Analytical Approach to Some Problems Involving Order Statistics

Wojciech Szpankowski
Purdue University

June 16, 1995

[summary by Danièle Gardy]

Abstract

Order statistics, such as the distribution of the maximum of n random variables, are usually studied from a probabilistic point of view. This talk presents an analytical approach that can be applied to a sequence of independent random variables, and to dependent variables. Applications include statistics on digital structures, the analysis of a leader election algorithm, and an extension of probabilistic counting.

1. Order statistics

Let $X_1, X_2, \ldots, X_n$ be a sequence of discrete random variables; the order statistics is the sequence arranged in nondecreasing order: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. The classical theory of order statistics takes place in a probabilistic frame; see for example [2] or [12].

Assume that the variables X_i are *exchangeable:* the $n!$ permutations $X_{i_1}, \ldots, X_{i_n}$ have the same joint distribution [3, p. 228].[1] Define $M_n = \max\{X_1, \ldots, X_n\}$ as the maximum of the n variables. By the inclusion-exclusion principle, we have that

$$\Pr\{M_n > k\} = \sum_{r=1}^{n} (-1)^{r+1} \binom{n}{r} \Pr\{X_1 > k, \ldots, X_n > r\}. \tag{1}$$

Define

$$\widehat{F_r}(z) = \sum_{k \geq 0} \Pr\{X_1 > k, \ldots, X_r > k\} z^k; \qquad \widehat{M_n}(z) = \sum_{k \geq 0} \Pr\{M_n > k\} z^k.$$

Then Equation (1) translates on the generating series as

$$\widehat{M_n}(z) = \sum_{r=1}^{n} (-1)^{r+1} \binom{n}{r} \widehat{F_r}(z).$$

Hence the generating function of M_n (or of the rth-ranked variable of the sequence) is expressed by an alternating sum, which suggests that a Mellin-Rice approach to the asymptotics might be successful (see for example [6] for a general introduction to this subject).

[1] Another definition of exchangeable variables might be: for any subsequence $\{i_j\}$ s.t. $1 \leq i_1 < \cdots < i_r \leq n$, $\Pr\{X_{i_1} < x_1, \ldots, X_{i_r} < x_r\} = \Pr\{X_1 < x_1, \ldots, X_r < x_r\}$.

2. The independent case: the probabilistic approach

2.1. Continuous random variables. In the continuous case, the X_i are i.i.d. and continuous; there exist two sequences of normalization constants $\{a_n\}$ and $\{b_n\}$ and a function H s.t. $\lim_{n\to+\infty} \Pr\{(M_n - a_n)/b_n < x\} = H(x)$; then H is the *limiting distribution* of the (normalized) maximum M_n.

The theory of asymptotic distribution of extremes was initiated by Fischer and Tippett in 1928, and further developed by Gnedenko around 1943; see for example the books by Galambos [7] or Resnick [12] or the presentation given by Sweeting in [13]. The behaviour of the tail determines the limiting distribution $H(x)$ (see for example [2, p. 210] or [7, p. 51-52] for complete conditions). The limiting distribution of the normalized maximum sample has one of the following types: *(i)* If $\Pr(X_i > tx)/\Pr(X_i > t) \to x^{-\alpha}$ $(\alpha > 0)$ for $t \to +\infty$, then $H(x) = \exp(-x^{-\alpha})$ for $x > 0$ and $H(x) = 0$ for $x \leq 0$; *(ii)* If there exists a finite λ s.t. $\Pr(X_i \leq \lambda) = 1$, and $\Pr(X_i > \lambda - \epsilon x)/\Pr(X_i > \lambda - \epsilon) \to x^{\alpha}$ for $\epsilon \to 0$, then $H(x) = \exp(-(-x)^{\alpha})$ for $x < 0$ and $H(x) = 1$ for $x \geq 0$; *(iii)* If $\Pr(X_i > t + xE(X_i - t | X_i > t))/\Pr(X_i > t) \to e^{-x}$ for $t \to +\infty$, then $H(x) = \exp(-e^{-x})$ for all x.

The normalization constants a_n and b_n might be seen respectively as a shift and a scaling factor. They are not necessarily unique: see [2, p. 209] or [12, p. 86] for a discussion of this point; Galambos devotes a whole section of his book to discussing possible choices [7, p. 57–63]. In good cases, a_n and b_n correspond to the limiting mean and variance; see [12, p. 84-85] for conditions that ensure that we can use the mean and variance as scaling factors.

When it is possible to prove conditions on the tail distribution, such as an exponential tail, the asymptotic mean can be computed as: $a_n = \inf\{x : \Pr(X_i \geq x) \leq 1/n\}$.

2.2. Discrete random variables. In this case, Anderson [1] (see also [7, p. 120, ex. 8]) gave a necessary condition for the existence of a_n and b_n s.t. the normalized random variable $(M_n - a_n)/b_n$ converges to a non-degenerate limiting distribution: $\Pr(X_i = k)/\Pr(X_i > k) \to 0$, $k \to \infty$.

The existence of a *limiting* distribution is a strong property, which is not always verified; in some cases we can prove a weaker result on the existence of an *asymptotic* distribution, which might not imply a limiting distribution because of fluctuations. An example of this happens when the X_i follow a geometric distribution: $\Pr(X_i = k) = p^k(1-p)$ for $k \geq 0$. Then it is possible to prove that $\Pr\{M_n < \lfloor \log_{1/p} n + m \rfloor\} \sim \exp\left(-p^{m - \{\log_{1/p} n + m\}}\right)$. Because of the fluctuating nature of the fractional part $\{t\} = t - \lfloor t \rfloor$, this expression oscillates between e^{-p^m} and $e^{-p^{m-1}}$.

3. The independent case: the analytical approach

For i.i.d. variables X_i, we have $\Pr\{X_1 > k, \ldots, X_r > k\} = (P\{X_1 > k\})^r$. Hence $\hat{F}^r = X^{ⓡ}$, with $X^{ⓡ}(z) = \sum_k (\Pr\{X > k\})^r z^k$ standing for r Hadamard products.

When the distribution of the X_i is a sum of geometric distributions, their g.f. is $\hat{X}(z) = a/(1 - \rho z)^d$ for some constants a, ρ and d. Now, for integer d, the coefficient $[z^n]\hat{X}(z)$ is equal to $a\binom{n+d-1}{d-1}\rho^n$ and we can compute a uniform approximation of $\hat{F}^r(z)$:

$$\hat{F}^r(z) = \frac{(dr-r)!a^r}{((d-1)!)^r(1-\rho^r z)^{dr-r+1}} + \Theta\left(\frac{1}{(1-\rho^r z)^{(d-1)r}}\right).$$

From now on a Mellin-Rice approach can be used, to obtain

THEOREM 1. *Let $a_n = \log n + (d-1)\log\log n - \log\Gamma(d)$; then for any integer k*

$$\Pr(M_n \leq a_n + k) = \frac{1}{\log 1/\rho} e^{-\rho^{k+1+\{a_n\}}}(1 + o(1)).$$

4. Digital structures

4.1. Depth in a trie. Consider a trie built on n independent random uniform binary sequences $S_i, 1 \le i \le n$, of 0 and 1. Define D_n as the average depth of an external node [14]; this quantity is related to the number of nodes visited during a successful search. The analysis of D_n leads to considering the length $C_{i,j}$ of the longest prefix common to the sequences S_i and S_j: D_n has the same distribution as $\max\{C_{1,2}, \ldots, C_{1,n}\}$. Although the $C_{i,j}$ are dependent, we still have

$$\Pr\{D_n \ge k\} = \frac{1}{n}\sum_{r=2}^{n}(-1)^r \binom{n}{r} r \Pr\{C_{1,2} \ge k, \ldots, C_{1,r} \ge k\}. \tag{2}$$

The Bernoulli model. In this model, $\Pr\{C_{1,2} \ge k, \ldots, C_{1,r} \ge k\} = (p^r + q^r)^k$. Define $\widehat{G_n}(z) = \sum_{k\ge 0} \Pr\{D_n \ge k\} z^k$; the relation (2) on D_n translates into an equation on $\widehat{G_n}(z)$, which gives after some computations

$$\widehat{G_n}(z) = \frac{1}{2i\pi}\int_{-3/2-i\infty}^{-3/2+i\infty} \frac{n^{-s-1}\Gamma(s+1)}{1 - z(p^{-s} + q^{-s})} ds \left(1 + O\left(\frac{1}{n}\right)\right).$$

In the asymmetric case, D_n follows a normal limiting distribution, with asymptotic mean $\mu_n \sim (1/h)\log n$ and variance $\sigma_n^2 \sim c \log n$; the constant c is $(h_2 - h^2)/h^3$, with $h = -p\log p - q \log q$ and $h_2 = p\log^2 p + q \log^2 q$. The proof relies on Goncharev's condition, characterizing a normal distribution from its g.f.: $\lim_{n\to\infty} e^{-\theta\mu_n/\sigma_n} G_n(e^{\theta/\sigma_n}) = e^{\theta^2/2}$. In the symmetric case ($p = q = 1/2$), the variance is $O(1)$ ($c = 0$), which suggests that Goncharev's condition does not hold and that we cannot expect a normal limiting distribution. Indeed, the asymptotic distribution fluctuates according to the fractional part of $\log_2 n$: $\Pr\{D_n \le \log_2 n + k\} \sim \exp(-2^{k+1+\{\log_2 n\}})/\log 2$.

The Markovian case. In the Markovian model, the next symbol depends on the previous one only; the probability $p_{i,j}$ of obtaining the letter i after the letter j is given by a matrix P. It is possible to write an equation on the g.f. of the depth D_n and a similar analysis [8] shows that D_n again tends to a normal limiting distribution, with a variance of order $\log n$, except for the symmetric independent model, where the variance is $O(1)$.

4.2. An open problem: height of a trie. The approach outlined in Section 4.1 fails when one considers the height of a trie, defined as the maximal depth of all leaves: $H_n = \max\{C_{i,j}, 1 \le i < j \le n\}$. The catch here is that the variables are not exchangeable.

4.3. Depth of a digital search tree. Consider a digital search tree built on n independent keys in the Bernoulli model; as for a trie, let D_n be the average depth of an external node, and define $\mathrm{E}\, B_n(k)$ as the average number of internal nodes at level k. Then $\Pr\{D_n = k\} = \mathrm{E}\, B_n(k)/n$. The generating function $B_n(u) := \sum_{k\ge 0} \mathrm{E}\, B_n(k) u^k$ satisfies the recurrence equation

$$B_{n+1}(u) = 1 + u\sum_{j=0}^{n}\binom{n}{j} p^j q^{n-j}(B_j(u) + B_{m-j}(u)),$$

whose solution can be expressed in terms of $Q_k(u) = \prod_{j=2}^{k+1}(1 - (p^j + q^j)u)$. Again the asymptotic distribution in the symmetric case fluctuates with n, and a central limit theorem can be proved in the asymmetric case [10].

The Lempel-Ziv algorithm for data compression can be modelled by a digital search tree built on independent keys, when the number n of parsed words is known, and its performance can be expressed in terms of parameters of the tree such as the average depth of an external node. A different model considers that the pertinent information is the length of the sequence to be parsed; again this can be modelled by a digital search tree, now with *dependent* keys. It is possible to prove (see [10]) that the distribution of the length of a random phrase is asymptotically the same as the limiting distribution of the depth in the first model, with a digital search tree built on $m = \lfloor nh \log n \rfloor$ nodes (here again h denotes the entropy of the alphabet: $h = -p \log p - q \log q$).

4.4. A leader election algorithm. The algorithm for the election of a loser, analyzed by Prodinger in [11], can be dealt with in a similar manner [4]. The principle of the algorithm is as follows: At the beginning, all players are active; at each step, the active players throw a coin randomly and independently and the set of active players for the next step is exactly those who throw *tail*, or all the former players if all of them throw *heads*; the algorithm ends when a single player throws *tail*. The number of steps required by the algorithm to choose a loser is the height H_n of the leftmost leaf of a trie. The analysis begins with the study of the Poisson model, where the number of keys follows a Poisson distribution, then goes on to extract the statistics for the Bernoulli model by a *Depoissonization Lemma*.

4.5. Probabilistic counting. This generalization of an algorithm by Flajolet and Martin [5], using an array of integers instead of a bitmap, is presented in [9].

Bibliography

[1] Anderson (C. W.). – Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *Journal of Applied Probability*, vol. 7, 1970, pp. 99–113.

[2] Arnold (B.), Balakrishnan (N.), and Nagaraja (H. N.). – *A first course in order statistics.* – Wiley series in Probability and Statistics, 1992.

[3] Feller (W.). – *An Introduction to Probability Theory and its Applications.* – Wiley & Sons, New York, 1971, vol. 2.

[4] Fill (J.), Mahmoud (H.), and Szpankowski (W.). – *On the distribution for the duration of a randomized leader election algorithm.* – Technical Report n° CSD-TR-95-038, Purdue University, 1995.

[5] Flajolet (P.) and Martin (G. N.). – Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, vol. 31, n° 2, October 1985, pp. 182–209.

[6] Flajolet (Philippe) and Sedgewick (Robert). – Mellin transforms and asymptotics: finite differences and Rice's integrals. *Theoretical Computer Science*, vol. 144, n° 1–2, June 1995, pp. 101–124.

[7] Galambos (J.). – *The asymptotic theory of extreme statistics.* – Wiley, New York, 1978.

[8] Jacquet (P.) and Szpankowski (W.). – Analysis of digital tries with Markovian dependency. *IEEE Transactions on Information Theory*, vol. 37, n° 5, 1991, pp. 1470–1475.

[9] Kirschenhofer (P.), Prodinger (H.), and Szpankowski (W.). – How to count quickly and accurately: A unified analysis of probabilistic counting and other related problems. In *Automata, Languages, and Programming, Lecture Notes in Computer Science*, pp. 211–222. – 1992. Proceedings of the 19th ICALP Conference.

[10] Louchard (G.) and Szpankowski (W.). – Average profile and limiting distribution for a phrase size in the Lempel-Ziv parsing algorithm. *IEEE Transactions on Information Theory*, vol. 41, 1995, pp. 478–488.

[11] Prodinger (H.). – How to select a loser. *Discrete Mathematics*, vol. 120, 1993, pp. 149–159.

[12] Resnick (S. I.). – *Extreme values, regular variation and point processes.* – Springer Verlag, 1987.

[13] Sweeting (T. J.). – On domains of uniform local attraction in extreme value theory. *Annals of Probability*, vol. 13, n° 1, 1985, pp. 196–205.

[14] Szpankowski (W.). – On the height of digital trees and related problems. *Algorithmica*, vol. 6, n° 2, 1991, pp. 256–277.