Monografías del Seminario Matemático García de Galdeano **31**, 331–336 (2004)

CENTRAL LIMIT THEOREMS FOR RECORDS

R. Gouet, F.J. López and G. Sanz

Abstract. Consider a sequence (X_n) of independent and identically distributed random variables, taking nonnegative integer values and call X_n a record if $X_n > \max\{X_1, \dots, X_{n-1}\}$. In Gouet et al. (2001), a martingale approach combined with asymptotic results for sums of partial minima was used to derive strong convergence results for the number of records among the first n observations. Now, in this paper we exploit the connection between records and martingales to establish a central limit theorem for the number of records in many discrete distributions, identifying the centering and scaling sequences.

Keywords: Extremes, Records, Martingales, Central Limit Theorem

AMS classification: AMS classification 60G70, 60G42

§1. Introduction

Let (X_n) be a sequence of nonnegative, independent and identically distributed (iid) random variables (rv's), with common distribution function F and let $M_n = \max\{X_1, \dots X_n\}, n \geq 1$ be the sequence of partial maxima; conventionally we write $M_0 = -1$. We say X_n is a (strict, upper) record if $X_n > M_{n-1}$, $n \geq 1$. The indicator of a record is denoted by $I_n = \mathbf{1}_{\{X_n > M_{n-1}\}}$ and the associated counting process by $N_n = \sum_{k=1}^{n} I_k$. General information on the theory of records can be found in [1]. We are interested here in the asymptotic normality of N_n, suitably centered and scaled, when the underlying distribution F is concentrated on the nonnegative integers.

A well known result of A. Renyi [6] states that the indicators I_n are independent, with $P[I_n = 1] = 1/n$, when F is continuous. Therefore, the central limit theorem (CLT)

$$\frac{N_n - \log n}{\sqrt{\log n}} \xrightarrow{d} N(0,1),$$

is readily obtained. When F is discontinuous the indicators I_n are not independent and their distributions depend on F. Therefore, this case is somewhat more complicated and results are rather scarce. W. Vervaat [7] obtains a variety of functional CLT's for records of nonnegative, integer valued random variables. In particular, his work contains the asymptotic normality of N_n for the geometric distribution.

In this paper we establish a central limit theorem for the number of records for a wide range of discrete distributions, identifying the centering and scaling sequences (Theorem 1 (a)) and we give a sketch of the proof. The whole proofs of the results can be checked in Gouet et al. [5].

We conclude this introduction with additional definitions and notation. Let (X_n) denote a sequence of iid rv's such that $P[X_n = k] = p_k > 0$, for $k \in \mathbb{Z}_+ = \{0, 1, \ldots\}$ and $n \geq 1$, with $\sum_{k \geq 0} p_k = 1$. Let $F(x) = P[X_n \leq x]$ be their common distribution function, $F(x^-) = P[X_n < x]$ the left limit function and $\tilde{F}(y) = \inf\{x \mid F(x) \geq y\}$ the (generalized) inverse of F, $x \geq 0, 0 \leq y \leq 1$. Clearly $\omega = \tilde{F}(1) = \infty$ and hence, $N_n \nearrow \infty$ a.s.

For $k \in \mathbb{Z}_+$, let $y_k = 1 - F(k) = \sum\limits_{i>k} p_i$ be the discrete survival function and define the discrete failure or hazard rate r_k by

$$r_k = \frac{P[X_1 = k]}{P[X_1 \geq k]} = \frac{p_k}{y_{k-1}}.$$

It is easily verified that $r_k = 1 - y_k/y_{k-1}$ and $y_k = \prod\limits_{i=0}^{k}(1 - r_i)$. Let also $\theta(k) = \sum_{i=0}^{k} r_i$ denote de cumulative hazard function and $m(t) = \min\{j \in \mathbb{Z}_+ | y_j < 1/t\}$ the quantile function, $k \in \mathbb{Z}_+, t > 0$.

Martingales are taken relative to the *natural* filtration $(\mathcal{F}_n)$, with $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, for $n \geq 1$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Convergence, almost sure, in probability and weak, will be denoted respectively by the arrows $\xrightarrow{a.s.}$, $\xrightarrow{P}$ and $\xrightarrow{d}$.

In Section 2 we state the main result (Theorem 1) and show some examples. In Section 3 we give a sketch of the proof of Theorem 1.

§2. Main result and examples

Our main result is the asymptotic normality of the counting process of records N_n suitably centered and scaled, applicable to a wide spectrum of discrete models. We use a martingale approach which connects the central limit theorem with convergence results from the theory of sums of partial minima of iid rv's, as developed by P. Deheuvels in [3].

Theorem 1. *Let* $z_k = \sum_{i>k} r_i y_i$ *and* $b_n^2 = \sum_{k=0}^{m(n)} z_k r_k / y_k$, *for* $k, n \in \mathbb{Z}_+$.

(a) Assume $\sum_{k=0}^{\infty}(1 - r_k) = \infty$. *If* $\limsup r_k < 1$ *or* $\liminf r_k > 0$, *then*

$$\frac{N_n - \theta(m(n))}{b_n} \xrightarrow{d} N(0, 1).$$

(b) If $\sum_{k=0}^{\infty}(1 - r_k) < \infty$, *then* $N_n - m(n)$ *is tight. In particular, there are no sequences* $(a_n), (b_n) \nearrow \infty$ *such that* $(N_n - a_n)/b_n$ *converges in distribution to a non-degenerate random variable.*

Proof. See [5] □

Remark 1. Theorem 1 gives a rather complete picture of the asymptotic normality of the number of records for discrete distributions. In fact, any sequence (r_k), $0 < r_k < 1$, $k \geq 0$ with $\sum\limits_{k=0}^{\infty} r_k = \infty$ is the failure rate sequence of a distribution on the nonnegative integers. Only the very special case of distributions whose failure rates (r_k) satisfy both $\liminf r_k = 0$ and $\limsup r_k = 1$ is left out of Theorem 1.

Example 1. Geometric with parameter p.

$$(\log n)^{-1/2}\left(N_n + \frac{p\log n}{\log(1-p)}\right) \xrightarrow{d} N\left(0, -\frac{p(1-p)}{\log(1-p)}\right). \tag{2.1}$$

Convergence in (2.1) was previously obtained by Vervaat [7] and Bai et al. [2] using completely different methods. To the best of our knowledge, the cases covered by the next examples are new.

Example 2. Converging failure rates $r_k \to r$, $0 < r < 1$, with $\sum_{i=1}^n |r_i - r|/\sqrt{n} \to 0$.

$$(\log n)^{-1/2}\left(N_n + \frac{r\log n}{\log(1-r)}\right) \xrightarrow{d} N\left(0, -\frac{r(1-r)}{\log(1-r)}\right). \tag{2.2}$$

A concrete example of random variable with converging r_k's is the negative binomial, with $p_k = (-1)^k\binom{-a}{k}p^a(1-p)^k$, $k \geq 0, 0 < p < 1, a > 1$. In this case, (2.2) holds with $r = p$.

Example 3. Alternating geometric with parameters p, q. Here, we mean $r_{2k} = p$ and $r_{2k+1} = q$, where $0 < p < q < 1$ and $k \geq 0$. This random variable can be seen as the number of failures of alternating coins, with respective success probabilities p and q, until the first head (success) shows up. In this case,

$$(\log n)^{-1/2}\left(N_n + \frac{(p+q)\log n}{\log(1-p)(1-q)}\right) \xrightarrow{d} N\left(0, -\frac{p(1-p)+q(1-q)}{\log(1-p)(1-q)}\right).$$

Example 4. Converging failure rates $r_k \to 0$, with $\sum_{k=1}^\infty r_k^2 < \infty$.

$$(\log n)^{-1/2}(N_n - \log n) \xrightarrow{d} N(0,1). \tag{2.3}$$

For a concrete example, consider the rv X with $y_k = (k+1)^{-d}, k \geq 0, d > 0$. Then, $r_k = d/(k+1) + O(k^{-2})$ and (2.3) applies.

Example 5. Converging failure rates $r_k \to 1$ with $\sum(1-r_k) = \infty$.

If $1 - r_k = ak^{-\alpha} + \delta_k$, $k \geq 1$, with $a \in \mathbb{R}_+$, $0 < \alpha \leq 1$ and $\sum|\delta_k| < \infty$, we have

$$(\log m(n))^{-1/2}(N_n - m(n) + a\log m(n)) \xrightarrow{d} N(0,a),$$

for $\alpha = 1$, and

$$(m(n))^{-\frac{1-\alpha}{2}}\left(N_n - m(n) + \frac{a}{1-\alpha}(m(n))^{1-\alpha}\right) \xrightarrow{d} N\left(0, \frac{a}{1-\alpha}\right),$$

for $\alpha < 1$. Also $m(n) \sim \frac{\log n}{\alpha\log\log n}$.

In the particular case of the Poisson distribution with parameter λ, we get

$$(\log\log n)^{-1/2}(N_n - m(n) + \lambda\log(m(n))) \xrightarrow{d} N(0,\lambda),$$

with $m(n) \sim \log n/\log\log n$.

Remark 2. Notice the differences between continuous and discrete distributions. For continuous distributions, the number of records is always asymptotically normal, with the variance growing as $\log n$, regardless of the parent distribution F. For discrete distributions, the asymptotic normality of the number of records depends on the distribution F via the failure rates (r_k): for distributions with very light tails (those with $\sum(1-r_k) < \infty$) the number of records is not asymptotically normal; moreover, when a CLT holds, the variance grows at a speed which depends on (r_k).

§3. Sketch of the proof of Theorem 1

The CLT for records of various discrete models is based on a single fundamental martingale, presented below. The original idea comes from the easily verifiable fact that $N_n - pM_n$ is a martingale, when the underlying rv's are geometric with parameter p.

Proposition 2. *(a) The process*

$$N_n - \theta(M_n) = N_n - \sum_{k=0}^{M_n} r_k, \quad n \geq 1 \tag{3.1}$$

is a square integrable martingale.

(b) Let $\xi_k = I_k - [\theta(M_k) - \theta(M_{k-1})]$, $k \geq 1$, then the increments of the processes of conditional variances in (3.1) are given by

$$E[\xi_k^2 | \mathcal{F}_{k-1}] = \sum_{i > M_{k-1}} p_i(1 - r_i) = \sum_{i > M_{k-1}} r_i y_i.$$

It is important to notice that the process of conditional variances in (3.1) behaves as a sum of partial minima of iid rv's. This is so because $u(M) = \sum_{i>M} r_i y_i$ is a decreasing function of M and therefore, $E[\xi_k^2 | \mathcal{F}_{k-1}] = u(M_{k-1}) = \min\{u(X_1), \ldots, u(X_{k-1})\}$, $k \geq 2$.

>From Proposition 2 above,

$$\sum_{k=2}^{n} E[\xi_k^2 | \mathcal{F}_{k-1}] = \sum_{k=2}^{n} \min\{Z_1, \ldots, Z_{k-1}\} = \sum_{k=2}^{n} z_{M_{k-1}},$$

where $Z_k = \sum_{i>X_k} r_i y_i = \sum_{i>X_k} p_i(1 - r_i), k \geq 1$. These random variables are iid, take values $z_j = \sum_{i>j} r_i y_i = \sum_{i>j} p_i(1 - r_i)$ with probability p_j and their common distribution function G is given by

$$G(z) = \sum_{i \geq j} p_i = y_{j-1}, \quad z_j \leq z < z_{j-1}.$$

Proposition 3. *Let (Z_n) be the sequence of iid r.v. defined above and let*

$$b_n^2 = \sum_{k=0}^{m(n)} \frac{z_k r_k}{y_k}. \tag{3.2}$$

(a) Assume $\sum_{k=0}^{\infty} (1 - r_k) = \infty$. If $\limsup r_k < 1$ or $\liminf r_k > 0$ then

$$\frac{1}{b_n^2} \sum_{k=1}^{n} \min\{Z_1, \ldots, Z_k\} \xrightarrow{P} 1.$$

(b) If $\sum_{k=0}^{\infty} (1 - r_k) < \infty$ then

$$\sum_{k=1}^{n} \min\{Z_1, \ldots, Z_k\} \xrightarrow{a.s.} Z, \tag{3.3}$$

where Z is a finite random variable.

We now get a central limit theorem for the martingale (3.1).

Theorem 4. *Assume* $\sum_{k=0}^{\infty}(1-r_k)=\infty$. *If* $\limsup r_k<1$ *or* $\liminf r_k>0$, *then*

$$\frac{N_n-\theta(M_n)}{b_n}\xrightarrow{d}N(0,1). \tag{3.4}$$

where (b_n) *is defined in (3.2). If* $\sum_{k=0}^{\infty}(1-r_k)<\infty$, *then* $N_n-\theta(M_n)$ *converge a.s. to a finite limit.*

We consider here the final step towards Theorem 1, namely, the substitution of $\theta(M_n)$ by a deterministic sequence (a_n) in (3.4). This amounts to showing that

$$\frac{\theta(M_n)-a_n}{b_n}\xrightarrow{P}0,$$

where (b_n) is defined in (3.2).

Proposition 5. *Assume* $\sum_{k=0}^{\infty}(1-r_k)=\infty$. *If* $\limsup r_k<1$ *or* $\liminf r_k>0$, *then*

$$\frac{\theta(M_n)-\theta(m(n))}{b_n}\xrightarrow{P}0.$$

Proof of Theorem 1

Conclusion (a) of Theorem 1 follows immediately from Theorem 4 and Proposition 5. For (b) note that the tightness of $N_n-m(n)$ is equivalent to

$$\frac{N_n-m(n)}{c_n}\xrightarrow{P}0,$$

for every $(c_n)\nearrow\infty$. Write $N_n-m(n)=N_n-\theta(M_n)+\theta(M_n)-M_n+M_n-m(n)$ and let $(c_n)\nearrow\infty$. The convergence of the series $\sum_{k=0}^{\infty}(1-r_k)$ yields, from Theorem 4, the convergence of the martingale and consequently, $(N_n-\theta(M_n))/c_n\to 0$ a.s. Also $M_n-\theta(M_n)=\sum_{i=0}^{M_n}(1-r_i)$ converges, so $(\theta(M_n)-M_n)/c_n\to 0$ a.s. Last, the same proof of Proposition 3 for the case $\sum_{k=0}^{\infty}(1-r_k)=\infty$ shows that $(M_n-m(n))/c_n\xrightarrow{P}0$.

Acknowledgements

The authors thank support from the FONDAP Project in Applied Mathematics, FONDECYT grants 1020836, 7020836 and MCYT Procject BFM 2001-2449 and CONSI+D Project 119/2001 and GC of D.G.A.

References

[1] ARNOLD, B.C., BALAKRISHNAN, N., NAGARAJA, H. N. (1998). *Records*. Wiley, New York.

[2] BAI, Z., HWANG, H. AND LIANG, W. (1998). Normal approximation of the number of records in geometrically distributed random variables. *Random Struct. Alg.* **13**, 319–334.

[3] DEHEUVELS, P. (1974). Valeurs extrémales d'échantillons croissants d'une variable aléatoire réelle. *Ann. Inst. Henri Poincaré* **X**, 89–114.

[4] GOUET, R., LÓPEZ, F.J. AND SAN MIGUEL, M. (2001). A martingale approach to strong convergence of the number of records and maxima. *Adv. Appl. Prob.* **33**, 864–873.

[5] GOUET, R., LÓPEZ, F.J. AND SANZ, G. (2004). Central limit theorems for the number of records in discrete models. *Preprint*

[6] RENYI, A. (1962). Théorie des éléments saillants d'une suite d'observations. *Ann. Fac. Sci. Univ. Clermont-Ferrand* **8**, 7–13.

[7] VERVAAT, W. (1973). Limit theorems for records from discrete distributions. *Stoch. Proc. Appl.* **1**, 317–334.

Gouet, R.
Dpto. de Ingeniería Matemática
Universidad de Chile y Centro de Modelamiento Matemático UMR 2071 UCHILE-CNRS
Casilla 170-3, Correo 3, Santiago, CHILE.
`rgouet@dim.uchile.cl`

López, F.J. and Sanz, G.
Dpto. de Métodos Estadísticos. Facultad de Ciencias. Universidad de Zaragoza
C/ Pedro Cerbuna, 12
5009 ZARAGOZA. SPAIN
`javier.lopez@unizar.es` and `gerardo@unizar.es`