).

Let $\underline{X} = (X_1, X_2, \dots, X_n, \dots)$ be a sequence of random variables which are used to form G_n . G_n is a function of n and \underline{X} which we can write as $G_n = \phi(n, \underline{X})$. To simplify the notation we assume that $\phi(0, \underline{X}) = 0$. Now let $k \geq 1$ be a fixed integer and let $0 = x_0 < x_1 < \dots < x_k = 1$ be fixed real numbers. We can define subintervals of $[0, 1]$ as follows:

$$I_1 = [x_0, x_1), \dots, I_{k-1} = [x_{k-2}, x_{k-1}), \quad I_k = [x_{k-1}, x_k].$$

The lengths of the intervals I_i will be denoted by d_i . Let $f_i > 0, i = 1, \dots, k$, be fixed numbers such that $\sum_{i=1}^k f_i d_i = 1$. These numbers together with

the intervals I_i define a step density f :

$$(15) \quad f(x) = \sum_{i=1}^k f_i \mathbf{1}_{I_i}(x)$$

Define $p_i = f_i d_i$. We have $\sum p_i = 1$ and hence there exist numbers $0 = x'_0 < x'_1 < \dots < x'_k = 1$ such that the intervals I'_i , defined similarly to I_i , have lengths p_i . There also exists a vector of random elements $(\underline{U}, \underline{Y}^1, \dots, \underline{Y}^k)$ such that:

- (a) the coordinates of this vector are stochastically independent,
- (b) $\underline{U} = (U_1, U_2, \dots)$ is a sequence of independent random variables uniformly distributed on $[0, 1]$,
- (c) $\underline{Y}^i = (Y_1^i, Y_2^i, \dots)$, for $i = 1, \dots, k$, are sequences of independent random variables with uniform distribution on I_i .

We can now define a sequence $\underline{Z} = (Z_1, Z_2, \dots)$ of independent random variables with density f :

$$(16) \quad Z_n = \sum_{i=1}^k \mathbf{1}_{I'_i}(U_n) Y_n^i, \quad n \geq 1.$$

Let us denote by $N_{i,n}$ the number of random variables Z_1, \dots, Z_n taking values in the interval I_i , that is,

$$(17) \quad N_{i,n} = \sum_{j=1}^n \mathbf{1}_{I_i}(Z_j).$$

It is easy to see that $N_{i,n} = \sum_{j=1}^n \mathbf{1}_{I'_i}(U_j)$ and thus the sequence of vectors $(N_{1,n}, \dots, N_{k,n})$ is independent of $(\underline{Y}^1, \dots, \underline{Y}^k)$. Using this we will prove in the next two lemmas that $\phi(n, \underline{Z})$ is asymptotically identically distributed as

$$(18) \quad \sum_{i=1}^k \phi(N_{i,n}, \underline{Y}^i).$$

In the sequel we will use the notation $\psi(x) = x \log x$.

LEMMA 6. *If τ is an arbitrary random variable taking values in the set $\{1, \dots, m\}$ then*

$$\sqrt{n} \psi(X_{(\tau)}) \xrightarrow{P} 0,$$

where $X_{(\tau)}$ denotes the τ -th order statistic from the sample of size n .

Proof. We have

$$P(\sqrt{n} |\psi(X_{(\tau)})| \geq \varepsilon) \leq \sum_{i=1}^m P(\sqrt{n} |\psi(X_{(i)})| \geq \varepsilon),$$

hence it is enough to prove that $\sqrt{n} \psi(X_{(i)}) \xrightarrow{P} 0$ for each fixed $i \geq 1$. It is known that $nX_{(i)}$ converges weakly to some random variable X (with Erlang's distribution). Since ψ is a continuous function we also have $\psi(nX_{(i)}) \xrightarrow{d} \psi(X)$. It follows that

$$\sqrt{n} \psi(X_{(i)}) = \frac{\psi(nX_{(i)})}{\sqrt{n}} - \frac{\log n}{\sqrt{n}}(nX_{(i)}) \xrightarrow{P} 0.$$

LEMMA 7. *The statistic $\phi(n, \underline{Z})$ has the same asymptotic distribution as the statistic (18).*

Proof. Let us consider the sequence $Z_{\alpha_{i,1}}, Z_{\alpha_{i,2}}, \dots$ of those successive random variables Z_n whose values belong to the interval I_i . The sequence $\alpha_{i,n}$, $n \geq 1$, is determined by the sequence \underline{U} , so it is independent of \underline{Y}^i . Since by (16), $Z_{\alpha_{i,n}} = Y_{\alpha_{i,n}}^i$, the sequence $Z_{\alpha_{i,n}}$ has the same probability distribution as \underline{Y}^i . It can be shown similarly that the joint distribution of the vector $(Z_{\alpha_{1,n}}, n \geq 1, \dots, Z_{\alpha_{k,n}}, n \geq 1)$ is the same as that of $(\underline{Y}^1, \dots, \underline{Y}^k)$. It follows that the sequence $X_{(1)}, \dots, X_{(n)}$ has the same probability distribution as $Y_{(1)}^1, \dots, Y_{(N_{1,n})}^1, \dots, Y_{(1)}^k, \dots, Y_{(N_{k,n})}^k$. To prove our lemma we need to exclude the m -spacings that span different intervals I_i . We have $N_{i,n} \xrightarrow{P,1} \infty$ so we can assume that each of the intervals I_i contains at least one full m -spacing. Then the m -spacings that are fully contained in I_i are built from the random variables

$$(19) \quad Y_{\tau_i}^i, Y_{\tau_i+1}^i, \dots, Y_{\tau_i+\eta_i}^i,$$

where $\tau_i = m + m[S_{i-1}/m] - S_{i-1}$, $\eta_i = m[S_i/m] - m[S_{i-1}/m] - m$ and $S_i = \sum_{j=1}^i N_{j,n}$. We have omitted the index n in the definition of τ_i, η_i, S_i to simplify the notation. Let $D = Y_{\tau_i+1}^i - Y_{\tau_i+\eta_i}^i$ be the m -spacing containing x_i which is the common endpoint of I_i and I_{i+1} . We can write $D = D_1 + D_2 = (x_i - Y_{\tau_i+\eta_i}^i) + (Y_{\tau_i+1}^i - x_i)$. We have $\sqrt{n} |\psi(D)| \leq \sqrt{n} |\psi(D_1)| + \sqrt{n} |\psi(D_2)|$ and from Lemma 6 we obtain $\psi(D) \xrightarrow{P} 0$. We have proven that the m -spacings spanning different intervals I_i are negligible.

Now τ_i and η_i are independent of \underline{Y}^i so the distribution of (19) will not change if we replace (19) by

$$(20) \quad Y_{(0)}^i, Y_{(1)}^i, \dots, Y_{(\eta_i)}^i.$$

It follows that $\phi(n, \underline{Z})$ has the same asymptotic distribution as $\sum_{i=1}^k \phi(\eta_i, \underline{Y}^i)$. It is easy to see that the latter statistic is asymptotically identically distributed as (18) because $0 < N_{i,n} - \eta_i < 2m - 1$ and $N_{i,n} \xrightarrow{P,1} \infty$.

We will need some lemmas to find the asymptotic distribution of (18).

LEMMA 8. Let X and $X_n, n \geq 1$, be random variables such that $X_n \xrightarrow{P,1} 0$ and $\sqrt{n} X_n \xrightarrow{d} X$. Then

$$(21) \quad \sqrt{n} |\log(1 + X_n) - X_n| \xrightarrow{P} 0.$$

Proof. (21) follows easily from the fact that for $|x| \leq 1/2$ we have

$$|\log(1 + x) - x| \leq x^2.$$

LEMMA 9. Let $X_{i,n}, i = 1, \dots, k, n \geq 0$, be the random variables defined as follows:

$$X_{i,n} = \frac{1}{\sigma_n} (\phi(n, \underline{Y}^i) - d_i (\log d_i + e_n)),$$

where e_n and σ_n are defined by (13) and (14). Then

$$(22) \quad \left(X_{1,N_{1,n}}, \dots, X_{k,N_{k,n}}, \frac{N_{1,n} - np_1}{\sqrt{n}}, \dots, \frac{N_{k,n} - np_k}{\sqrt{n}} \right) \xrightarrow{d} (d_1 X_1, \dots, d_k X_k, W_1, \dots, W_k),$$

where the X_i are independent and normally $N(0, 1)$ distributed random variables, the vector (W_1, \dots, W_k) is independent of (X_1, \dots, X_k) and has the multivariate normal distribution $N(0, \Sigma)$, where $\Sigma = [\sigma_{i,j}]$ and

$$\sigma_{i,j} = \begin{cases} -p_i p_j & \text{for } i \neq j, \\ p_i - p_i^2 & \text{for } i = j, \end{cases}$$

$i, j = 1, \dots, k$.

Proof. We will first show that

$$(23) \quad (X_{1,n}, \dots, X_{k,n}) \xrightarrow{d} (d_1 X_1, \dots, d_k X_k).$$

We have $\sigma_n^{-1} (\phi(n, \underline{Y}^i / d_i) - e_n) \xrightarrow{d} X_i$, where e_n and σ_n are given by (13) and (14) respectively. Transforming $\phi(n, \underline{Y}^i / d_i)$ into $\phi(n, \underline{Y}^i)$ we get $X_{i,n} \xrightarrow{d} d_i X_i$. As $X_{i,n}$ are independent we also have (23). Now we show that

$$(24) \quad \left(\frac{N_{1,n} - np_1}{\sqrt{n}}, \dots, \frac{N_{k,n} - np_k}{\sqrt{n}} \right) \xrightarrow{d} (W_1, \dots, W_k).$$

This follows from the central limit theorem because $N_{i,n} = \sum_{j=1}^n \mathbf{1}_{I'_i}(U_j)$, $E(\mathbf{1}_{I'_i}(U_1)) = p_i$ and $\text{Cov}(\mathbf{1}_{I'_i}(U_1), \mathbf{1}_{I'_j}(U_1)) = \sigma_{i,j}$. Since the sequence $N_{i,n}$ is independent of \underline{Y}^i , (23) and (24) together give (22).

LEMMA 10. We have

$$\left(\sqrt{n} \log \left(\frac{N_{1,n}}{np_1} \right), \dots, \sqrt{n} \log \left(\frac{N_{k,n}}{np_k} \right) \right) \xrightarrow{d} \left(\frac{1}{p_1} W_1, \dots, \frac{1}{p_k} W_k \right).$$

Proof. Let $\alpha(x) = \log(1+x) - x$. Then

$$\sqrt{n} \log \left(\frac{N_{i,n}}{np_i} \right) = \frac{N_{i,n} - np_i}{\sqrt{np_i}} + \sqrt{n} \alpha \left(\frac{N_{i,n} - np_i}{np_i} \right).$$

From Lemma 8 it follows that $\sqrt{n} \alpha((N_{i,n} - np_i)(np_i)^{-1}) \xrightarrow{P} 0$. The rest of the proof results from Lemma 9.

Now we can compute the asymptotic distribution of the statistic G_n when the underlying distribution has density f given by (15).

THEOREM 3. *If the $X_1, X_2, \dots, X_n, \dots$ are i.i.d. random variables with the density f given by (15) then*

$$(25) \quad \frac{G_n - e_{f,n}}{\sigma_{f,n}} \xrightarrow{d} N(0, 1),$$

where

$$(26) \quad e_{f,n} = E \left(\frac{1}{f} \log \left(\frac{1}{f} \right) \right) - \sum_{m+1}^{n+1} \frac{1}{i},$$

$$(27) \quad \sigma_{f,n}^2 = \frac{1}{n+2} \left[F(m+1) E \left(\frac{1}{f^2} \right) - 1 \right].$$

An asymptotically equivalent form of $e_{f,n}$ is

$$(28) \quad e_{f,n} = E \left(\frac{1}{f} \log \left(\frac{1}{f} \right) \right) - \log n + \sum_{i=1}^m \frac{1}{i} - \gamma.$$

Proof. As proved in Lemma 7 we can replace G_n by $\sum \phi(N_{i,n}, \underline{Y}^i)$. Set $C_m = F(m+1) - 1$. From Lemmas 9 and 10 we get

$$\begin{aligned} \frac{1}{\sigma_n} \left[\sum_{i=1}^k \phi(N_{i,n}, \underline{Y}^i) - \sum_{i=1}^k d_i \log \left(\frac{d_i}{np_i} \right) - \left(\sum_{i=1}^m \frac{1}{i} - \gamma \right) \right] \\ \xrightarrow{d} \sum_{i=1}^k \frac{d_i}{\sqrt{p_i}} X_i - \frac{1}{\sqrt{C_m}} \sum_{i=1}^k \frac{d_i}{p_i} W_i, \end{aligned}$$

where W_i and X_i are defined in Lemma 9. It remains to compute the variance of the right side:

$$\begin{aligned} \text{Var} \left(\sum_{i=1}^k \frac{d_i}{\sqrt{p_i}} X_i - \frac{1}{\sqrt{C_m}} \sum_{i=1}^k \frac{d_i}{p_i} W_i \right) &= \sum_{i=1}^k \frac{d_i^2}{p_i} + \frac{1}{C_m} \sum_{i=1}^k \sum_{j=1}^k \frac{d_i d_j}{p_i p_j} \sigma_{ij} \\ &= \sum_{i=1}^k \frac{d_i^2}{p_i} + \frac{1}{C_m} \left(\sum_{i=1}^k \frac{d_i^2}{p_i} - 1 \right). \end{aligned}$$

From this we get

$$\frac{\sum_{i=1}^k \phi(N_{i,n}, \underline{Y}^i) - E(f^{-1} \log f^{-1}) + \log n - \sum_{i=1}^m i^{-1} + \gamma}{\sqrt{(n+2)^{-1}[F(m+1)E(f^{-2}) - 1]}} \xrightarrow{d} N(0, 1).$$

Replacing $\log n + \gamma$ in the expression above by $\sum_{i=1}^{n+1} i^{-1}$, which is asymptotically equivalent as was shown in the proof of Theorem 2, we obtain (26).

References

- N. Cressie (1976), *On the logarithms of high-order spacings*, *Biometrika* 63, 343–355.
 F. Czekala (1993), *Asymptotic distributions of statistics based on logarithms of spacings*, *Zastos. Mat.* 21, 511–519.
 G. E. Del Pino (1979) *On the asymptotic distribution of k-spacings with applications to goodness of fit tests*, *Ann. Statist.* 7, 1058–1065.
 J. R. Gebert and B. K. Kale (1969), *Goodness of fit tests based on discriminatory information*, *Statist. Hefte* 3, 192–200.
 S. R. Jammalamadaka and R. C. Tiwari (1986), *Efficiencies of some disjoint spacings tests relative to a χ^2 test*, in: M. L. Puri, J. Vilaplana and W. Wertz (eds.) *New Perspectives in Theoretical and Applied Statistics*, Wiley, New York, 311–318.
 B. K. Kale (1969), *Unified derivation of tests of goodness of fit based on spacings*, *Sankhyā Ser. A* 31, 43–48.

Franciszek Czekala
 Mathematical Institute
 University of Wrocław
 Pl. Grunwaldzki 2/4
 50-384 Wrocław, Poland
 E-mail: czekala@math.uni.wroc.pl

Received on 24.1.1995