*The Annals of Statistics*
1986, Vol. 14, No. 4, 1485–1496

# ASYMPTOTICALLY INDEPENDENT SCALE-FREE SPACINGS WITH APPLICATIONS TO DISCORDANCY TESTING

BY T. J. SWEETING

*University of Surrey*

Motivated by the construction of independent scale-free spacings for the exponential distribution, analogous quantities are defined for an arbitrary distribution, and their joint asymptotic behaviour studied under suitable tail conditions. These scale-free spacings provide a unified approach to the problem of consecutive discordancy testing when location and scale parameters are unknown. A normal example is discussed and the accuracy of the approximations assessed via some Monte Carlo studies.

**1. Introduction.** When testing for the existence of an unspecified number of upper or lower outliers, it is desirable to have a procedure which will not only detect the existence of outliers, but which will also indicate the number of outliers present. A type of consecutive discordancy test aimed at doing just this in the case of normal samples was proposed by Rosner (1975); a similar procedure for the exponential case is discussed in Kimber (1982). A major problem with such procedures, however, is the difficulty of calculating the null distributions. The process is greatly simplified if the consecutive test statistics involved are all *independent*. In the case of exponential samples, it is well known that the spacings (differences between adjacent order statistics) are all independent, and so form an appealing set of quantities on which to base a consecutive test. When the scale parameter $\sigma$ is unknown, however, it was shown in Sweeting (1983) that, somewhat surprisingly, the independence is retained if one uses an appropriate sequence of consecutive estimators of $\sigma$.

Turning now to samples from arbitrary distributions $F$ belonging to the domain of attraction of the Type III extreme-value distribution, in the same way one can use the result that the $k$ upper spacings when suitably normalized are asymptotically independent and exponentially distributed (Weissman, 1978). When, as is often the case, an unknown scale parameter $\sigma$ is present, it is shown here that one can construct a natural set of *scale-free* spacings by using consecutive estimators $\hat{\sigma}_i$ of $\sigma$ analogous to the exponential case. The consistency of these estimators, proved in the appendix, certainly guarantees the asymptotic independence and exponentiality of these quantities, but one would anticipate superior approximations if one were to use the *exact* distributions available in the exponential case. The method of proof here, based on a well-known exponential representation of spacings, yields a direct proof of Weissman's result; furthermore, it enables one to study the joint asymptotic behaviour of $k_n$ spacings where $k_n \propto n$.

Received October 1984; revised July 1985.

*AMS* 1980 *subject classifications*. Primary 62E20; secondary 62F05.

*Key words and phrases*. Scale-free spacings, consecutive discordancy tests, extreme value theory, regular variation.

These asymptotic results and their application to the problem of consecutive discordancy testing are discussed in Section 2; some details for specific distributions are given in Section 3. In Section 4 the consecutive test procedure is applied to some contaminated normal data given by Rosner (1975). Finally, in order to assess the accuracy of the approximate null distributions, some results from Monte Carlo investigations are presented in Section 5. These studies indicate that the approximations are of practical value, and hence provide a unified approach to the problem of consecutive discordancy testing when location and scale parameters are unknown.

**2. Asymptotic distributions and applications to discordancy testing.** In this section we introduce the scale-free spacings, present the required asymptotic results for these quantities and discuss their application to the problem of consecutive discordancy testing. Let $X_1, \dots, X_n$ be independent observations from a distribution $F$ with continuous density $f(x)$, positive over the range of $F$. For $t \geqq 1$ define the functions

$$U(t) = tfF^{-1}(1 - t^{-1}), \qquad L(t) = tfF^{-1}(t^{-1}).$$

Let $X_{(1)}, \dots, X_{(n)}$ be the order statistics and define

$$D_i = c_{i,n}\big(X_{(i)} - X_{(i-1)}\big), \qquad W_i = \sum_{j=2}^{i} D_j$$

for $i = 2, \dots, n$, where $c_{i,n} = nfF^{-1}((i-1)/n) = (n-i+1)U(n/(n-i+1))$.

When $F$ is the exponential distribution with mean $\sigma$, denoted here by $E(1/\sigma)$, then $c_{i,n} = (n-i+1)/\sigma$ and the quantities $D_i$ are all independent and $E(1)$. It was shown in Sweeting (1983) that the scale-free spacings $D_i/W_i$ are also independent with distribution functions given by $1-(1-z)^{i-2}$, $i = 2, \dots, n$. In this case it can be seen that $W_i = \sum_{j=1}^{i}(X_{(j)} - X_{(1)}) + (n-i)(X_{(i)} - X_{(1)})$ is the numerator of a one-sided Winsorized estimator $W_i/(i-1)$ of $\sigma$. Now let $F$ be an arbitrary distribution. Notice that if one writes $F_1(x) = F((x-\mu)/\sigma)$, then, with obvious notation, $U_1(x) = U(x)/\sigma$. Therefore the ratios $D_i/W_i$, $i = 2, \dots, n$, are scale free; we refer to such quantities as *scale-free spacings*. The question we wish to investigate is whether the known joint distribution of these scale-free spacings in the exponential case can be used as a suitable approximation more generally. When $F$ belongs to the domain of attraction of the Type III extreme-value distribution [see Galambos (1978) for example], it is known that for fixed $k$ the normalized spacings

$$(1) \qquad (n-i+1)U(n)\big(X_{(i)} - X_{(i-1)}\big), \qquad i = n-k+1, \dots, n,$$

are jointly asymptotically independent and $E(1)$; see Weissman (1978) for example. Moreover, using a representation given by Pyke (1965), one can deduce the asymptotic independence and exponentiality of any finite number of "central" $D_i$'s. These results suggest that the exact distributions in the exponential case may give useful approximations in the general case. Notice that it will be possible to replace the coefficient $U(n)$ in (1) by $U(n/(n-i+1))$, in line with our definition of $D_i$, whenever $U$ is *slowly varying*. In fact, under this condition one can give a rather elementary derivation of Weissman's result by using the exponential representation in Pyke (1965); the asymptotic independence and exponentiality of both upper and/or lower outliers may be deduced from Theorem 1 below. Moreover, the same method enables one to study the joint asymptotic behaviour of $k_n$ spacings when $k_n \propto n$ (Theorem 3). It is of interest to note that the slow variation of $U$ is actually a necessary and sufficient condition for uniform local convergence to the Type III extreme-value distribution (Sweeting, 1985).

THEOREM 1. *Suppose $U$ is slowly varying. Then there is a sequence $(Y_i;\ i \geqq 1)$ of independent $E(1)$ random variables such that*

$$D_{n-k+1}/Y_{n-k+1} \to_P 1$$

*as $n \to \infty$ for every fixed integer $k$.*

The consistency of the consecutive estimators of $\sigma$ based on the $W_i$ is given in the next result.

THEOREM 2. *Suppose that $L$ and $U$ are both regularly varying with finite exponents and let $k, l \geqq 1$ be fixed integers. Then*

$$\frac{W_{n-k+1} - W_l}{n-k-l+1} \to_P \sigma$$

*as $n \to \infty$ where $W_1 = 0$.*

The proofs are given in the Appendix. Theorems 1 and 2 immediately imply the following result for upper scale-free spacings. Suppose $L$ is regularly varying with finite exponent, and $U$ is slowly varying. Then for fixed $k \geqq 1$ the quantities

$$(2) \qquad Z_i = (n-i+1)D_{n-i+1}/W_{n-i+1}, \qquad i = 1, \dots, k,$$

are asymptotically $E(1)$ and independent. In the exponential case it can be shown as in Sweeting (1983) that these quantities are exactly independent with distributions $P(Z_i > z) = 1 - \{z/(n-i+1)\}^{n-i-1}$, and these distributions may be used in the general case as an alternative to the $E(1)$ distribution. The quantities $Z_i$ are convenient to use for consecutive discordancy testing of the "inside-out" type proposed by Rosner (1975), who discussed Gaussian samples. Specifically, to test for up to $k$ upper outliers, one first carries out a test for $k$ outliers based on $Z_k$. If $Z_k > A_k$, a predetermined critical value, the procedure is terminated and $k$ upper outliers declared. Otherwise one carries out a test for $k-1$ outliers based on $Z_{k-1}$, and so on. The choice of constants $A_1, \dots, A_k$ for a specified overall significance level is particularly simple if we use the approximate independence of the $Z_i$, and their approximate sampling distributions. For a size-$\alpha$ test, choose any constants $\lambda_1, \dots, \lambda_k \geqq 0$ with $\sum_{i=1}^k \lambda_i = 1$ and then determine the $A_i$ from $P(Z_i \leqq A_i) = (1-\alpha)^{\lambda_i}$, $i = 1, \dots, k$. The size $\alpha_i$ of the

$i$th test is $1 - (1 - \alpha)^{\lambda_i}$ and the overall significance level (= probability of wrongly declaring one or more outliers) is $1 - \prod_{i=1}^{k}(1 - \alpha_i) = \alpha$. The procedure is illustrated in Section 4 by means of an example discussed in Rosner (1975).

An alternative procedure would be to base the test of $k$ upper outliers on $X_{(n-k+1)}/W_{n-k+1}$, rather than on $Z_k$. This would be a reasonable approach where "masking" by a $(k + 1)$th outlier is a possibility; in such a case a test based on $D_{n-k+1}$ may fail to detect any outliers. The slow variation of $U$ ensures the convergence in distribution of $X_{(n-k+1)}$, suitably normalized (Sweeting, 1985). Furthermore, under the same condition it may be verified that

$$Y_k = U(n/k)\left[(n - k + 1)X_{(n-k+1)}/W_{n-k+1} - F^{-1}((n - k + 1)/n)\right]$$

has an asymptotic log Gamma $(k, k)$ distribution, and is asymptotically independent of $Z_i$, $i = 1, \ldots, k - 1$. (When an unknown location parameter is present, a similar statistic based on $X_{(n-k+1)} - X_{(1)}$ may be constructed.) It is not, however, the purpose of the present paper to examine the relative merits of alternative test procedures, but rather to point out the good approximations available when discordancy tests are based on the statistics proposed here. It is also interesting to note that the regular variation of $U$ with finite negative (positive) exponent is implied by the uniform local convergence to a Type I (Type II) extreme-value distribution. These facts may be deduced from Theorem 1 in Sweeting (1985).

For testing lower outliers, simply apply the above results to $-X_1, \ldots, -X_n$ to see that if $U$ is regularly varying with finite exponent and $L$ is slowly varying, then for fixed $k$ the quantities

$$Z_i' = (n - i + 1)D_{i+1}/(W_n - W_i), \qquad i = 1, \ldots, k,$$

are asymptotically $E(1)$ and independent. As an alternative, one can use the exact distributions in the exponential case (which are the same as those for the $Z_i$). Finally, the case of both lower and upper outliers can be tackled by "working from the middle." Thus if $L$ and $U$ are both slowly varying, then Theorems 1 and 2 imply that for fixed $k$ the quantities

$$L_i = (n - 2i + 1)D_{i+1}/(W_{n-i} - W_i),$$
$$U_i = (n - 2i + 2)D_{n-i+1}/(W_{n-i+1} - W_i)$$

for $i = 1, \ldots, k$ are asymptotically $E(1)$ and independent. In the exponential case these quantities are exactly independent with distributions given by

$$P(L_i > l) = \left(1 - \{l/(n - 2i + 1)\}\right)^{n-2i-1},$$
$$P(U_i > u) = \left(1 - \{u/(n - 2i + 2)\}\right)^{n-2i}$$

(Sweeting, 1983). One possible procedure would consist of performing a sequence of tests based on $L_k, U_k, L_{k-1}, \ldots$ and so on. If a significant result is obtained at any stage, say at $U_j$, then $j$ upper outliers are declared and the lower outlier procedure continues, based on the approximate distributions of $L'_{j-1}, \ldots, L'_1$ where $L'_i = (n - j - i + 1)D_{i+1}/(W_{n-j} - W_i)$.

In practice, one usually wishes to test for the existence of a greater number of outliers in a larger sample. The main drawback to the above results is that $k$ must remain fixed as $n \to \infty$. It is possible however to obtain stronger results using the method of proof of Theorems 1 and 2. The following theorem (proved in the Appendix) is expressed in terms of the upper scale-free spacings; similar results hold for the ordinary spacings, lower spacings etc.

Let $\alpha > 0$ be the predicted overall size of a consecutive discordancy test for $k_n$ upper outliers based on exponential distributions; that is

$$1 - \alpha = P(Y_{n-i+1} \leqq x_{i,n},\ i = 1, \dots, k_n),$$

where $Y_1, \dots, Y_n$ are independent $E(1)$ r.v.'s and $x_{i,n}$, $i = 1, \dots, k_n$, are the critical values. Let $\alpha_{i,n} = P(Y_{n-k+1} > x_{i,n}) = e^{-x_{i,n}}$, the size of the single test based on $Y_{n-i+1}$.

THEOREM 3. *Suppose that $L$ is regularly varying with finite exponent and that $U$ is slowly varying. Let $k_n = [(1-\pi)n]$ for some $\pi > 0$. Then under the condition*

$$\sum_{i=1}^{k_n} \alpha_{i,n}^{1-\delta} \text{ bounded for some } \delta > 0, \tag{3}$$

*one has*

$$P(Z_i \leqq x_{i,n},\ i = 1, \dots, k_n) \to 1 - \alpha$$

*as $n \to \infty$.*

Thus if a fixed *proportion* of outliers is to be investigated, the above result tells us that, provided the sizes of consecutive tests satisfy (3), the size of our approximate test will be asymptotically correct. Note that $\prod_{i=1}^{k_n}(1-\alpha_{i,n}) = 1 - \alpha$ implies that $0 < (1-\alpha)|\log(1-\alpha)| < \sum_{i=1}^{k_n}\alpha_{i,n} < |\log(1-\alpha)| < \infty$. In particular, if $\alpha_{i,n} = \alpha_i$ then $\sum_{i=1}^{\infty}\alpha_i < \infty$ and we require $\sum_{i=1}^{\infty}\alpha_i^{1-\delta} < \infty$, which will of course be the case for a large choice of sequences $(\alpha_i)$. In many cases, however, where $\alpha_{i,n}$ depends on $n$, condition (3) is not satisfied; generally there needs to be a few terms which "dominate." In particular, we see that if $\max_{1 \leqq i \leqq k_n}\alpha_{i,n} \to 0$ then (3) cannot hold. The extent to which (3) may be weakened would appear to the author to depend on the underlying distribution $F$, and it is conjectured that Theorem 3 cannot *generally* be improved.

**3. Specific distributions.** We briefly discuss some of the main distributions for which the procedures of the preceding section are available. In each case we give expressions for the constants $c_{i,n}$ appearing in the definition of the $D_i$'s and consider the variation of the functions $U(t)$ and $L(t)$.

*Normal distribution.* Here $U(t) \sim (2\log t)^{1/2}$, so by symmetry both $U$ and $L$ are slowly varying, and $c_{i,n} = \exp(-\frac{1}{2}\{\Phi^{-1}((i-1)/n)\}^2)$. One can therefore readily construct consecutive tests for upper and/or lower outliers in the case of a normal population with unknown mean and variance. Consecutive discordancy tests for normal populations discussed in Rosner (1975, 1977, 1983), Hawkins (1979) and Prescott (1979) rely on Monte Carlo evaluations of percentage points, which become computationally prohibitive for large values of $k$.

*Weibull distribution*. Suppose $P(T > t) = \exp(-\{\lambda t\}^{\gamma})$ where $\lambda, \gamma > 0$ are unknown, and let $X = \log T$. Then $X \sim F((x - \mu)/\sigma)$ where $F(x) = 1 - \exp(-e^{x})$, $\sigma = \gamma^{-1}$, $\mu = -\log \lambda$. Here (taking $\mu = 0$, $\sigma = 1$) $U(t) = \log t$ and $L(t) = (t - 1)\log(t/(t - 1))$ are both slowly varying, and $c_{i,n} = (n - i + 1)\log(n/(n - i + 1))$. Both lower and upper outliers may therefore be tested.

Alternatively, suppose that $\gamma$ is known, but that there is possibly an unknown location parameter $\nu$, i.e., $P(T > t) = \exp(-\lambda\{t - \nu\}^{\gamma})$. Taking $\nu = 0$, $\lambda = 1$, we have $U(t) = \gamma(\log t)^{(\gamma-1)/\gamma}$, which is slowly varying, and $L(t) \sim \gamma t^{1/\gamma}$, which is regularly varying with exponent $\gamma^{-1}$. Thus in this case one is only able to apply the procedure for upper outliers.

*Logistic distribution*. Here $F(x) = 1 - (1 + e^{x})^{-1}$ and $U(t) = (t - 1)/t$. Thus, by symmetry, $U$ and $L$ are both slowly varying, $c_{i,n} = (i - 1)(n - i + 1)/n$, and one can test for both upper and lower outliers.

*Gamma distribution*. Suppose $f(x) \propto (x - \nu)^{\gamma-1}\exp(-\lambda\{x - \nu\})$, $\lambda > 0$, and the shape parameter $\gamma > 0$ is *known*. Taking $\nu = 0$, $\lambda = 1$ one finds that $U(t) \to 1$ and $L(t) \sim \Gamma(1 + \gamma)^{1-(1/\gamma)}t^{1/\gamma}$, which is regularly varying with exponent $1/\gamma$. Upper outliers may therefore be treated on taking $c_{i,n} = \{F^{-1}((i - 1)/n)\}^{\gamma-1}\exp\{-F^{-1}((i - 1)/n)\}$. If the location parameter $\nu$ is also known, it is possible to test for both upper and lower outliers by transforming to $X_1 = \log X$. Then $U_1(t) \sim F^{-1}(1 - t^{-1})$ and $L_1(t) \sim \Gamma(1 + \gamma)$, which are both slowly varying.

**4. An example.** In order to illustrate the procedure in the normal case and also to compare it with the extreme studentized deviate (ESD) many outlier procedure proposed by Rosner (1975), we consider the simulated data given by Rosner. Twenty pseudo-random $N(0, 1)$ deviates were generated and two perturbed samples created. For perturbed sample A, 5 was added to $x_{20}$, while for perturbed sample B, 5 was added to both $x_{19}$ and $x_{20}$. Rosner applies the ESD procedure for up to two upper outliers to each of the three samples. Here we apply a consecutive test based on the upper scale-free spacings; for illustration and comparison with Rosner, we take $\lambda_1 = \lambda_2 = 0.5$.

The critical values are calculated from $(1 - A_i/\{n - i + 1\})^{n-i-1} = \alpha_i$, where $\alpha_i = 1 - (1 - \alpha)^{\lambda_i}$, $i = 1, 2$ and $\alpha$ is the overall size of the test. For $\alpha = 0.05$ we find $A_1 = A_2 = 3.69$, and for $\alpha = 0.01$, we have $A_1 = 5.10$, $A_2 = 5.09$. The values $c_{i,n}$ are calculated from the normal scores as given in Section 3. For the original uncontaminated sample we find $W_{19} = 2.693$, $D_{19} = 0.220$ and $D_{20} = 0.039$. From (2) the consecutive test statistics are therefore $Z_2 = 1.55$, $Z_1 = 0.29$ and we declare no outliers present.

TABLE 1

*Probability of declaring* (i) 0, 1, 2 *or* 3 *outliers in a consecutive test based on* 5000 *random samples of size* $n = 50$; (ii) 0, 1 *or* 2 *outliers in a consecutive test based on* 5000 *random samples of size* $n = 20$

| | Number of outliers declared | | | | | | |
|---|---|---|---|---|---|---|---|
| | $n = 50$ | | | | $n = 20$ | | |
| | 0 | 1 | 2 | 3 | 0 | 1 | 2 |
| Predicted (exact for exponential ($U$)) | 0.950 | 0.020 | 0.015 | 0.015 | 0.950 | 0.025 | 0.025 |
| Gamma ($U$), shape = 2 | 0.956 | 0.018 | 0.011 | 0.015 | 0.952 | 0.024 | 0.024 |
| Gamma ($U$), shape = 3 | 0.960 | 0.014 | 0.014 | 0.012 | 0.962 | 0.018 | 0.020 |
| Logistic ($U$ or $L$) | 0.952 | 0.020 | 0.013 | 0.015 | 0.964 | 0.018 | 0.018 |
| Normal ($U$ or $L$) | 0.969 | 0.008 | 0.011 | 0.012 | 0.968 | 0.014 | 0.018 |
| Weibull ($U$), shape = 2 | 0.964 | 0.012 | 0.011 | 0.013 | 0.964 | 0.017 | 0.019 |
| Weibull ($U$), shape = 3 | 0.970 | 0.009 | 0.009 | 0.012 | 0.970 | 0.013 | 0.017 |
| Log–Weibull ($L$) | 0.954 | 0.018 | 0.012 | 0.016 | 0.952 | 0.024 | 0.024 |
| Log–Weibull ($U$) | 0.973 | 0.007 | 0.010 | 0.010 | 0.976 | 0.011 | 0.013 |

$L$ = lower outliers, $U$ = upper outliers.

For perturbed sample A, the corresponding quantities are $W_{19} = 2.946$, $D_{19} = 0.066$, $D_{20} = 0.664$ giving $Z_2 = 0.43$ and $Z_1 = 3.68$, which is very nearly significant at the 5% level, indicating one upper outlier. For perturbed sample B, $W_{19} = 4.214$, $D_{19} = 1.129$, $D_{20} = 0.287$ giving $Z_2 = 5.09$ and $Z_1 = 1.28$. Thus $Z_2$ is just significant at the 1% level, and there is strong evidence of two upper outliers. The corresponding test statistics based on the ESD procedure of Rosner (1975) lie above the 5% and 1% critical values (as calculated by Monte Carlo) for samples A and B, respectively, suggesting that the ESD procedure is slightly more sensitive. However, the Monte Carlo results presented in the next section show that our procedure is actually conservative, which is likely to be the reason for the apparent relative inefficiency. The true critical values will be slightly lower than the approximate values given here, in line with the results from the ESD procedure.

**5. Monte Carlo results.** We conclude by presenting some Monte Carlo results indicating the level of approximation error to be expected for the null distributions in small to moderate sized samples. Data sets of various sizes were generated from the distributions discussed in Section 3 and listed in Table 1, and consecutive discordancy tests based on the scale-free spacings carried out. Approximate probabilities of rejecting $0, 1, \ldots, k$ outliers were obtained from the exact joint distribution of the scale-free spacings in the exponential case, as discussed in Section 2. The relative frequencies with which the test rejected outliers over a large number of repetitions were compared with these probabilities for various values of $k$ and $\alpha_1, \ldots, \alpha_k$. All computations were carried out on a PRIME 750 computer, using FORTRAN programmes. The data simulations used pseudo-random number and associated routines available on the Numerical Algorithms Group (NAG) library.

Table 1 presents results based on 5000 random samples of size $n = 50$ for each of the indicated cases. The predicted relative frequencies here are the exact probabilities in the exponential case, taking $\lambda_1 = 0.4$, $\lambda_2 = 0.3$ and $\lambda_3 = 0.3$. The accuracy of the figures is $\pm 0.006$ for the overall significance level and $\pm 0.004$ for the individual rejection probabilities (at a 95% confidence level). The best results here were obtained from the logistic distribution and the lower tail of the log–Weibull distribution (equivalently, the upper tail of the Gumbel distribution). The worst case was the log–Weibull, upper outliers test, which gave an overall significance level of approximately 3%. However, in the poorest cases the tests were always found to be conservative. These results can be improved upon slightly by using alternative arguments in the function $U$ of Section 2; for example, by replacing $i - 1$ by $i - \frac{1}{2}$. The optimal choice of argument does depend on the underlying distribution, however, although $i - \frac{1}{2}$ appears to give generally good results. Corresponding results are presented in the same table for $n = 20$ and $k = 2$, taking $\lambda_1 = \lambda_2 = 0.5$, and it can be seen that the approximations are still quite acceptable. Again, the log–Weibull distribution (lower tail) fared well here, as did the Gamma distribution with shape parameter 2.

The difference between using an exponential or a Beta approximation will be most marked, of course, for smaller sample sizes. For example, in the normal case when $n = 10$, $k = 3$ ($\lambda_1 = 0.4$, $\lambda_2 = 0.3$, $\lambda_3 = 0.3$ as before) an exponential approximation gives a true overall significance level of less that 1% (based on 5000 repetitions) when supposedly testing at the 5% level. The true level using the Beta approximation is just under 3%, which is surprisingly good considering the small sample size.

## APPENDIX

*Proofs of asymptotic results.* The joint asymptotic distribution of $k$ upper or lower spacings was obtained by Weissman (1978), from which the joint asymptotic distribution of the $D_i$ may be deduced using the slow variation of $U$ and Lemma 2 in Sweeting (1985). For the joint asymptotic distribution of the scale-free spacings it then suffices to show that $(W_{n-k+1} - W_l)/(n - k - l + 1)$ is a consistent estimator of $\sigma$ for fixed $k$ and $l$. Under appropriate conditions this may be deduced from general results on linear combinations of functions of order statistics in Chernoff et al. (1967). However, we shall give an elementary argument based on an exponential representation of the spacings of a general distribution. Moreover, this approach is a very natural one for the study of the joint asymptotic distribution of the extreme spacings, and it is possible to study the asymptotic behaviour when the number of spacings is not fixed.

Let $Y_1, \dots, Y_n$ be independent $E(1)$ random variables and let $U_i = H(\sum_{j=1}^{i} Y_j/(n - j + 1))$ where $H(z) = 1 - e^{-z}$ is the $E(1)$ distribution function. Note that the $U_i$ are the order statistics from a uniform $[0, 1]$ sample. Pyke (1965) gives the following representation of the spacings of $F$ when $F$ possesses a continuous density $f$ positive throughout its range:

$$X_{(i)} - X_{(i-1)} = (n-i+1)^{-1} Y_i r(A_i), \tag{4}$$

where $U_{i-1} < A_i < U_i$ and $r(u) = (1-u)/f(F^{-1}(u)) = [U((1-u)^{-1})]^{-1}$ in our notation. Set $B_i = (1-A_i)^{-1}$; then (4) becomes

$$(n-i+1)U(B_i)\left(X_{(i)} - X_{(i-1)}\right) = Y_i.$$

Let $\xi_i = (i-1)/n$. Since $E(U_i) = i/(n+1)$ and $B_i \approx (1-U_i)^{-1}$, it is reasonable to expect that under suitable conditions $U(B_i)$ may be replaced by $U((1-\xi_i)^{-1})$ for large $n$. We need the following two simple lemmas.

LEMMA 1. *$M_n = \max_{R_n(\varepsilon)}|(1-\xi_i)B_i - 1| \to_P 0$, where $R_n(\varepsilon) = \{i\colon 1 \le i \le [(1-\varepsilon)n]\}$ and $\varepsilon > 0$.*

PROOF. We have $M_n \le (\max_{R_n(\varepsilon)}|A_i - \xi_i|)/(1 - U_{[(1-\varepsilon)n]})$, and the result follows immediately from $U_{i-1} < A_i < U_i$ and the Glivenko–Cantelli lemma $\max_{1\le i\le n}|U_i - (i/n)| \to_P 0$. □

LEMMA 2. *For all $\eta > 0$ there exist positive constants $c_1, c_2$ such that*

$$P\left(\bigcap_{i=i}^{n}\{c_1 < (1-\xi_i)B_i < c_2\}\right) < 1 - \eta$$

*for all $n$.*

PROOF. All maxima are taken over $1 \le i \le n$. Let $V_i = -\log(1-U_i)$; then the $V_i$ are the order statistics from an exponential sample, and the lemma will follow if we prove that $\max|V_i - \log(n/(n-i+1))|$ is stochastically bounded. The representation $V_i = \sum_{j=1}^{i} Y_j/(n-j+1)$ gives $\mu_i = E(V_i) = \sum_{j=1}^{i}(n-j+1)^{-1}$, $\kappa_{i,2} = \sum_{j=1}^{i}(n-j+1)^{-2} \le 2(n-i+1)^{-1}$ and $\kappa_{i,4} = 6\sum_{j=1}^{i}(n-j+1)^{-4} < 8(n-i+1)^{-3}$, where $\kappa_{i,r}$ is the $r$th cumulant of $V_i$. It follows that $E(V_i - \mu_i)^4 = \kappa_{i,4} + 3\kappa_{i,2}^2 < 20(n-i+1)^{-2}$ and so

$$\begin{aligned} P\left(\max|V_i - \mu_i| > c\right) &\le \sum_{i=1}^{n} P\left(|V_i - \mu_i| > c\right) \\ &\le c^{-4}\sum_{i=1}^{n} E(V_i - \mu_i)^4 > 20c^{-4}\sum_{k=1}^{\infty} k^{-2}. \end{aligned}$$

Thus $\max|V_i - \mu_i|$ is stochastically bounded, and the result follows since $|\log(n/(n-i+1)) - \mu_i| \le 1$ for all $i = 1, \dots, n$. □

We now give the proofs of Theorems 1–3 stated in Section 2.

PROOF OF THEOREM 1. Let $\eta > 0$ and $\mathscr{F}_n$ the set on which $0 < c_1 < (1-\xi_i)B_i < c_2 < \infty$ for $i = 1, \dots, n$ [we actually only need $c_1 < (k/n)B_{n-k+1} < c_2$ here]. Then

$$Y_{n-k+1}/D_{n-k+1} = U(n/k)/U(B_{n-k+1}) \to 1,$$

with probability one on $\mathscr{F}_n$, since $U(t)/U(tx) \to 1$ as $t \to \infty$ uniformly on compact intervals of $\mathbb{R}^+$ from the slow variation of $U$. We can therefore choose $n_0$ so large that $|Y_{n-k+1}/D_{n-k+1} - 1| < \eta$ on $\mathscr{F}_n$ for $n > n_0$, and hence $P(|Y_{n-k+1}/D_{n-k+1} - 1| > \eta) \le P(\bar{\mathscr{F}}_n) < \eta$ and the result follows. □

PROOF OF THEOREM 2. We show that

$$R_n = \sum_{i=l+1}^{n-k+1} D_i \Big/ \sum_{i=l+1}^{n-k+1} Y_i \to_P 1 \tag{5}$$

as $n \to \infty$, from which Theorem 2 follows. Let $0 < \varepsilon < \frac{1}{2}$ and write $\mathscr{S}_n = \{i\colon l + 1 \leqq i \leqq n - k + 1\}$. We have

$$\begin{aligned}\Big|\sum_{\mathscr{S}_n}(D_i - Y_i)\Big| \le &\sum_{\mathscr{S}_n \cap \mathscr{B}_n} \Big|U\big((1-\xi_i)^{-1}\big)/U(B_i) - 1\Big| Y_i \\ &+ \sum_{\mathscr{A}_n \cup \mathscr{C}_n} \Big(U\big((1-\xi_i)^{-1}\big)/U(B_i) + 1\Big) Y_i,\end{aligned}$$

where

$$\begin{aligned}\mathscr{A}_n &= \{i\colon 1 \le i < [\varepsilon n]\},\\ \mathscr{B}_n &= \{i\colon [\varepsilon n] \le i < [(1-\varepsilon)n],\\ \mathscr{C}_n &= \{i\colon [(1-\varepsilon)n] \le i \le n\}.\end{aligned}$$

Then for $n$ sufficiently large

$$\begin{aligned}|R_n - 1| \le &\max_{\mathscr{B}_n} \Big|U\big((1-\xi_i)^{-1}\big)/U(B_i) - 1\Big| \\ &+ \max_{\mathscr{A}_n \cup \mathscr{C}_n} \Big(U\big((1-\xi_i)^{-1}\big)/U(B_i) + 1\Big)\Big(\sum_{\mathscr{A}_n \cup \mathscr{C}_n} Y_i \Big/ \sum_{\mathscr{B}_n} Y_i\Big) \qquad (6)\\ = &\, I_n + J_n, \quad \text{say}.\end{aligned}$$

$I_n \to_P 0$ follows from Lemma 1, since $U(x)$ is uniformly continuous and $\inf U(x) > 0$ over compact sets of $(0, \infty)$ (because $f$ is continuous and positive throughout the range of $F$). Let $\eta = \varepsilon$ and $\mathscr{F}_n$ be the set on which $0 < c_1 < (1 - \xi_i)B_i < c_2 < \infty$, $i = 1, \dots, n$. Then on $\mathscr{F}_n$, $M_n = \max U((1 - \xi_i)^{-1})/U(B_i)$ is bounded a.s. This follows from the regular variation of $U$ and the fact that if $S(t)$ is slowly varying then $S(tx)/S(t) \to 1$ as $t \to \infty$ uniformly on compact intervals of $\mathbb{R}^+$. Similarly $M_n^- = \max U((1 - \xi_i)^{-1})/U(B_i)$ bounded follows by considering the variables $-X_i$, $i = 1, \dots, n$ and noting that $L(t) = U^-(t)$, where $U^-$ is the function $U$ for the variables $-X_i$. Finally, using Markov's inequality,

$$P\Big(\sum_{\mathscr{A}_n \cup \mathscr{C}_n} Y_i > u \sum_{\mathscr{B}_n} Y_i\Big) \le 2\varepsilon n E\Big(\sum_{\mathscr{B}_n} Y_i\Big)^{-1} \Big/ u \le c_5 \varepsilon / u,$$

and so $\limsup_{n \to \infty} P(|R_n - 1| > u) \le c_6 \varepsilon / u + \varepsilon$ and (5) follows since $\varepsilon$ was arbitrary. □

PROOF OF THEOREM 3. We first show that

$$\max_{i>[\pi n]} |Y_i/Z_{n-i+1} - 1| \to_P 0. \tag{7}$$

We have

$$Y_i/Z_{n-i+1} = \left[U(B_i)/U\left((1-\xi_i)^{-1}\right)\right]\left[\sum_{j=2}^{i} D_j \Big/ \sum_{j=2}^{i} Y_j\right]\left[\sum_{j=2}^{i} Y_j/(i-1)\right],$$

and we consider each term in square brackets in turn. The notation here is taken from the proof of Theorem 2.

(a) Let $E_i = |U(B_i)/U((1-\xi_i)^{-1})|$; we show that $\max_{i>[\pi n]} E_i \to_P 0$. First, $\max_{\mathscr{B}_n} E_i \to_P 0$ since $I_n \to_P 0$ from the proof of Theorem 2, and it suffices to show that $\max_{\mathscr{C}_n} E_i \to_P 0$. But on the set $\mathscr{F}_n$ we may choose $\varepsilon$ so small that $E_i \le \eta$ for all $i \in \mathscr{C}_n$, since $U(tx)/U(t) \to 1$ as $t \to \infty$ uniformly on compact intervals of $\mathbb{R}^+$. Hence $P(\max_{\mathscr{C}_n} E_i > \eta) \le P(\bar{\mathscr{F}}_n) < \eta$ as required.

(b) We show that $R'_n = \max_{i>[\pi n]} |\sum_{j=2}^{i} D_j / \sum_{j=2}^{i} Y_j - 1| \to_P 0$. Take $\varepsilon < \frac{1}{2}\pi$. Inspecting the proof of Theorem 2, we see that $R'_n$ is less than the right-hand side of (6) on replacing $\sum_{\mathscr{B}_n} Y_i$ by $\sum_{\mathscr{B}'_n} Y_i$ where $\mathscr{B}'_n = \{i\colon [\varepsilon n] \le i \le [\pi n]\}$ and, exactly as in the proof of Theorem 2, we find $P(\sum_{\mathscr{A}_n \cup \mathscr{C}_n} Y_i > u \sum_{\mathscr{B}'_n} Y_i) \le c_7 \varepsilon/(u\pi)$, and hence $R'_n \to_P 0$.

(c) We have

$$P\left(\max_{i>[\pi n]}\left|\sum_{j=2}^{i} Y_j/(i-1) - 1\right| > u\right) \le \sum_{i>[\pi n]} P\left(\left|\sum_{j=2}^{i} Y_j - (i-1)\right| > (i-1)u\right)$$
$$\le 9u^{-4} \sum_{i>[\pi n]} (i-1)^{-2} \to 0$$

as $n \to \infty$, using Markov's inequality and $E(T-\alpha)^4 = 3\alpha(\alpha+2)$ when $T \sim \Gamma(\alpha, 1)$. Thus $\max_{i>[\pi n]} |\sum_{j=2}^{i} Y_j/(i-1) - 1| \to_P 0$.

The convergence in (7) now follows from (a), (b) and (c).

Let $0 < \varepsilon < \frac{1}{2}\delta$, $T_n = \max_{i<k_n}(Y_{n-i+1}/x_{i,n})$ and write $\Delta_n^+ = P(1 < T_n \le 1 + \varepsilon)$, $\Delta_n^- = P(1 - \varepsilon < T_n \le 1)$. Standard manipulations give

$$\begin{aligned}&\left|P\left(\max_{i\le k_n}(Z_i/x_{i,n}) \le 1\right) - (1-\alpha)\right| \\ &\qquad \le \max(\Delta_n^+, \Delta_n^-) + P\left(\max_{i\le k_n}|(Y_{n-i+1}/Z_i) - 1| > \varepsilon\right),\end{aligned} \tag{8}$$

recalling that $1 - \alpha = P(T_n \le 1)$. Suppressing $i \le k_n$ in all sums and products and writing $\alpha_{i,n} = \alpha_i$ for brevity, we have

$$\begin{aligned}\Delta_n^+ &= (1-\alpha)\left[\left(P(T_n \le 1+\varepsilon)/P(T_n \le 1)\right) - 1\right] \\ &= (1-\alpha)\left[\left(\Pi\left(1 - \alpha_i^{1+\varepsilon}\right)/(1-\alpha_i)\right) - 1\right] \\ &= (1-\alpha)\left[\exp\left(\sum \log\{1 + \alpha_i(1-\alpha_i^{\varepsilon})/(1-\alpha_i)\}\right) - 1\right] \\ &\le (1-\alpha)\left[\exp\left(\varepsilon \sum \alpha_i |\log \alpha_i|/(1-\alpha_i)\right) - 1\right] < c_8 \varepsilon\end{aligned}$$

(using $1 - e^{-x} \le x,\ x > 0$) provided $\Sigma\alpha_i|\log\alpha_i|$ is bounded, since all $\alpha_i \le \alpha$. Also

$$\begin{aligned}\Delta_n^- &= (1-\alpha)\left[1 - \left(P(T_n \le 1-\varepsilon)/P(T_n \le 1)\right)\right]\\ &= (1-\alpha)\left[1 - \exp\left(-\sum\log\left\{(1-\alpha_i)/\left(1-\alpha_i^{1-\varepsilon}\right)\right\}\right)\right]\\ &\le (1-\alpha)\sum\log\left\{1 + \alpha_i\left(\alpha_i^{-\varepsilon}-1\right)/\left(1-\alpha_i^{1-\varepsilon}\right)\right\}\\ &\le c_9\sum\alpha_i\left(\exp\{\varepsilon|\log\alpha_i|\} - 1\right) \le c_9\varepsilon\sum\alpha_i^{1-\varepsilon}|\log\alpha_i|,\end{aligned}$$

using $e^x - 1 \le xe^x,\ x > 0$, and from (7) and (8) the theorem will be proved if we show that $\Sigma\alpha_i^{1-\delta/2}|\log\alpha_i|$ is bounded. Let $\mathscr{E}_n = \{i\colon\ \alpha_{i,n} < a\}$ where $0 < a < 1$ is any number satisfying $a^{\delta/2}|\log a| < 1$. Splitting the sum over $\mathscr{E}_n$ and $\bar{\mathscr{E}}_n$ we have

$$\begin{aligned}\sum\alpha_i^{1-\delta/2}|\log\alpha_i| &\le \sum_{\mathscr{E}_n}\alpha_i^{1-\delta} + |\bar{\mathscr{E}}_n||\log a|\\ &\le \sum\alpha_i^{1-\delta} + c_{10}|\log a|/a,\end{aligned}$$

since $c_{10} \ge \Sigma\alpha_i \ge |\bar{\mathscr{E}}_n|a$, and the result follows. □

## REFERENCES

CHERNOFF, H., GASTWIRTH, J. L. and JOHNS, M. V., JR. (1967). Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *Ann. Math. Statist.* **38** 52–72.

GALAMBOS, J. (1978). *The Asymptotic Theory of Extreme Order Statistics.* Wiley, New York.

HAWKINS, D. M. (1979). Fractiles of an extended multiple outlier test. *J. Statist. Comput. Simulation* **8** 227–236.

KIMBER, A. C. (1982). Tests for many outliers in an exponential sample. *Appl. Statist.* **31** 263–271.

PRESCOTT, P. (1979). Critical values for a sequential test for many outliers. *Appl. Statist.* **28** 36–39.

PYKE, R. (1965). Spacings. *J. Roy. Statist. Soc. Ser. B* **27** 395–436.

ROSNER, B. (1975). On the detection of many outliers. *Technometrics* **17** 221–227.

ROSNER, B. (1977). Percentage points for the RST many outlier procedure. *Technometrics* **19** 307–312.

ROSNER, B. (1983). Percentage points for a generalized ESD many outlier procedure. *Technometrics* **25** 165–172.

SWEETING, T. J. (1983). Independent scale-free spacings for the exponential and uniform distributions. *Statist. Probab. Lett.* **1** 115–119.

SWEETING, T. J. (1985). On domains of uniform local attraction in extreme value theory. *Ann. Probab.* **13** 196–205.

WEISSMAN, I. (1978). Estimation of parameters and large quantiles based on the $k$ largest observations. *J. Amer. Statist. Assoc.* **73** 812–815.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SURREY
GUILDFORD SURREY GU2 5XH
ENGLAND