J. Appl. Prob. **23**, 922–936 (1986)
Printed in Israel

ASYMPTOTIC BEHAVIOR OF HILL'S ESTIMATE AND APPLICATIONS

GANE SAMB LO,* *Université Paris VI*

Abstract

The problem of estimating the exponent of a stable law is receiving an increasing amount of attention because Pareto's law (or Zipf's law) describes many biological phenomena very well (see e.g. Hill (1974)). This problem was first solved by Hill (1975), who proposed an estimate, and the convergence of that estimate to some positive and finite number was shown to be a characteristic of distribution functions belonging to the Fréchet domain of attraction (Mason (1982)). As a contribution to a complete theory of inference for the upper tail of a general distribution function, we give the asymptotic behavior (weak and strong) of Hill's estimate when the associated distribution function belongs to the Gumbel domain of attraction. Examples, applications and simulations are given.

ORDER STATISTICS; DOMAIN OF ATTRACTION; FUNCTIONS SLOWLY VARYING AT ZERO; NORMING CONSTANTS; CENTRAL LIMIT THEOREM; WEAK LAW OF LARGE NUMBERS

1. Introduction and results

For the last 100 years, Zipf's (or Pareto's) law, defined by

$$G(x) = 1 - C \cdot x^{-1/c}, \qquad C > 0, \quad c > 0, \quad \text{as } x \uparrow +\infty \tag{1.1}$$

has received an increasing amount of attention (cf. Boulenger (1885)). As noted by Hill (1974), who gives in Hill (1970), (1974), (1975) a useful survey of this topic, many biological phenomena are well described by the probabilistic model (1.1).

This has been the justification for much work on the problem of estimating c. First, Hill (1975) proposed, for a fixed integer k, $1 \leqq k < n$,

$$T_n = k^{-1} \sum_{i=1}^{i=k} (X_{n-i+1,n} - X_{n-k,n})$$

as the conditional maximum likelihood estimate of c under the assumption that $X_1, X_2, \cdots, X_n$ are independent and identical copies of a random variable X

Received 7 February 1985; revision received 10 October 1985.

* Postal address: 81 Résidence d'Athis, 26 Rue de la Plaine Basse, 91200 Athis-Mons, France.

whose distribution function $F(x) = P(X \leqq x)$ is such that $F(\log(\cdot))$ satisfies (1.1), where $X_{1,n} \leqq X_{2,n} \leqq \cdots \leqq X_{n,n}$ denote the order statistics of $X_1, \cdots, X_n$.

The asymptotic study of T_n (consistency and asymptotic normality) has been considerably developed by various authors such as Hall (1982), Csörgő et al. (1985), etc. In fact, Mason (1982) proved that for an unbounded random variable X, i.e., $\sup\{x, F(x) < 1\} = +\infty$, and for any real c, $0 < c$, one has

$$T_n \to c, \quad \text{in probability } (\xrightarrow{\text{P}}) \text{ for any sequence } k_n = k$$

satisfying

$$\text{(K)} \qquad 0 < k_n < n, \quad k_n \to +\infty, \quad k_n/n \to 0, \qquad \text{as } n \to +\infty$$

if and only if

$$\text{(Ac)} \qquad \forall t > 0, \quad \lim_{x \uparrow +\infty} \frac{1 - F(\log(tx))}{1 - F(\log(x))} = t^{-1/c}$$

i.e., if and only if $F(\log(\cdot))$ belongs to the Fréchet domain of attraction. He proved also that for $k_n = [n^\delta]$, $0 < \delta < 1$, ([ ] denoting the integer part), the limits are almost sure.

This provides a suitable indicator of the models described by (1.1). This motivated us to find out what happens if (Ac) fails. Specifically, if in some area related to problems of maximum values we want to identify a phenomenon, what can we say if (Ac) is not true? So, by studying the behavior of T_n for a wide range of standard random variables such as Gaussian, lognormal, exponential law, etc., we can derive new decision tools such as statistical tests and comparison methods. Illustrations of such applications and the results of simulations are given in Section 3.

In this paper, we treat the behavior of T_n in the case where $F \in D(\Lambda)$ or $F(\log(\cdot)) \in D(\Lambda)$, here $D(\Lambda)$ denotes the domain of attraction of Gumbel's law:

$$\Lambda(x) = \exp(-e^{-x}), \qquad \forall x, \quad -\infty < x < +\infty.$$

To begin with, we define

$$A = \inf\{x, F(x) = 1\}; \quad B = \sup\{x, F(x) = 0\},$$

$$F^{-1}(s) = Q(s) = \inf\{x, F(x) \geqq s\}, \qquad 0 < s < 1,$$

$$R(t) = (1 - F(t))^{-1} \int_t^A (1 - F(v))dv, \qquad B < t < A.$$

We shall, when appropriate, make the following assumptions:

$$\text{(H1)} \qquad F \in D(\Lambda), \text{ and } F(x) \text{ is strictly increasing as } x \uparrow A;$$

$$\text{(H2)} \qquad F(x) \text{ is continuous for } x \in (A, B).$$

If (H1) holds, we have the following (see e.g. Lemma 3):
(1.2) there exist some constant c_0 and a positive function $s(\cdot)$ such that

$$Q(1-u)=c_0+s(u)+\int_u^1 (s(t)/t)dt,$$

where $s(\cdot)$ is a function slowly varying at 0 (SVZ).
Therefore, we can make the following assumption:
(H3) (H1) and (1.2) hold and $s(u)$ is ultimately non-increasing as $u \downarrow 0$.
Our main results are as follows.

Theorem 1. Let

(H4) $\quad F(\log(\cdot))\in D(\Lambda)$ and $F(x)$ is ultimately continuous as $x \uparrow A$

be satisfied. Then for any sequence $k_n = k$ satisfying (K), we have

$$k^{-1}\sum_{i=1}^{i=k}(X_{n-i+1,n}-X_{n-k,n})\xrightarrow{P} 0, \qquad \text{as } n\to+\infty.$$

Theorem 2. Let (H1) be satisfied. Let (1.2) be reduced to

$$Q(1-u)=c_0+\int_u^1 \frac{s(t)}{t}\,dt, \tag{1.2'}$$

then for any sequence satisfying (K), we have

$$k^{1/2}\left\{\left\{k^{-1}\sum_{i=1}^{i=k} i\cdot C_{i,n}\cdot(X_{n-i+1,n}-X_{n-i,n})\right\}-1-\alpha_n\right\}\xrightarrow{d} \mathrm{N}(0,1), \tag{1.3}$$

where

$$\alpha_n\xrightarrow{P} 0,\quad C_{i,n}^{-1}=s(i/n), \qquad i=1,\cdots,k$$

and $\xrightarrow{d}$ denotes the convergence in distribution.

Corollary 1. Let the assumptions of Theorem 2 and (H3) be satisfied. If k satisfies

$$s(k/n)/s(1/n)\to 1, \qquad \text{as } n\to+\infty, \tag{K1}$$

we have

$$C_{k,n}\cdot k^{-1}\sum_{i=1}^{i=k}(X_{n-i+1,n}-X_{n-k,n})\xrightarrow{P} 1, \quad \text{as } n\to+\infty.$$

Corollary 2. Let (H2) and (H4) be satisfied. Let $k=(n^{\delta})$, $0<\delta<1$. If for some ν, $0<\nu<\delta/2$, we have

$$n^{\nu}s(k/n)\to+\infty, \quad \text{as } n\to\infty \tag{K2}$$

then we get

$$(1.4) \qquad C_{k,n} \cdot k^{-1} \sum_{i=1}^{i=k} (X_{n-i+1,n} - X_{n-k,n}) \to 1, \quad \text{almost surely.} \quad \text{(a.s.)}$$

Corollary 3. Let $A > 0$ and (H2) be satisfied. Then if (H1) and (1.2′) (or (H4)) are satisfied, the statement (1.3) (or (1.4)) remains true if we replace

$$X_{n-i+1,n} \text{ by } \log X_{n-i+1,n}, \qquad i = 1, \cdots, k+1$$

$$s(u) \text{ by } r(u) \sim \{R(Q(1-u))/Q(1-u)\} \text{ as } u \downarrow 0,$$

$$C_{i,n} \text{ by } r(i/n),$$

and if k satisfies (K) (or (K) and (K2) with $s(\cdot)$ replaced by $r(\cdot)$). Moreover if $\log A > 0$, we may derive similar results for the second logarithm, etc.

Remark. If $A > 0$, we have for large values of n

$$0 < X_{n-i+1,n} < A, \quad 1 \leqq i \leqq k_n + 1, \quad \text{since } k_n/n \to 0 \quad \text{and} \quad k_n \to +\infty.$$

Corollary 3 may be inverted as follows.

Corollary 4. Let $R(t) \to 0$ as $t \uparrow A$ and let (H2) be satisfied. Then if (H1) and (1.2′) (or (H4)) are satisfied, the statement (1.3) (or (1.4)) remains true if we replace

$$X_{n-i+1,n} \text{ by } \exp(X_{n-i+1,n})$$

$$s(u) \text{ by } t(u) \sim \{R(Q(1-u))\exp(Q(1-u))\} \text{ as } u \downarrow 0$$

$$C_{i,n} \text{ by } t(i/n)$$

and if k satisfies (K) (or (K) and (K2) with $s(\cdot)$ replaced by $t(\cdot)$). Moreover, if $t(u) \to 0$ as $u \downarrow 0$, we may repeat the operation, etc.

Now, we give some examples via the expressions of the norming constants $C_{k,n}$.

Corollary 5 (particular cases). In each case, (i) (or (ii)) will correspond to the choice of $k_n = (\log n)$ (or $k_n = (n^\delta)$, $0 < \delta < 1$).

1. *Normal case*: $X \sim \mathrm{N}(0,1)$

(i) $(2\log n)^{1/2} T_n \xrightarrow{P} 1$,

(ii) $(2(1-\delta)\log n) T_n \to 1$, a.s.

2. *Exponential (or gamma case)*: $\exp(X) \sim \mathrm{E}(1)$

(i) $(\log n) T_n \xrightarrow{P} 1$,

(ii) $((1-\delta)\log n) T_n \to 1$, a.s.

3. *Log_p-normal*: $X = \log_p \sup(b, Z)$, $Z \sim \mathrm{N}(0,1)$, $\log_p$ stands for the pth log and $\log_p b = 0$: Let $D_n = (2\log(k_n/n)) \Pi_{j=0}^{j=p-1} \log_j (2\log(k_n/n))^{1/2}$, $\log_0 x = 1$, $\forall x$.

Then

(i) $D_n T_n \xrightarrow{P} 1$,

(ii) $D_n T_n \to 1$, a.s.

4. Let T_{n1} (or T_{n2}) denote T_n for $X = \log\sup(0, Z)$, $Z \sim N(0,1)$ (or $Z \sim E(1)$), then

(i) $T_{n2}/T_{n1} \to 2$, in probability,

(ii) $T_{n2}/T_{n1} \to 2$, a.s.

2. Proofs of theorems and corollaries

Before we proceed any further, we give two useful lemmas.

Lemma 1. Let $F \in D(\Lambda)$, then

$$(k(t') - k(t) \to a) \Leftrightarrow \left(\frac{t'-t}{R(t)} \to a\right) \quad \text{as } t', t \uparrow A$$

for any $-\infty \leqq a \leqq +\infty$, where $k(t) = -\log(1 - F(t))$.

Proof of Lemma 1. It is well known that $F \in D(\Lambda)$ (see e.g. De Haan (1970)) iff

$$\text{(B)} \qquad (\forall x) \quad \frac{1 - F(t + xR(t))}{1 - F(t)} \to \exp(-x), \quad \text{as } t \uparrow A$$

which may be written

$$\text{(B)} \qquad (\forall x) \quad k(t + xR(t)) - k(t) \to x, \quad \text{as } t \uparrow A.$$

(i) Suppose that $\lim_{n\to\infty} (t'_n - t_n)/R(t_n) \geqq a$, where a is finite and (t'_n, t_n) is a subsequence extracted from (t', t) and $t', t \uparrow A$. Then for any $\varepsilon > 0$, we have for large values of n: $t'_n \geqq t_n + (a - \varepsilon)R(t_n)$. The fact that $k(\cdot)$ is non-decreasing and (B) imply that $\lim_{n\to\infty} \inf k(t'_n) - k(t_n) \geqq a - \varepsilon$. Hence

$$\lim_{n\to\infty} \inf k(t'_n) - k(t_n) \geqq a.$$

(ii) In the same manner, we prove that

$$\left\{\lim_{n\to\infty} (t'_n - t_n)/R(t_n) \leqq a\right\} \Rightarrow \left\{\lim_{n\to\infty} \sup k(t'_n) - k(t_n) \leqq a\right\}.$$

Now, suppose that $k(t') - k(t) \to a$, $-\infty \leqq a \leqq +\infty$.

(a) Let a be finite: by (i) and (ii) it follows that for any sequence (t'_n, t_n) such that $t'_n, t_n \uparrow A$, we have necessarily that $(t'_n - t_n)/R(t_n)$ is bounded. Furthermore if $(t'_n - t_n)/R(t_n) \to d$, where d is finite, the same points (i) and (ii) imply that $d = a$.

(b) Let $a = +\infty$. Suppose that there exists a sequence (t'_n, t_n), $t'_n, t_n \uparrow A$ such that $(t'_n - t_n)/R(t_n)$ is bounded. (ii) would imply that $k(t'_n) - k(t_n)$ is also bounded, which is not possible at the same time as $a = +\infty$.

(c) Let $a = -\infty$. We use (i) at the place of (ii) in the above case and get that $(t' - t)/R(t) \to -\infty$.

So, we have proved that

$$(k(t') - k(t) \to a) \Rightarrow ((t' - t)/R(t) \to a) \quad \text{at } t', t \uparrow A.$$

Conversely, suppose that $(t' - t)/R(t) \to a$.

(α) Let a be finite. Then for any $\varepsilon > 0$, we have for t', t near A, $t + (a - \varepsilon)R(t) \leqq t' \leqq t + (a + \varepsilon)R(t)$. Therefore, (B) implies

$$a - \varepsilon \leqq \liminf_{t',t \uparrow A} k(t') - k(t) \leqq \limsup_{t',t \uparrow A} k(t') - k(t) \leqq a + \varepsilon.$$

(β) Let $a = +\infty$, for any $d > 0$, we have for t', t near $A : t' \geqq t + dR(t)$. Therefore (B) implies

$$\liminf_{t',t \uparrow A} k(t') - k(t) \geqq d.$$

(γ) Let $a = -\infty$; similarly to the preceding case, we get

$$\limsup_{t',t \uparrow A} k(t') - k(t) \leqq -d.$$

By letting $\varepsilon \to 0$, and $d \to +\infty$, we get the other implication of the equivalence we had to prove.

Lemma 2. Let $F \in D(\Lambda)$. If in addition $F(x)$ is continuous for x near A, then $F^{-1}(1-u) = Q(1-u)$ is slowly varying at 0 (SVZ).

Proof of Lemma 2. Let $t' = Q(1-u)$ and $t = Q(1-uv)$, with v fixed and $v > 0$. Because of the continuity of $F(\cdot)$, we have $u = 1 - F(t') = \exp(-k(t'))$ and $uv = 1 - F(t) = \exp(-k(t))$. Hence $k(t') - k(t) = \log(1/v)$. Lemma 1 implies then $(t' - t)/R(t) \to \log(1/v)$, which in turn implies that

$$\frac{Q(1-u)}{Q(1-uv)} = 1 + (\log(1/v) + o(1)) \frac{R(Q(1-u))}{Q(1-u)}.$$

But, by Lemma 5, $R(t)/t \to 0$ as $t \uparrow A$, whenever $F \in D(\Lambda)$. Hence $\lim_{u \downarrow 0} Q(1-u)/Q(1-uv) = 1$, which is the announced result.

Proof of Theorem 1. Since (H5) holds, Lemma 2 implies that the function $H(1-u)$ associated with $F(\log(\cdot))$ is SVZ.

Note that $Q(1-u) = \log H(1-u)$ as $u \downarrow 0$. Recall also the well-known representations

$$\{X_{i,n}, 1 \leqq i \leqq n\} \overset{d}{=} \{Q(U_{i,n}), 1 \leqq i \leqq n\} \tag{2.1}$$
$$\{U_{i,n}, 1 \leqq i \leqq n\} \overset{d}{=} \{1 - U_{n-i+1,n}, 1 \leqq i \leqq n\}$$

where $U_{1,n} \leqq U_{2,n} \leqq \cdots \leqq U_{n,n}$ denote the order statistics associated with $U_1, U_2, \cdots, U_n$, a sequence independent and uniformly distributed on $(0,1)$. Finally, recall that $H(1-u)$ admits Karamata's representation since it is SVZ:

$$H(1-u) = c(u)\exp\left(\int_u^1 (r(s)/s)ds\right), \tag{2.2}$$
$$c(u) \to c, \quad 0 < c < +\infty, \quad r(u) \to 0 \quad \text{as } u \downarrow 0.$$

Combining (2.1) and (2.2), we have for each i, $1 \leqq i \leqq k$,

$$X_{n-i+1,n} - X_{n-k,n} \overset{d}{=} \log\{c(U_{i,n})/c(U_{k,n})\} + \int_{U_{i,n}}^{U_{k,n}} (r(s)/s)ds = A_{i,n}. \tag{2.3}$$

Remark that $0 < U_{i,n} \leqq U_{k,n} \overset{P}{\to} 0$, if $1 \leqq i \leqq k$, $k/n \to 0$. It follows that for any $c < \varepsilon < 0$, for large values of n, $(\forall i, 1 \leqq i \leqq k, c - \varepsilon \leqq c(U_{i,n}) \leqq c + \varepsilon)$ holds with probability greater than ε. So, for large n, we have, with probability greater than ε,

$$A_{i,n} \leqq \log\left(\frac{c+\varepsilon}{c-\varepsilon}\right) + \sup_{0 \leqq s \leqq U_{k,n}} |r(u)| \cdot \log(U_{k,n}/U_{i,n}) = B_{i,n}.$$

But statements (5), (6) and (7) in Mason (1982) yield

$$k_n^{-1} \sum_{i=1}^{i=k} \log(U_{k_n,n}/U_{i,n}) \overset{d}{=} k_n^{-1} \sum_{i=1}^{i=k} \xi_i, \tag{2.4}$$

where $\xi_1, \xi_2, \cdots, \xi_n$ is a sequence of independent and standard exponential random variables. It follows then that for large n, we have, with probability greater than ε,

$$T_n \overset{d}{=} k^{-1} \sum_{i=1}^{i=k} A_{i,n} \leqq k^{-1} \sum_{i=1}^{i=k} B_{i,n} \overset{d}{=} \log\frac{c+\varepsilon}{c-\varepsilon} + \sup_{0 \leqq s \leqq U_{k,n}} |r(s)| \left(\frac{1}{k} \sum_{i=1}^{i=k} \xi_i\right).$$

Now, by applying the strong law of large numbers to the average of ξ_i, $i = 1, 2, \cdots, k$ and by using the fact that $r(u) \to 0$ as $u \downarrow 0$ we may conclude from the last inequality that if we let $\varepsilon \to 0$, we get $T_n \overset{P}{\to} 0$.

Proof of Theorem 2. To begin with, we need three lemmas.

Lemma 3. Let F be strictly increasing as $x \uparrow A$. Then the following assertions are equivalent:

(i) $F \in D(\Lambda)$.

(ii) There exist some constant c_0 and a positive function $s(\cdot)$ SVZ such that

$$Q(1-u) = c_0 + s(u) + \int_u^1 (s(t)/t)dt.$$

Proof of Lemma 3. The proof follows from Theorem 1.4.1 and Theorem 2.4.1 of De Haan (1970).

Lemma 4. Let the assertion (ii) of Lemma 3 be satisfied, then $F \in D(\Lambda)$. If in addition F is continuous as $x \uparrow A$, we get

$$\lim_{u \downarrow 0} \frac{s(u)}{R(Q(1-u))} = 1.$$

Proof of Lemma 4. The proof is given in Lo (1986), Lemma 4.

Lemma 5. Let (H1) and (H2) be satisfied, then $R(t)/t \to 0$ as $t \uparrow A$ and $R(Q(1-u))$ is SVZ.

Proof of Lemma 5. First, we remark that Lemma 4 implies that $R(Q(1-u))$ is SVZ and by Lemma 3 of Lo (1986), $R(t)/t \to 0$ as $t \uparrow A$.

Proof of Theorem 2 (continued). Since (H1) is satisfied, suppose that (1.2) is reduced to

$$Q(1-u) = c_0 + \int_u^1 (s(t)/t)dt.$$

Thus, for each i, $1 \leqq i \leqq k$, k satisfying (K), one has

$$X_{n-i+1,n} - X_{n-i,n} \overset{\mathrm{d}}{=} Q(1-U_{i,n}) - Q(1-U_{i+1,n}) = \int_{U_{i,n}}^{U_{i+1,n}} (s(u)/u)du. \tag{2.5}$$

Now, since $s(u)$ is SVZ, it admits the Karamata representation:

$$\begin{gathered} s(u) = z(u)\exp\left(\int_u^1 (w(v)/v)dv\right), \\ z(u) \to z, \quad 0 < z < +\infty, \quad w(u) \to 0, \quad \text{as } u \downarrow 0. \end{gathered} \tag{2.6}$$

So, for each i, $1 \leqq i \leqq k$, for any $a_{i,n}$, $U_{i,n} \leqq a_{i,n} \leqq U_{i+1,n}$, we have for large n,

$$\{s(a_{i,n})/s(i/n)\} = \{z(a_{i,n})/z(i/n)\}\exp\left(-\int_{i/n}^{a_{i,n}} (w(v)/v)dv\right). \tag{2.7}$$

Also

$$\begin{aligned} \left|\int_{i/n}^{a_{i,n}} \frac{w(v)}{v}dv\right| &\leqq \sup_{v \in I_n} |w(v)| \cdot \sup_{1 \leqq i \leqq n} \max\left(\left|\log \frac{n}{i} U_{i,n}\right|, \left|\log \frac{n}{i} U_{i+1,n}\right|\right) \\ &=: R_n^1 \cdot R_n^2 \end{aligned} \tag{2.8}$$

where $I_n = (0, \max(k/n, U_{k+1,n}))$ is a random interval.

However, it is not difficult to see that the sequence R_n^2 is bounded in probability (see e.g. Lemma 13 of Csörgő and Mason (1984)). Furthermore, since $w(u) \to 0$ as $u \downarrow 0$, it follows that $\sup\{w(v), v \in I_n\} \overset{\mathrm{P}}{\to} 0$ independently of $a_{i,n}$ and independently of i, $1 \leqq i \leqq k$. Moreover, since $a_{i,n} \leqq U_{i+1,n} \leqq U_{k+1,n} \xrightarrow{P} 0$, we get also that $\{z(a_{i,n})/z(i/n)\} \xrightarrow{P} 1$, uniformly with respect to $a_{i,n}$ and with respect to i, $1 \leqq i \leqq k$. Therefore, we can see that the following equalities are independent of $a_{i,n}$ and of i, $1 \leqq i \leqq k$:

$$(2.9) \qquad \int_{i/n}^{a_{i,n}} \frac{w(v)}{v}\, dv = o_p(1), \quad \{z(a_{i,n})/z(i/n)\} = 1 + o_p(1).$$

Applying (2.9) to (2.7), and recalling that $s(\cdot)$ is positive, we get

$$(2.10) \qquad 1 + \beta_{ni}^1 = \inf_{v \in J_n} \frac{s(v)}{s(i/n)} \leqq \sup_{v \in J_n} \frac{s(v)}{s(i/n)} = 1 + \beta_{ni}^2, \quad \text{as } n \uparrow +\infty$$

where $J_n = (U_{i,n}, U_{i+1,n})$ and $\beta_{ni}^j = o_p(1)$, $j = 1, 2$, independently of i, $1 \leqq i \leqq k$. Apply (2.10) in an appropriate manner to (2.5) and get

$$(2.11) \qquad iC_{i,n}(X_{n-i+1,n} - X_{n-i,n}) = \log\left(\frac{U_{i+1,n}}{U_{i,n}}\right)^i (1 + \gamma_{ni})$$

where $\gamma_{ni} = o_p(1)$ independently of i, $1 \leqq i \leqq k$, k satisfying (K). However, it is also well known that

$$\{\log(U_{i+1,n}/U_{i,n}), 1 \leqq i \leqq n\} \stackrel{d}{=} \{\xi_i/i, 1 \leqq i \leqq n\} \quad \text{with } U_{n+1,n} = 1.$$

Applying this and again applying (2.9), we get

$$iC_{i,n}(X_{n-i+1,n} - X_{n-i,n}) \stackrel{d}{=} \xi_i(1 + \gamma_{ni}).$$

It follows that

$$V_n = k^{-1} \sum_{i=1}^{i=k} iC_{i,n}(X_{n-i+1,n} - X_{n-i,n}) \stackrel{d}{=} k^{-1} \sum_{i=1}^{i=k} \xi_i + \alpha_n.$$

It is obvious that $\alpha_n \xrightarrow{P} 0$, because of (2.11). Therefore, the weak law of large numbers implies that $V_n \xrightarrow{P} 1$. Moreover the central limit theorem gives $k^{1/2}(V_n - 1 - \alpha_n) = k^{1/2}(k^{-1}\sum_{i=1}^{i=k} \xi_i - 1) \xrightarrow{d} \mathrm{N}(0, 1)$.

Proof of Corollary 1. This corollary is immediate since $V_n \xrightarrow{P} 1$ and

$$k^{-1} \sum_{i=1}^{i=k} i(X_{n-i+1,n} - X_{n-i,n}) = k^{-1} \sum_{i=1}^{i=k} (X_{n-i+1,n} - X_{n-k,n}).$$

Proof of Corollary 2. If (H4) is satisfied, Theorem 1 implies that $T_n \xrightarrow{P} 0$. On the other hand, Mason (1982) has proved that if $T_n = T_n(\delta)$, where $k = k_n = (n^\delta)$, is bounded in probability, then

$$T_n(\delta) - R(Q(1 - U_{k,n})) = O(n^{-\nu}) \quad \text{a.s., } 0 < \nu < \delta/2,$$

where $\{T_2^*(\delta), n \geqq 1\} \stackrel{d}{=} \{T_n(\delta), n \geqq 1\}$ (as processes).

By Lemma 7, (H4) and (H2) $\Rightarrow$ (H1). Hence, by Lemma 4,

$R(Q(1-u))/s(u) \to 1$, as $u \downarrow 0$. Now, since $s(u)$ is SVZ and $(n/k)U_{k,n} \to 1$, a.s., we get by Lemma 5

$$R(Q(1-U_{k,n}))/s(k/n) \to 1, \quad \text{a.s.}$$

which completes the proof.

Proof of Corollaries 3 *and* 4. These two corollaries will follow from the lemmas below. Define by L the set of distribution functions F satisfying (H1) and (H2).

Lemma 7. Let $A > 0$. Then if X has a distribution function $F \in L$, $\log \sup(0, X)$ has a distribution function $G \in L$ and

$$S(G^{-1}(1-u)) \sim \{R(Q(1-u))/Q(1-u)\}, \quad \text{as } u \downarrow 0,$$

where

$$S(t) = \int_t^{\log A} \frac{(1-G(v))}{(1-G(t))} \, dv, \qquad -\infty < t < \log A.$$

Conversely, we have the following.

Lemma 8. Let $R(t) \to 0$ as $t \uparrow A$. Then if X has a distribution function $F \in L$, then $\exp(X)$ has a distribution function $Z \in L$ and

$$R(Q(1-u)) \sim \{T(Z^{-1}(1-u))/Z^{-1}(1-u)\} \quad \text{as } u \downarrow 0,$$

where

$$T(t) = \int_t^{e^A} \frac{1-Z(v)}{1-Z(t)} \, dv, \qquad e^B < t < e^A.$$

Proof of Lemmas 7 *and* 8. These lemmas are proved in Lo (1986), via Lemmas 9 and 10. Note that (H2) implies that $\log \sup(0, X)$ exists almost surely if $A > 0$.

Proof of Corollaries 3 *and* 4. By Lemma 7, we see that if for instance (H1) is satisfied, the same property is also true for $\log \sup(0, X)$. So, we can write (1.3) with $s(\cdot)$ replaced by $r(\cdot)$, where $r(u)$ is derived from De Haan's representation for $G^{-1}(1-u)$ as in (1.2). But, we see also that $Q(1-u) = c_0 + \int_u^1 (s(t)/t)dt \Rightarrow s(u) = uQ'(1-u)$. Since

$$G^{-1}(1-u) = \log Q(1-au), \quad a = P(X > 0),$$

$$r(u) = \frac{uQ'(1-u)}{Q(1-u)} = u(G^{-1}(1-u)')$$

is SVZ by Lemma 2 and (1.2). Hence,

$$G^{-1}(1-u) = -G^{-1}(1-y_1) + \int_u^1 \frac{r(t)}{t} \, dt, \qquad \text{for } u \leqq y_1 \leqq 1.$$

The preceding means that the De Haan representation of $G^{-1}(1-\cdot)$ is reduced. We may apply Theorem 2 for $G(\cdot)$. Note that $\sup(X_{n-i+1,n}, 0) = X_{n-i+1,n}$, a.s., as $k \to \infty$, $n \to \infty$, $k/n \to 0$. Corollary 4 is proved in a similar way.

Proof of Corollary 5. It may be easily verified that in all our particular cases, we have that $l(u) = uQ'(1-u)$ is slowly varying at 0, where $Q'(u)$ is the derivative function of Q for u near 1. So, for some y near 0, $0 < y < 1$, we have

$$Q(1-u) = -Q(1-y) + \int_u^y (l(t)/t)dt, \qquad 0 < u < y. \tag{2.12}$$

That means that (1.2) is reduced. As remarked in the proof of Theorem 2, (2.12) may entirely replace (1.2). Moreover, if (2.12) holds with a function SVZ, it follows that $F \in D(\Lambda)$ and $l(u) \sim R(Q(1-u))$ (see Lo (1986), Lemma 4). Note also that $Q(1-u)$ is continuous as $u \downarrow 0$ whenever (2.12) holds.

By the above explanations, the proof of Corollary 5 consists in determining the function $l(u)$. The verification of (K1) or (K2) is immediate.

1. *Normal case*: $X \sim \mathrm{N}(0,1)$. We use the well-known expansion of the tail of the standard Gaussian random variable:

$$\exists x_0, \forall x \geqq x_0, \left\{\frac{1}{x} - \frac{1}{x^3}\right\} M \exp(-x^2/2) < 1 - F(x) < \frac{M}{x}\exp(-x^2/2), \; M = (2\pi)^{-1/2}.$$

Using that, one verifies that $W(x) = (1-F(x))/F'(x)$ has a negative derivative for $x \geqq x_0$. Then $uQ'(1-u) = (1-F(x))/(F'(x))$ is strictly decreasing as $u = 1 - F(x) \downarrow 0$. This shows that (H3) is satisfied. Moreover, it is also known that

$$Q(1-u) = \left(2\log\frac{1}{s}\right)^{1/2}\left\{1 + \frac{\log\log(1/s) - 4\pi + o(1)}{4\cdot\log(1/s)}\right\}, \quad \text{as } s \downarrow 0. \tag{2.13}$$

Routine calculations show that

$$Q'(1-s) = \frac{1}{s(\log(1/s))^{1/2}}(1+o(1)) \quad \text{as } s \downarrow 0.$$

So $l(u) = (2\log(1/s))^{-1/2}(1+o(1))$.

At this step, an appropriate application of Corollaries 1 and 2 gives the stated results.

2. *Exponential case*: $\exp(X) \sim \mathrm{E}(1)$. Here $Q(1-s) = \log\log(1/s)$; therefore, we apply Corollaries 1 and 2. Remark that for a general gamma law, $Q(1-s) \sim \log\log(1/s)$.

3. This part follows from a typical application of Corollary 3 p times.

4. This part follows from Part 2 and Part 3 in the case where $p = 1$.

3. Applications and simulations

As remarked above, Hill (1974) described some basic models which follow (1.1) or (Ac). We note that all these models are closely related to problems based on extreme values. As already noticed, the works of several authors (Csörgő and Mason (1985), Csörgő et al. (1985), Hall (1982), Hill (1970), (1974), (1975), Mason (1982)) have entirely settled the properties of T_n under the assumption (Ac). It follows from their results that if $X_1, X_2, \cdots, X_n$ are the observations of X, we can verify if (Ac) holds. In that case, we proceed as follows:

3.1. *Identification of the upper tail of a distribution.*

(i) Choose δ, $0 < \delta < 1$,

(ii) Choose $k_n = (n^\delta)$,

(iii) Calculate $T_n = k^{-1} \Sigma_{i=1}^{i=k} (X_{n-i+1,n} - X_{n-k,n})$ for large values of n.

(iv) If T_n is very near c, $0 < c < +\infty$, then by Theorem 1 of Mason (1982), (Ac) holds.

(v) But (Ac) $\Rightarrow F \in D(\Lambda) \Rightarrow c^{-1}(X_{n,n} - Q(1-1/n)) \xrightarrow{d} \Lambda$,

and therefore, we can use (v) for predictions about a critical value of $X_{n,n}$. However, it is not always certain that T_n converges to a finite strictly positive number. For example, if we want to know whether F satisfies (Ac) or if $F(\log(\cdot))$ is the distribution function of the standard Gaussian random variable, how could we proceed?

3.2. *Comparison between a regular tail and a gaussian tail.* We want to know if

(L1) $F(\log(x)) = x^{-1/c}(1 + DO(x^{-b}))$, $c > 0$, and $b = 1/2c$

or

($\bar{\text{L}}\bar{1}$) $X_1 = \log \sup(0, Z)$, $Z \sim \text{N}(0, 1)$

where $F(\cdot)$ is the unknown distribution function associated with the observations $X_1, X_2, \cdots, X_n$ of X.

(i) Choose $k = (n^{1/2})$, then

(ii) If (L1) holds, we have

  (a) (Mason (1982)) $T_n \to c$, a.s.

  (b) (Hall (1982)) $n^{1/4}(T_n - c) \xrightarrow{d} \text{N}(0, 1)$.

(iii) If ($\bar{\text{L}}\bar{1}$) holds, we have

  (a) (Corollary 5, Part 3, $p = 1$), $\log nT_n \to 1$, a.s.

  (b) (Lo (1986)) $n^{1/4}(D_nT_n - 1) \xrightarrow{d} \text{N}(0, 1)$, $D_n = \log n(1 + o(1))$.

Thus, we see that we are now able to test (L1) against ($\bar{\text{L}}\bar{1}$). If we choose $D = \{(c - \varepsilon) \leqq T_n \leqq (c + \varepsilon)\}$ as the accepted region of our test, (i) and (ii) give the characteristics of that test. To test ($\bar{\text{L}}\bar{1}$) against (L1), one can choose $\bar{D} = \{1 - \varepsilon \leqq T_n \cdot \log n \leqq 1 + \varepsilon\}$ as the accepted region with a small value of ε.

3.3. *Comparison between an exponential and a gaussian tail.* Let

$X_1, X_2, \cdots, X_n$, $Y_1, Y_2, \cdots, Y_n$ be two independent samples respectively associated with the distribution functions F and G. We want to know if

(L2) $F(\log(x)) = 1 - \exp(-x) = G(\log(x))$

or

($\bar{\text{L}}\bar{2}$) $X_1 = \log\sup(0, Z)$, $Z \sim \text{N}(0,1)$, $G(\log(x)) = 1 - \exp(-x)$.

In Lo (1986), we have also given the limiting law of T_n under the assumption $F(\log(x)) = 1 - \exp(-x)$. Denote by T_{n1} (T_{n2}) the Hill's estimate associated with $X_1, X_2, \cdots, X_n$ ($Y_1, Y_2, \cdots, Y_n$). Then from Corollary 5, we get $T_{n1}/T_{n2} \to 1/2$, a.s. under ($\bar{\text{L}}\bar{2}$) and $T_{n1}/T_{n2} \to 1$, a.s. under (L2). The independence of the two samples and the limiting laws obtained in Lo (1986) thus enable us to construct statistical tests. After the simulations, numerical applications of that test will be given.

Finally, we give numerical applications using simulations.

3.4. *Simulations.* Here we have used an ordered sample generated from a uniform random variable $u_1, u_2, \cdots, u_{4000}$. We have constructed the following order statistics:

Exponential case: $y_i = -\log(1 - u_i)$.

Normal case: $x_i = \left(2\log\left(\frac{1}{1-u_i}\right)\right)^{1/2}$, for large values of i, $1 \leqq i \leqq 4000$.

Pareto case: $z_i = (1 - u_i)^{-1}$.

Define for $k = k_n = (n^{1/2})$,

$$T_{n1} = k^{-1} \sum_{i=1}^{i=k} (\log x_{n-i+1} - \log x_{n-k}), \qquad 3990 \leqq n \leqq 4000.$$

We write $T_{n1} = B_n(x_i)$. Therefore, we define $T_{n2} = B_n(y_i)$ and $T_{n3} = B_n(z_i)$.

Our simulations are given in Table 1.

TABLE 1

1	2	3	4	5
N	$\log nT_{n1}$	$\frac{1}{2}\log nT_{n2}$	T_{n3}	$u_{4000-n+1}$
3991	0.5222	0.5222	0.6780	0.002435
3992	0.6090	0.6090	0.6988	0.001631
3993	0.6225	0.6225	0.7215	0.001620
3994	0.6443	0.6443	0.7035	0.001337
3995	0.6204	0.6204	0.7162	0.000988
3996	0.6464	0.6464	0.7460	0.000437
3997	0.6683	0.6683	0.7182	0.000418
3998	0.6898	0.6898	0.8102	0.000308
3999	0.7093	0.7093	0.8412	0.000297
4000	0.7554	0.7554	0.9086	0.000095

Remarks.

1. One might be surprised to find that our simulations are not sufficiently good, considering the large size of the sample space ($n = 4000$). However, only the k extreme observations ($k = 62$ or 63) are used for the calculation of T_n. Taking that into account, the theoretical part of this paper is relatively well illustrated by the simulations. Specifically:

2. Column 4 illustrates the almost sure convergence of Hill's estimate for the Pareto law: T_{n3} converges almost surely to 1.

3. The identity of columns 2 and 3 is a consequence of the choice of x_i. It is clear that if $x_i = (-2\log(1-u_i))^{1/2}$, we get $T_{n2} = \frac{1}{2}T_{n1}$. However, this choice is not arbitrary. Indeed, if T_{n2} is the true value obtained from the use of the true quantile function, we have $T_{n2}^* = B_n(t_i)$, where (see e.g. statement (2.13)).

$$\log t_i = \log\left\{(-2\log(1-u_i))^{1/2}\left(1 + \frac{\log(-\log(1-u_i) - 4\pi + o(1))}{4\log(1-u_i)}\right)\right\}$$

$$= \tfrac{1}{2}\log 2 + \tfrac{1}{2}\log y_i + O\left(\left(\log\log\frac{1}{1-u_i}\right)\Big/ 4\log\frac{1}{1-u_i}\right).$$

With our data, $3937 \leqq i \leqq 4000$, $k = 62$ or 63, we have

$$\log t_i = \tfrac{1}{2}\log 2 + \tfrac{1}{2}\log y_i \pm 0.07,$$

therefore

$$T_{n2}^* = \tfrac{1}{2}T_{n1} \pm 0.07.$$

4. We now test (L3): (x_i) are the order statistics of a N(0, 1) random variable, against (L4): (x_i) are the order statistics of the standard exponential law. If we choose $R_n = \{1 - a \leqq \log nT_{n1} \leqq 1 + a\}$ as our accepted region and choose $\alpha = 0.29$ as the significance level of our test, the power of the test will be $\beta \sim 0.74$, and R_n will be $R_n = \{0.75 \leqq T_{n1} \cdot \log n \leqq 1.2533\}$. Here, we accept (L3) since the table gives the value 0.7554 for $(\log n \cdot T_{n1})$ for $n = 4000$.

5. Column 5 gives the ten first values of the order statistics of the uniform random variable. One may work with the highest or the lowest values since

$$\{1 - u_i, 1 \leqq i \leqq 4000\} \overset{d}{=} \{u_{4000-i+1}, 1 \leqq i \leqq 4000\}.$$

Conclusion. De Haan and Resnick (1980) and Csörgő et al. (1985) have also given estimates of c under the assumption (Ac). In future papers, we shall describe their asymptotic behavior using the assumptions of this paper.

5. Remarks and further generalizations

Remark 1. Deheuvels et al. (1986) have recently shown that the De Haan representation (1.2) holds whenever $F(\cdot)$ belongs to $D(\Lambda)$.

(This remark is derived from a discussion with Professor D. M. Mason and Dr. Haeusler. Many thanks to them.)

Remark 2. We have used the continuity of $F(\cdot)$ just to obtain that $F(F^{-1}(x)) = x$ for large x. But this is true for any distribution function.

Remark 3. We have proved in Lo (1986) (see Lemma 12′) that the assumption K2 in Corollary 2 is always satisfied.

These three simple remarks yield the following generalization.

Generalization. Throughout the paper, the hypothesis (H3) may be replaced by $F \in D(\Lambda)$, (H4) by $F(\log(\cdot)) \in D(\Lambda)$, and both (K1) and (K2) may be dropped.

Acknowledgments

I should like to thank Professor Deheuvels for having suggested to me the problem which led to this investigation, and for his moral support. I am also indebted to M. Der Mergreditchian for giving me permission to use the EERM's computers, to N. Nader, M. Davis and R. Moradkhan for their comments concerning the English text, and finally to the referee for his helpful suggestions.

References

BOULENGER, G. A. (1885) *Catalogue of the Lizards in the British Museum.* British Museum, London.

CSÖRGŐ, S. AND MASON, D. (1984) Central limit theorems for sums of extreme values. *Math. Proc. Camb. Phil. Soc.* **98**, 547–558.

CSÖRGŐ, S., DEHEUVELS, P. AND MASON, D. (1985) Kernel estimates of the tail index of a distribution. *Ann. Statist.* **13**, 1467–1487.

DE HAAN, L. (1970) *On Regular Variation and Applications to the Weak Convergence of Sample Extremes.* Mathematical Centre Tracts, **32**, Amsterdam.

DE HAAN, I. AND RESNICK, S. I. (1980) A simple asymptotic estimate for the index of a stable law. *J. R. Statist. Soc.* B **44**, 83–87.

DEHEUVELS, P., HAEUSLER, E. AND MASON, D. M. (1986) Laws of the iterated logarithm when the maximum is attracted to a Gumbel law. Unpublished.

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

HALL, P. (1982) On some simple estimates of an exponent of regular variation. *J. R. Statist. Soc.* B **44**, 37–42.

HILL, B. M. (1970) Zipf's law and prior distributions for the composition of a population. *J. Amer. Statist. Assoc.* **65**, 1220–1232.

HILL, B. M. (1974) The rank-frequency form of Zipf's law. *J. Amer. Statist. Assoc.* **69**, 1017–1026.

HILL, B. M. (1975) A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163–1174.

LO, G. S. (1986) Ph.D. Dissertation, University of Paris VI.

MASON, D. (1982) Law of large numbers for sums of extreme values. *Ann. Prob.* **10**, 754–764.