Quickselect and Dickman function

Hsien-Kuei Hwang* and Tsung-Hsi Tsai
Institute of Statistical Science
Academia Sinica
Taipei 115
Taiwan

September 1, 2001

Abstract

We show that the limiting distribution of the number of comparisons used by Hoare's quickselect algorithm when given a random permutation of n elements for finding the m-th smallest element, where $m = o(n)$, is the Dickman function. The limiting distribution of the number of exchanges is also derived.

1 Quickselect

Quickselect is one of the simplest and efficient algorithms in practice for finding specified order statistics in a given sequence. It was invented by Hoare [19] and uses the usual partitioning procedure of quicksort: choose first a partitioning key, say x; regroup the given sequence into two parts corresponding to elements whose values are less than and larger than x, respectively; then decide, according to the size of the smaller subgroup, which part to continue recursively or to stop if x is the desired order statistics; see Figure 1 for an illustration in terms of binary search trees. For more details, see Guibas [15] and Mahmoud [26].

This algorithm[1], although inefficient in the worst case, has linear mean when given a sequence of n independent and identically distributed continuous random variables, or equivalently, when given a random permutation of n elements, where, *here and throughout this paper*, all $n!$ permutations are equally likely.

Let $C_{n,m}$ denote the number of comparisons used by quickselect for finding the m-th smallest element in a random permutation, where the first partitioning stage uses $n-1$ comparisons. Knuth [23] was the first to show, by some differencing argument, that

$$E(C_{n,m}) = 2\left(n + 3 + (n+1)H_n - (m+2)H_m - (n+3-m)H_{n+1-m}\right),$$

for $1 \le m \le n$, where $H_m = \sum_{1\le k\le m} k^{-1}$. A more transparent asymptotic approximation is

$$\frac{E(C_{n,m})}{n} \sim \mu(\alpha), \qquad \mu(\alpha) := 2 - 2\alpha\log\alpha - 2(1-\alpha)\log(1-\alpha),$$

*Part of the work of this author was done while he was visiting School of Computer Science, McGill University. He thanks the School for hospitality and support.

[1]For simplicity of presentation, we only sketched the algorithm of quickselect; when dealing with random input, we assume that the randomness is preserved in each partitioning stage, which is the case of many partitioning algorithms; see Sedgewick (1980) for a detailed discussion.

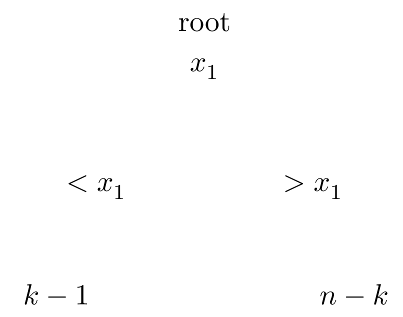

Figure 1: *A binary search tree illustration of quickselect. Given a sequence $\{x_1, \dots, x_n\}$. The binary search tree associated with this sequence is constructed by putting x_1 in the root node, the remaining elements being compared sequentially to x_1. Those smaller (larger) elements go to the left (right) branch and they form (recursively) a binary search tree. Assume that we are selecting the m-th smallest element in this sequence and that the size of the left subtree is $k-1$. If $m = k$ then x_1 (root) is the desired order statistics; if $m < k$ then the search for the m-th smallest element go on recursively in the left subtree, otherwise, the search for the $(m-k)$-th smallest element is conducted in the right subtree.*

for $0 \le \alpha = m/n \le 1$. In particular, $\mu(0) = \mu(1) = 2$. The variance was derived by Kirschenhofer and Prodinger [21], using generating functions,

$$\frac{\mathrm{Var}(C_{n,m})}{n^2} \sim \sigma^2(\alpha),$$

where $\sigma(\alpha) > 0$ is defined by

$$\begin{aligned}
\sigma^2(\alpha) &= \frac{1}{2} - 2\alpha^2 \log^2 \alpha + 4(\alpha^2 - \alpha - 1)\log\alpha \log(1-\alpha) - 2(1-\alpha)^2 \log^2(1-\alpha) \\
&\quad - 4\alpha \log\alpha - 4(1-\alpha)\log(1-\alpha) + \frac{2}{3}\pi^2 \alpha(1-\alpha) + 5\alpha(1-\alpha) \\
&\quad + 4\alpha \int_\alpha^1 \frac{\log t}{1-t}\,\mathrm{d}t + 4(1-\alpha)\int_0^\alpha \frac{\log(1-t)}{t}\,\mathrm{d}t;
\end{aligned}$$

see also Paulsen [31]. Note that $\sigma^2(0) = \sigma^2(1) = 1/2$.

The limiting distribution of $C_{n,m}/n$ was studied independently by Grübel and Rösler [14] and Kodaj and Móri [24]; see also Grübel [12]. Although several (different) characterizations of the limiting distribution of $C_{n,m}/n$ were derived, none of them is simple. Our aim of this paper is to show that when $1 \le m = o(n)$, the limiting distribution can be described in a more transparent way via the Dickman function, extensively studied in number theory and probability theory. By symmetry, the same results hold for $0 \le n - m = o(n)$. While all previous approaches are based essentially on the recurrence relations of $C_{n,m}$, our approach is more combinatorial (in contrast to computational) in nature and relies on proper decomposition of the random variable in question. For more methodological interests, we also sketch another computational approach, using recurrence and generating functions. The number of exchanges is discussed in Section 5. We conclude with some remarks.

2 The Dickman function

The Dickman function $\rho(u)$ is defined as the continuous solution of the differential-difference equation

$$u\rho'(u) + \rho(u-1) = 0 \qquad (u > 1),$$

with the initial condition $\rho(u) = 1$ for $0 \le u \le 1$. It originated in the study by Dickman (1930) who showed that

$$\lim_{n\to\infty} \frac{1}{n} \# \left\{ k \,:\, 1 \le k \le n, \text{the largest prime factor of } k \text{ is } \le n^{1/u} \right\} = \rho(u),$$

for $u \ge 1$; in words, the number of positive integers less than n whose largest prime factor is less than $n^{1/u}$ has the limiting density $\rho(u)$ for $u \ge 1$. The Dickman function plays an important role in analytic number theory, especially for problems in connection with the so-called psixylogy; see Tenenbaum [35], Hildebrand and Tenenbaum [17], and Moree [29] for further information and more instances.

Besides its appearance and applications in number theory (see also Hirth [18]), the Dickman function also arises in a large number of problems like the degree of the largest irreducible factors in random polynomials over finite fields (see Arratia et al. [2], Car [5], Knopfmacher and Manstavicius [22]), the size of the largest cycle in random permutations (see Shepp and Lloyd [34], Gourdon [11]), the sum of products of uniform random variables (see Goldie and Grübel [10], Devroye [7]), and allele frequencies in the infinitely-many neutral alleles diffusion model (see Watterson [36]). See also Arratia [1] and Arratia et al. [3] for a comprehensive survey on scale invariant Poisson processes in which Dickman function appeared in several different forms. Our example of quickselect is a new addition to this list of Dickman function.

Note that (see Tenenbaum [35, §III.5.4])

$$\int_0^\infty \rho(x)\,\mathrm{d}x = e^\gamma,$$

where γ is Euler's constant. For simplicity of reference, we call the distribution with the density function $e^{-\gamma}\rho(x)$ the *Dickman distribution*; see Figure 2 for a plot of the density.

Let Z be a random variable with the Dickman distribution. Some known properties of this distribution are listed as follows.

1. The distribution of Z is infinitely divisible; see Hensley [16].

2. The moment generating function of Z satisfies

$$E(e^{sZ}) = \exp\left(\int_0^s \frac{e^v - 1}{v}\,\mathrm{d}v\right) = -\frac{e^{-\gamma}}{s} \exp\left(\int_{-s}^\infty v^{-1}e^{-v}\,\mathrm{d}v\right), \tag{1}$$

   the second equality holding for $s \in \mathbb{C} \setminus [0,\infty)$.

3. The k-th cumulant of Z is equal to $1/k$; and the k-th moment μ_k of Z satisfies the recurrence

$$\mu_k = k^{-1} \sum_{0\le j\le k-1} \binom{k}{j} \mu_j \qquad (k \ge 1),$$

   with $\mu_0 := 1$.

One simple way of describing the Dickman function is the following. Let $\{X_i\}_{1\le i\le n}$ be a sequence of independent random variables such that

$$P(X_k = k) = k^{-1} \quad \text{and} \quad P(X_k = 0) = 1 - k^{-1},$$

for $k = 1, \dots n$.

Proposition 1. *The limiting distribution of the random variable*

$$D_n := \sum_{1\le j\le n} X_j$$

is Dickman:

$$\lim_{n\to\infty} P(n^{-1}D_n < x) = e^{-\gamma}\int_0^x \rho(v)\,\mathrm{d}v \qquad (x > 0).$$

This result will be used to prove our main result for $C_{n,m}$.
Proof. From (1) and

$$E\left(e^{itD_n}\right) = \prod_{1\le k\le n}\left(1 + \frac{e^{ikt}-1}{k}\right),$$

it suffices, by Lévy's continuity theorem, to show that

$$\lim_{n\to\infty} E\left(e^{itD_n/n}\right) = \exp\left(\int_0^{it}\frac{e^v-1}{v}\,\mathrm{d}v\right), \tag{2}$$

uniformly for finite and real t. Now

$$E\left(e^{itD_n/n}\right) = \exp\left(\sum_{1\le k\le n}\frac{e^{ikt/n}-1}{k} + R_n(t)\right),$$

where

$$R_n(t) := \sum_{j\ge 2}\frac{(-1)^{j-1}}{j}\sum_{1\le k\le n}\frac{(e^{ikt/n}-1)^j}{k^j} = O\left(|t|^2 n^{-1}\right).$$

The dominant term satisfies

$$\sum_{1\le k\le n}\frac{e^{ikt/n}-1}{k} = \int_0^{it}\frac{e^v-1}{v}\,\mathrm{d}v + O\left(|t|^2n^{-1}\right),$$

by a simple application of Euler-MacLaurin summation formula or by the integral representation of the left-hand side. This implies (2) and the proposition. ∎

With more calculations, we can prove a local limit theorem for D_n of the form (see Arratia et al. [4, Corollary 2.8])

$$\lim_{n\to\infty} nP(D_n = k) = e^{-\gamma}\rho(x),$$

where $k/n \to x > 0$; see Figure 2. Interestingly, the histograms of D_n indicate that $P(D_n = k)$ is highly fluctuating for $k \le n$, a fact due to the wide gap of individual X_i.

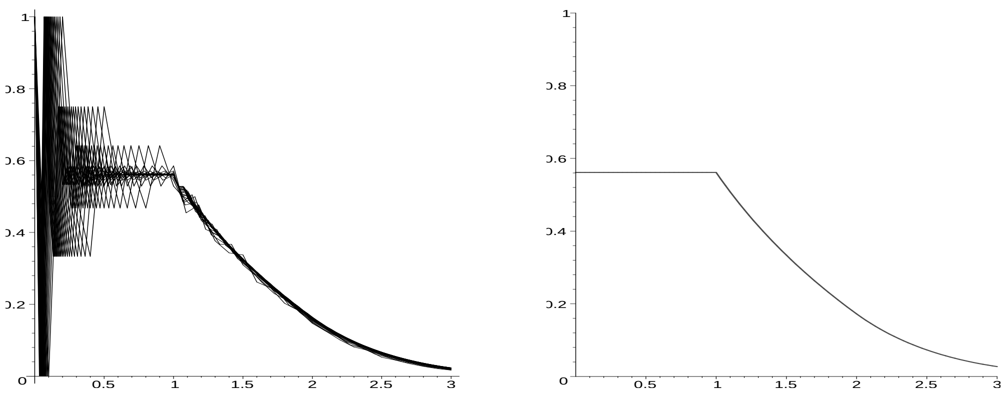

Figure 2: *Histograms of* $nP(D_n = \lfloor xn \rfloor)$ *for* n *from 10 to 30 and* $0 \le x \le 3$ *and the Dickman distribution* $e^{-\gamma}\rho(x)$.

3 Cost of quickselect

Our main result is as follows.

Theorem 1. *If* $1 \le m = o(n)$, *then the limiting distribution of* $n^{-1}C_{n,m} - 1$ *is Dickman:*

$$\lim_{n\to\infty} P\left(\frac{C_{n,m} - n}{n} < x\right) \to e^{-\gamma}\int_0^x \rho(v)\,\mathrm{d}v \qquad (x > 0). \tag{3}$$

By Proposition 1, it suffices to prove the following result. We write $X_n \overset{\mathcal{D}}{\sim} Y_n$ if $X_n \overset{\mathcal{D}}{=} Y_n + Z_n$ and $Z_n \to 0$ in distribution.

Proposition 2. *If* $m = o(n)$ *then*

$$n^{-1}C_{n,m} \overset{\mathcal{D}}{\sim} 1 + n^{-1}D_n.$$

Proof. Let $\mathcal{S}_n$ denote the set of all permutations of n elements. Since we consider quickselect algorithms whose partitioning procedures preserve randomness, we can, for the purpose of analysis, restrict our analysis only to quickselect algorithms that do not exchange keys in the whole process. We thus consider algorithms in which the keys in the subproblems we continue the search for the desired order statistics retain their original (relative) ordering after each partitioning stage; this essentially corresponds to the structure of a binary search tree (we use the first element in each subproblem as the partitioning key).

Let $\xi(i,j)$, $1 \le i,j \le n$, be the indicator that key i (as a partitioning key) compares with key j during the whole quickselect process when given a random permutation. Then

$$C_{n,m} = \sum_{1\le i,j\le n} \xi(i,j).$$

We first prove the case $m = 1$. Observe that in this case $\xi(i,j)$ depends only on the relative ordering of $\{1, 2, \dots, \max\{i,j\}\}$. Let $\sigma \in \mathcal{S}_n$ be the given permutation. Let $A_k \subset \mathcal{S}_n$ denote the event that k appears before $\{1, 2, \ldots, k-1\}$ in σ for $2 \le k \le n$. Then $P(A_k) = 1/k$ and k is used as a partitioning key (at some stage of the quickselect process) if and only if $\sigma \in A_k$. Thus

$$\sum_{1\le j<k} \xi(k,j)(\sigma) = \begin{cases} k-1, & \text{if } \sigma \in A_k; \\ 0, & \text{if } \sigma \notin A_k. \end{cases}$$

On the other hand, we have

$$\sum_{1\le j<k} \xi(j,k)(\sigma) = \begin{cases} 0, & \text{if } \sigma \in A_k; \\ 1, & \text{if } \sigma \notin A_k, \end{cases}$$

since we are selecting the smallest element, and the key k is not present in the remaining process once it is used as a partitioning key ($\sigma \in A_k$) or once it is compared to a partitioning key that is less than k ($\sigma \notin A_k$).

Summing over k from 2 to n, we obtain, by the independence of the sets A_k,

$$n^{-1} \sum_{1\le i,j\le n} \xi(i,j) \stackrel{\mathcal{D}}{\sim} 1 + n^{-1} \sum_{1\le j\le n} X_j. \tag{4}$$

The referee pointed out an alternative proof for this simple case by using binary search trees that proceeds as follows. Observe first that the smallest element in a binary search tree is the node reachable by following only left branches from the root. Thus the number of comparisons used by quickselect to find the smallest element is essentially the sum of the subtree sizes of all nodes lying along the left-most path (or "left-arm") of the tree, yielding

$$C_{n,1} \stackrel{\mathcal{D}}{\sim} n\left(1 + \mathcal{U}_1 + \mathcal{U}_1\mathcal{U}_2 + \cdots\right),$$

where the $\mathcal{U}_j$'s are iid uniform $[0,1]$ random variables. The limiting distribution of the right-hand side is Dickman; see Arratia et al. [3], Chamayou [6], Devroye [7].

For $2 \le m = o(n)$, we observe that the distribution of the sum

$$\sum_{m\le i,j\le n} \xi(i,j)$$

is identical to the distribution of the number of comparisons used to find the smallest element in the set $\{m, \ldots, n\}$, which, by (4), satisfies

$$n^{-1} \sum_{m\le i,j\le n} \xi(i,j) \stackrel{\mathcal{D}}{\sim} 1 + n^{-1} \sum_{1\le j\le n-m+1} X_j \stackrel{\mathcal{D}}{\sim} 1 + n^{-1} \sum_{1\le j\le n} X_j.$$

Thus it remains to prove that the remaining cost is negligible:

$$\sum_{1\le \min\{i,j\}<m} E(\xi(i,j)) = o(n).$$

To this aim, we observe that the sum on the left-hand side is bounded above by

$$\begin{aligned} \sum_{1\le \min\{i,j\}\le m} \xi(i,j) &= \sum_{1\le i,j\le m} \xi(i,j) + \sum_{m<k\le n} \sum_{1\le i<m} \xi(i,k) + \sum_{m<k\le n} \sum_{1\le j<m} \xi(k,j) \\ &=: Z_1 + Z_2 + Z_3. \end{aligned}$$

The first part satisfies

$$E(Z_1) \sim 2m = o(n),$$

since the distribution of Z_1 is identical to the cost of finding the largest (or smallest) element in a random permutation of m elements.

For Z_2 (and Z_3), we note that for fixed k, $m < k \leq n$, the indicators $\{\xi(k,j) : 1 \leq j < k\}$ and $\{\xi(j,k) : 1 \leq j < k\}$ depend only on the relative ordering of $\{1,\ldots,k\}$. Thus it suffices to consider permutations in $\mathcal{S}_k$. Given $\sigma \in \mathcal{S}_k$, let B_ℓ be the event that there are exactly ℓ elements in $\{1,\ldots,m-1\}$ that all appear before $\{m,\ldots,k\}$; let $\{b_1(\sigma),\ldots,b_\ell(\sigma)\} \subset \{1,\ldots,m-1\}$ denote these ℓ elements (in their left-to-right orders). Then the sum

$$\sum_{1\leq j<m} \xi(j,k)(\sigma)$$

is equal to the number of records (or left-to-right maxima) in the sequence $\{b_1(\sigma),\ldots,b_\ell(\sigma)\}$, whose mean is of logarithmic order for $1 \leq \ell < m$; the sum is zero if $\ell = 0$. To see this, observe first that if m is used as the partitioning key then the process stops; otherwise if any element in $\{m+1,\ldots,k\}$ is used as a partitioning key, then k is eliminated from the remaining selection process. So the records in the sequence $\{b_1(\sigma),\ldots,b_\ell(\sigma)\}$ are exactly the partitioning keys in the process in which k is present. Thus

$$P(B_\ell) = \frac{(k-m+1)\binom{m-1}{\ell}}{(k-\ell)\binom{k}{\ell}} = \frac{(m-1)\cdots(m-\ell)(k-m+1)}{k(k-1)\cdots(k-\ell)} \qquad (1 \leq \ell < m);$$

and, consequently,

$$\begin{aligned}
E(Z_2) &= \sum_{m<k\leq n}\sum_{1\leq \ell<m} P(B_\ell)\cdot O\left(1+\log \ell\right) \\
&= O\left(\sum_{m<k\leq n}\frac{m(k-m+1)}{k^2} + \sum_{2\leq\ell<m}\log\ell\sum_{k>m}\frac{k-m+1}{k}(m/k)^\ell\right) \\
&= O\left(m\log\frac{n}{m} + \sum_{2\leq\ell<m}\log\ell\sum_{k>m}\left((m/k)^\ell-(m/k)^{\ell+1}\right)\right) \\
&= o(n) + O\left(m\sum_{2\leq\ell<m}\frac{\log\ell}{\ell(\ell-1)}\right) \\
&= o(n).
\end{aligned}$$

For the final part Z_3, let J_k denote the event that k appears before $\{m,\ldots,k-1\}$. Then $P(J_k) = 1/(k-m+1)$ and k is a partitioning key iff $\sigma \in J_k$. Thus

$$\sum_{1\leq j<m} \xi(k,j)(\sigma) = \begin{cases} m-1-U_{m,k}, & \text{if } \sigma \in J_k; \\ 0, & \text{if } \sigma \notin J_k, \end{cases}$$

where $U_{m,k}$ denotes the maximum element among those elements $\{1,\ldots,m-1\}$ that appear before $\{m,\ldots,k\}$, and $U_{m,k}=0$ if no such element exists. It follows that

$$\begin{aligned}
E(U_{m,k}) &= \sum_{1\le j<m}(m-j)P\left(U_{m,k}=m-j\right)\\
&= \sum_{1\le j<m}\frac{m-j}{k-m+2}\left(\frac{k-m+1}{k-m+2}\right)^{j-1}\\
&\ge m-1-(k-m+2).
\end{aligned}$$

By this inequality and by noting that J_k is independent of $U_{m,k}$, we obtain

$$\begin{aligned}
E(Z_3) &= \sum_{m<k\le n}E\left(\sum_{1\le j<m}\xi(k,j)\right)\\
&= \sum_{m<k\le n}\frac{m-1-E(U_{m,k})}{k-m+1}\\
&\le \sum_{m<k\le 2m}\frac{k-m+2}{k-m+1}+\sum_{2m<k\le n}\frac{m}{k-m+1}\\
&= o(n).
\end{aligned}$$

This completes the proof of the proposition and the theorem. ▮

4 Recurrences and generating functions

We give another direct proof of Theorem 1 in this section. This proof relies on the recurrence relations in question and uses only suitable differencing arguments. It is thus elementary in nature. This method is easily amended for other cost measures like the number of exchanges used by quickselect or other versions of quickselect; see Section 5.

Let $P_{n,m}(y)=E(y^{C_{n,m}})$ denote the probability generating function of $C_{n,m}$. Then $P_{n,m}$ satisfies the recurrence (see [21] and [27])

$$P_{n,m}(y)=\frac{y^{n-1}}{n}\left(1+\sum_{1\le k<m}P_{n-k,m-k}(y)+\sum_{m\le k<n}P_{k,m}(y)\right), \tag{5}$$

for $1\le m\le n$, with the initial condition $P_{n,0}(y)=\delta_{n,0}$, the Kronecker symbol.

Proposition 3. *For $n\ge 1$,*

$$P_n(y):=P_{n,1}(y)=P_{n,n}(y)=y^{n-1}\prod_{1\le j\le n-1}\frac{j+y^{j-1}}{j+1}; \tag{6}$$

and, for $2\le m\le n-1$,

$$P_{n,m}(y)=P_n(y)+y^{n-1}\sum_{m\le j<n}\frac{\prod_{j<h<n}\frac{h+y^{h-1}}{h+1}}{j+1}\sum_{1\le \ell<m}\left(P_{j-\ell+1,m-\ell}(y)-P_{j-\ell,m-\ell}(y)\right). \tag{7}$$

Proof. By multiplying both sides of (5) by n and by taking the difference of $nP_{n,m}(y) - (n-1)yP_{n-1,m}(y)$, we obtain

$$P_{n,m}(y) = \frac{(n-1)y + y^{n-2}}{n} P_{n-1,m}(y) + \frac{y^{n-1}}{n} \sum_{1\le \ell<m} \left(P_{n-\ell,m-\ell}(y) - P_{n-1-\ell,m-\ell}(y)\right), \tag{8}$$

from which (6) and (7) follow. ▮

Proposition 4. *If $m = o(n)$, then*

$$e^{-it} P_{n,m}(e^{it/n}) = \exp\left(\int_0^{it} \frac{e^v - 1}{v}\, \mathrm{d}v\right) + o(1),$$

uniformly for real t.

Proof. For $m = 1$, the result follows from (6) using the same method of proof of Proposition 1. Also by Proposition 1, it suffices that we prove for $m = o(n)$

$$P_{n,m}(e^{it/n}) = P_n(e^{it/n}) + o(1) \qquad (t \in \mathbb{R}),$$

or, equivalently, by (7), $R_{n,m}(e^{it/n}) = o(1)$, where

$$R_{n,m}(y) := \sum_{m\le j<n} \frac{\prod_{j<h<n} \frac{h+y^{h-1}}{h+1}}{j+1} \sum_{1\le\ell<m} \left(P_{j-\ell+1,m-\ell}(y) - P_{j-\ell,m-\ell}(y)\right). \tag{9}$$

We need first an estimate for the difference $|P_{n,m}(y) - P_{n-1,m}(y)|$ when $|y| = 1$. We have, by (8),

$$\begin{aligned} P_{n,m}(y) - P_{n-1,m}(y) \quad &= \quad \left((y-1) + \frac{y^{n-1} - y}{n}\right) P_{n-1,m}(y) \\ &\quad + \frac{y^{n-1}}{n} \sum_{1\le\ell<m} \left(P_{n-\ell,m-\ell}(y) - P_{n-\ell-1,m-\ell}(y)\right), \end{aligned}$$

implying that for $1 \le m \le n$

$$\left|P_{n,m}(e^{iu}) - P_{n-1,m}(e^{iu})\right| \le 2|u| + \frac{1}{n} \sum_{1\le\ell<m} \left|P_{n-\ell,m-\ell}(e^{iu}) - P_{n-\ell-1,m-\ell}(e^{iu})\right| \qquad (u \in \mathbb{R}).$$

Defining $\Delta_{n,m} = \Delta_{n,m}(|u|)$ by $\Delta_{n,1} = 2|u|$ and for $2 \le m \le n$

$$\Delta_{n,m} = 2|u| + \frac{1}{n} \sum_{1\le\ell<m} \Delta_{n-\ell,m-\ell},$$

then $|P_{n,m}(e^{iu}) - P_{n-1,m}(e^{iu})| \le \Delta_{n,m}$ and $\Delta_{n,m}$ can be easily solved by taking the difference $n\Delta_{n,m} - (n-1)\Delta_{n-1,m-1}$, giving

$$\Delta_{n,m} = 2|u| + 2|u| \sum_{n-m+2\le j\le n} j^{-1}.$$

It follows that

$$\left|P_{n,m}(e^{iu}) - P_{n-1,m}(e^{iu})\right| = O(|u|),$$

uniformly for $m = o(n)$. Substituting this estimate in (9), we obtain

$$\left|R_{n,m}(e^{it/n})\right| = O\left(\frac{m|t|}{n} \sum_{m\le j<n} j^{-1}\right) = O\left(\frac{m|t|}{n} \log\frac{n}{m}\right) = o(1),$$

for $m = o(n)$. This completes the proof. ▮

5 The number of key exchanges

Let $W_{n,m}$ denote the number of key exchanges made by quickselect to select the m-th smallest element in a random permutation of n elements. Let $T_{n,m}(y)$ denote the probability generating function of $W_{n,m}$. Then, by conditioning on the value of the partitioning element, we have $T_{n,m}(y) = 1$ for $0 \le n, m \le 2$ and[2]

$$T_{n,m}(y) = \frac{1}{n}\left(\sum_{1\le j<m} T_{n-j,m-j}(y)V_{n,j}(y) + V_{n,m}(y) + \sum_{m<j\le n} T_{j-1,m}(y)V_{n,j}(y)\right), \tag{10}$$

for $n \ge 3$ and $1 \le m \le n$, where $V_{n,j}(y)$ denotes the probability generating function of the number of exchanges used to partition the random permutation into two parts when the partitioning element is j (see Sedgewick [33]):

$$V_{n,j}(y) = \sum_s \frac{\binom{n-j}{s}\binom{j-1}{s}}{\binom{n-1}{j-1}} y^s \qquad (1 \le j \le n),$$

since $\binom{j-1}{s}/\binom{n-1}{j-1}$ is the probability that there are exactly s elements whose indices are less than the rank of the partitioning key and whose values are larger than the partitioning key. Note that

$$\begin{aligned} V_{n,j}^{(k)}(1) &= \sum_s \frac{\binom{n-j}{s}\binom{j-1}{s}}{\binom{n-1}{j-1}} s(s-1)\cdots(s-k+1) \\ &= \frac{(n-j)!(j-1)!(n-k-1)!}{(n-j-k)!(j-1-k)!(n-1)!}. \end{aligned} \tag{11}$$

From (10) and (11), we obtain

$$T'_{n,m}(1) = \frac{n-2}{6} + \frac{1}{n}\sum_{1\le j<m} T'_{n-j,m-j}(1) + \frac{1}{n}\sum_{m<j\le n} T'_{j-1,m}(1),$$

for $1 \le m \le n$ and $n \ge 2$, with $T'_{0,m}(1) = T'_{1,m}(1) = 0$. By the same differencing argument of Knuth (1971), we have for $n \ge 2$ and $1 \le m \le n$

$$\begin{aligned} E(W_{n,m}) &= \frac{1}{3}\left(n + \frac{14}{3} + (n+1)H_n - \left(n - m + \frac{7}{2}\right)H_{n-m+2} - \left(m + \frac{5}{2}\right)H_m\right) \\ &\quad + \frac{1_{1<m<n}}{2(n-m+1)} + \frac{1_{m=1} + 1_{m=n}}{36} \\ &\sim \frac{n}{3}\left(1 - \alpha\log\alpha - (1-\alpha)\log(1-\alpha)\right), \end{aligned} \tag{12}$$

uniformly for $\alpha := m/n \in [0,1]$.

Let $\mathcal{U}$ denote a uniform $[0,1]$ random variable.

Theorem 2. *For $1 \le m = o(n)$,*

$$\frac{W_{n,m}}{n} \xrightarrow{\mathcal{D}} W, \tag{13}$$

[2]Note that the number of exchanges, unlike the number of comparisons, depends on the partitioning key used. For simplicity, we count only the main random part of the number of exchanges.

where

$$W \stackrel{\mathcal{D}}{=} \mathcal{U}W + \mathcal{U}(1-\mathcal{U}),$$

$\mathcal{U}$ and W being independent.

The distribution is uniquely characterized by its moments and not Dickman; see Figure 3. For simulations of the limit law W, see Devroye and Neininger [8]. Our approach is to first prove the result (13) for $m = 1$, then to show that $W_{n,m}$ and $W_{n,1}$ are close in distribution.

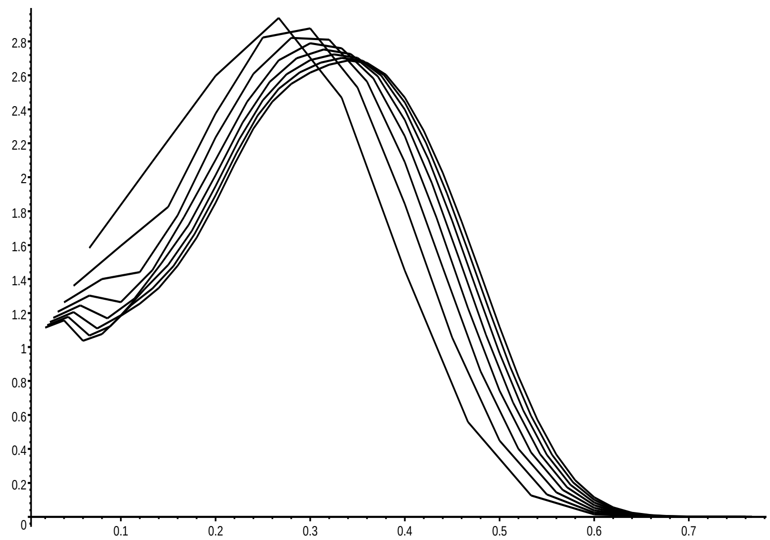

Figure 3: *Histograms of $nP(W_{n,1} = \lfloor xn \rfloor)$ for $n = 15, 20, 25, \ldots, 50$ and $0 \le x < 0.8$.*

Proposition 5. *For $2 \le m = o(n)$,*

$$\left| T_{n,m}(e^{it/n}) - T_{n,1}(e^{it/n}) \right| = o(1), \tag{14}$$

uniformly for $t \in \mathbb{R}$.

We prove the convergence in distribution of $T_{n,1}/n$ by the method of moments, and then prove Proposition (5) by a differencing argument similar to the proof for the number of comparisons. Although the method of moments can also be applied in the case when $2 \le m = o(n)$, the proof given here seems simpler and extends the arguments used in the proof of Proposition 4.

Proof of (13) for $m = 1$. Write for simplicity $T_n(y) = T_{n,1}(y)$ and $\tau_{n,k} := T_n^{(k)}(1)$, the k-th factorial moment of $W_{n,1}$. Then by (10) with $m = 1$

$$\tau_{n,k} = \frac{1}{n} \sum_{1 \le j < n} \tau_{j,k} + v_{n,k} \qquad (n \ge 2),$$

with $\tau_{0,k} = \tau_{1,k} = 0$, where

$$v_{n,k} := \frac{1}{n} \sum_{0 \le \ell < k} \binom{k}{\ell} \sum_{1 \le j < n} \tau_{j,\ell} V_{n,j+1}^{(k-\ell)}(0) \qquad (k \ge 1).$$

We now show, by induction, that

$$\tau_{n,k} \sim g_k n^k \qquad (k \geq 1),$$

for some g_k that will be determined recursively. The case $k = 1$ holds by (12) with $g_1 = 1/3$. Define $g_0 = 1$. By induction and (11)

$$\begin{aligned} v_{n,k} &\sim \frac{1}{n} \sum_{0 \leq \ell < k} \binom{k}{\ell} g_\ell \sum_{1 \leq j < n} j^k (n-j)^{k-\ell} n^{-k+\ell} \\ &\sim n^k \sum_{0 \leq \ell < k} \binom{k}{\ell} g_\ell \int_0^1 x^k (1-x)^{k-\ell} \, \mathrm{d}x \\ &= n^k \sum_{0 \leq \ell < k} \frac{k!k!}{\ell!(2k-\ell+1)!} \, g_\ell. \end{aligned} \tag{15}$$

We need a simple lemma.

Lemma 1. *The solution to the recurrence $a_0 = 0$ and*

$$a_n = b_n + \frac{1}{n} \sum_{1 \leq j < n} a_j \qquad (n \geq 1), \tag{16}$$

where b_n is a given sequence, is given by

$$a_n = b_n + \sum_{1 \leq j < n} \frac{b_j}{j+1}. \tag{17}$$

Proof. Take the difference $na_n - (n-1)a_{n-1}$ and iterate. ∎

Corollary 1 (Asymptotic transfer). *Assume a_n satisfies (16). If $b_n \sim cn^\beta$, where $\beta > 0$, then*

$$a_n \sim c \frac{\beta + 1}{\beta} \, n^\beta. \tag{18}$$

Proof. By (17)

$$a_n \sim cn^\beta + c \sum_{j \leq n} j^{\beta - 1} \sim c \frac{\beta+1}{\beta} \, n^\beta. \quad ∎$$

Applying the asymptotic transfer (18) to $\tau_{n,m}$ using (15), we obtain for $k \geq 2$

$$\frac{\tau_{n,k}}{n^k} \sim \frac{k+1}{k} \sum_{0 \leq \ell < k} \frac{k!k!}{\ell!(2k-\ell+1)!} \, g_\ell.$$

Thus if we define g_k recursively by $g_0 = 1$ and

$$g_k = \frac{k+1}{k} \sum_{0 \leq \ell < k} \frac{k!k!}{\ell!(2k-\ell+1)!} \, g_\ell \qquad (k \geq 1),$$

then

$$E(W_{n,1}^k) \sim \tau_{n,k} \sim g_k n^k \qquad (k \geq 1).$$

We now show that the sequence $\{g_k\}$ uniquely characterizes a distribution by proving that $g_k/k! \le 1$ for $k \ge 0$. By induction using the inequality,

$$\frac{a!b!}{(a+b+1)!} \le \frac{1}{a+b+1} \qquad (a,b \ge 0),$$

we have for $k \ge 1$

$$\frac{g_k}{k!} \le \frac{k+1}{k} \sum_{0 \le \ell < k} \frac{1}{(k-\ell)!(2k-\ell+1)} \le \frac{2}{3} \le 1.$$

Consequently, the distribution W is uniquely characterized by its moments $\{g_k\}$ by Carleman's criterion, which states that *the moment sequence* $\{g_k\}$ *uniquely characterizes the distribution if* $\sum_k g_{2k}^{-1/(2k)} = \infty$. The result (13) for $m = 1$ is thus proved by the Frechet-Shohat moment convergence theorem (see Loève [25]).

Note that the moment generating function $G(z) = E(e^{Wz})$ satisfies

$$G(z) = \int_0^1 G(xz)e^{x(1-x)z}\,\mathrm{d}x.$$

Also

$$\left|T_{n,1}(e^{it/n}) - G(it)\right| = o(1), \tag{19}$$

uniformly in t.

Proof of (5). Consider first the difference $nT_{n,m}(y) - (n-1)T_{n-1,m-1}(y)$. By (10), we have

$$\begin{aligned} n(T_{n,m}(y) - T_{n-1,m-1}(y)) \quad &= \sum_{2 \le j < m} \Big(V_{n,j}(y) - V_{n-1,j-1}(y)\Big)T_{n-j,m-j}(y) + V_{n,m}(y) - V_{n-1,m-1}(y) \\ &\quad + \sum_{m < j \le n} \Big(T_{j-1,m}(y) - T_{j-2,m-1}(y)\Big)V_{n,j}(y) \\ &\quad + \sum_{m < j \le n} \Big(V_{n,j}(y) - V_{n-1,j-1}(y)\Big)T_{j-2,m-1}(y). \end{aligned}$$

Then

$$n\left|T_{n,m}(y) - T_{n-1,m-1}(y)\right| \le \sum_{2 \le j \le n} |V_{n,j}(y) - V_{n-1,j-1}(y)| + \sum_{m < j \le n} \left|T_{j-1,m}(y) - T_{j-2,m-1}(y)\right|.$$

We need an upper bound for the "total variation" $\sum_{2 \le j \le n} |V_{n,j}(y) - V_{n-1,j-1}(y)|$. Observe first that

$$V_{n,j}(y) = \sum_s \frac{\binom{j-1}{s}\binom{n-j}{s}}{\binom{n-1}{j-1}} y^s = \sum_s \frac{\binom{j-1}{s}\binom{n-j}{s}}{\binom{n-1}{s}} (y-1)^s.$$

From this and the inequalities

$$\frac{\binom{j-1}{s}}{\binom{n-1}{s}} \ge \frac{\binom{j-2}{s}}{\binom{n-2}{s}} \qquad (2 \le j \le n; 0 \le s \le j-2),$$

it follows that

$$\begin{aligned}
&\sum_{2\le j\le n} \left|V_{n,j}(e^{iu}) - V_{n-1,j-1}(e^{iu})\right| \\
&\le \sum_{2\le j\le n}\sum_{s\ge 1}\binom{n-j}{s}|e^{iu}-1|^s\left(\frac{\binom{j-1}{s}}{\binom{n-1}{s}} - \frac{\binom{j-2}{s}}{\binom{n-2}{s}}\right) \\
&= \sum_{1\le s\le n/2}|e^{iu}-1|^s\left(\sum_j \frac{\binom{j-1}{s}\binom{n-j}{s}}{\binom{n-1}{s}} - \sum_j \frac{\binom{j-2}{s}\binom{n-j}{s}}{\binom{n-2}{s}}\right) \\
&= \sum_{1\le s\le n/2}|e^{iu}-1|^s\left(\frac{\binom{n}{2s+1}}{\binom{n-1}{s}} - \frac{\binom{n-1}{2s+1}}{\binom{n-2}{s}}\right) \\
&\le \sum_{1\le s\le n/2}|e^{iu}-1|^s\frac{s!}{(2s+1)!}(n-s-2)\cdots(n-2s)\left(n(s+1)-2s-1\right) \\
&\le \sum_{s\ge 1}\frac{(s+1)!}{(2s+1)!}|u|^s n^s \\
&\le n|u|\phi(n|u|),
\end{aligned}$$

where $\phi(z)$ is an entire function with nonnegative coefficients. Now define $\Delta_{n,m}(u)$ by $\Delta_{n,0}(u) = 0$ and for $n \ge m \ge 1$

$$\Delta_{n,m}(u) = |u|\phi(n|u|) + \frac{1}{n}\sum_{m<j\le n}\Delta_{j-1,m}(u). \tag{20}$$

Then $|T_{n,m}(e^{iu}) - T_{n-1,m-1}(e^{iu})| \le \Delta_{n,m}(u)$. The recurrence (20) is easily solved using (17):

$$\begin{aligned}
\Delta_{n,m}(u) &= |u|\phi(n|u|) + |u|\sum_{m<j<n}\frac{\phi(k|u|)}{k+1} \\
&\le |u|\phi(n|u|) + |u|\phi(n|u|)\log\frac{n}{m},
\end{aligned}$$

for $m < n$. Now

$$\begin{aligned}
\left|T_{n,m}(e^{iu}) - T_{n,1}(e^{iu})\right| &\le \sum_{0\le j\le m-2}\left|T_{n-j,m-j}(e^{iu}) - T_{n-j-1,m-j-1}(e^{iu})\right| \\
&\quad + \left|T_{n,1}(e^{iu}) - T_{n-m+1,1}(e^{iu})\right| \\
&\le m|u|\phi(n|u|) + |u|\phi(n|u|)\sum_{0\le j\le m-2}\log\frac{n-j}{m-j} \\
&\quad + \left|T_{n,1}(e^{iu}) - T_{n-m+1,1}(e^{iu})\right|.
\end{aligned}$$

Thus, substituting $u = t/n$ and using (19), we obtain

$$\begin{aligned}
\left|T_{n,m}(e^{it/n}) - T_{n,1}(e^{iu/n})\right| &\le \frac{m}{n}|t|\phi(|t|) + \frac{|t|}{n}\phi(|t|)\sum_{0\le j\le m-2}\log\frac{n-j}{m-j} \\
&\quad + \left|T_{n,1}(e^{it/n}) - G(it)\right| + \left|T_{n-m+1,1}(e^{it/n}) - G(it)\right| \\
&= o(1) + \left|T_{n-m+1,1}(e^{it/(n-m+1)}) - G(it)\right| \\
&= o(1),
\end{aligned}$$

uniformly for $m = o(n)$ and $t \in \mathbb{R}$. This completes the proof of (5). ▮

Note that the number of exchanges, denoted by $\mathcal{V}_n$, used at the first partitioning stage satisfies, by (11),

$$E(\mathcal{V}_n^k) \sim \frac{k!k!(n-k-1)!}{(2k+1)!(n-2k-1)!} \sim \frac{k!k!}{(2k+1)!}\, n^k,$$

for $k \geq 1$. Thus we deduce that $\mathcal{V}_n/n$ has in the limit a beta distribution:

$$P\left(\frac{\mathcal{V}_n}{n} < x\right) \to 1 - \sqrt{1-4x} \qquad (0 < x < 1/4).$$

6 Extensions

Random rank selection. Mahmoud et al. [27] considered the problem of selecting the τ-th smallest element using quickselect, where τ takes any of the n values $\{1, \ldots, n\}$ with equal probability. Let B_n denote the number of comparisons used by quickselect when given a random permutation. Then the bivariate generating function $B(x,y) := \sum_n E(y^{B_n})x^n$ satisfies (see [27])

$$x\frac{\partial^2}{\partial x^2}B(x,y) + \frac{\partial}{\partial x}B(x,y) = \frac{2x}{1-xy}\frac{\partial}{\partial x}B(xy,y) + \frac{1}{(1-xy)^2},$$

with the initial conditions $B(0,y) = 0$ and $(\partial/\partial x)B(x,y)|_{x=0} = 1$. This equation can be *explicitly* solved as for $P_{n,m}(y)$ above; the result is

$$E(y^{B_n}) = \frac{n+1}{n}\, y^{n-1} \sum_{1\leq j\leq n} \frac{1}{j(j+1)} \prod_{j\leq h<n} \frac{h+2y^{h-1}}{h+2} \qquad (n \geq 1).$$

From this expression, we can proceed as above to show that

$$\lim_{n\to\infty} E\left(e^{it(B_n-n)/n}\right) = \exp\left(2\int_0^{it} \frac{e^v-1}{v}\, \mathrm{d}v\right),$$

(see [27]) and thus

$$\lim_{n\to\infty} P\left(\frac{B_n-n}{n} < x\right) = e^{-2\gamma}\int_0^x \rho(v)\rho(x-v)\, \mathrm{d}v \qquad (x > 0).$$

By the same approach, it can be shown that the multiple quickselect problems in Mahmoud and Smythe [28] introduce higher convolutions of the Dickman distribution; see also Panholzer and Prodinger [32].

Quickselect with median-of-three. If instead of choosing a random element to partition the random input, we take the median of three random elements as the partitioning key, then the total cost is in most cases reduced. We can extend our differencing method to describe the limiting distribution of the cost measures of this algorithm. For example, the moments, μ_k, of the limiting distribution of the number of comparisons satisfy (when $m = o(n)$) $\mu_0 = 1$ and

$$\mu_k = 6\frac{(k+2)(k+3)}{k(k+5)} \sum_{0\leq j<k} \binom{k}{j} \frac{\mu_j}{(j+2)(j+3)} \qquad (k \geq 1).$$

From this we deduce that the moment generating function $M(z)$ satisfies the second-order differential equation

$$z^2M''(z) - 2z(z-3)M'(z) + (z^2 - 6z + 6)M(z) = 6e^zM(z),$$

with the initial conditions $M(0) = 1$ and $M'(0) = 2$. Note that the mean is asymptotically the same as the quickselect proper; see Kirschenhofer et al. [20], Grübel [13], Neininger [30] for related materials.

Other Dickman "relatives". Most examples we mentioned in the Introduction actually involve the Dickman function as the asymptotic distribution function instead of as the limiting density function that we encountered in this paper. Although different, there are common features these problems share. The definition of D_n reflects to some extent the "origin" of the Dickman distribution, which is the sum of independent random variables with wide "gaps" with probabilities proportional to the values they take. This situation also appears in other Dickman problems. For example, the distribution function of the longest cycle, say L_n, in a random permutation satisfies

$$\sum_n P(L_n \le m)z^n = \exp\left(\sum_{1\le j\le m} \frac{z^j}{j}\right).$$

This seems to indicate that the Dickman distribution is a rather special instead of universal (like Gaussian) phenomenon. See Gourdon [11] for other variants of the Dickman distribution.

From normal to Dickman? Consider the random variable I_n whose probability generating function is given by

$$E(y^{I_n}) = \prod_{1\le j\le n}\left(1 + \frac{1}{j}\sum_{\ell\in E_j}\left(y^\ell - 1\right)\right),$$

where $E_j \subset \{1,\ldots,j\}$ for $1 \le j \le n$. If $E_j = \{1\}$, then I_n is the number of records, or cycles, in a random permutation (Stirling numbers of the first kind); if $E_j = \{1,\ldots,j\}$, then I_n is the number of inversions in a random permutation; if $E_j = \{j\}$, then $I_n = D_n$. So I_n may be interpreted as generalized inversions in permutations (counting only inversions produced by some specified numbers). While the distributions in the cases of records and inversions are asymptotically normal, the limiting distribution of D_n is not normal. In general, for what sets of E_j will I_n be asymptotically normally distributed? When will it be Dickman?

Acknowledgements

We thank the referee for helpful comments.

References

[1] Arratia, R. (1998) On the central role of scale invariant Poisson processes on $(0,\infty)$. Microsurveys in discrete probability (Princeton, NJ, 1997), 21–41, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 41, AMS, Providence, RI.

[2] Arratia, R., Barbour, A. D. and Tavaré, S. (1993) On random polynomials over finite fields. *Mathematical Proceedings of the Cambridge Philosophical Society*, **114** 347–368.

[3] Arratia, R., Barbour, A. D. and Tavaré, S. (1999) The Poisson-Dirichlet distribution and the scale-invariant Poisson process. *Combinatorics, Probability and Computing*, **8** 407–416.

[4] Arratia, R., Barbour, A. D. and Tavaré, S. (2000) Limits of logarithmic combinatorial structures. *Annals of Probability*, **28** 1620–1644.

[5] Car, M. (1987) Théorèmes de densité dans $F_q[X]$. *Acta Arithmetica*, **48** 145–165.

[6] Chamayou, J.-M.-F. (1973) A probabilistic approach to a differential-difference equation arising in analytic number theory. *Mathematics of Computations*, **27** 197–203.

[7] Devroye, L. (2001) Simulating perpetuities. *Methodologies and Computing in Applied Probability*, **3** 97–115.

[8] Devroye, L. and Neininger, R. (2001) Exact simulation of random variables that are solutions of fixed-point equations. Preprint.

[9] Dickman, K. (1930) On the frequency of numbers containing prime factors of a certain relative magnitude. *Arkiv för Matematik, Astronomi och Fysik*, **22** 1–14.

[10] Goldie, C. M. and Grübel, R. (1996) Perpetuities with thin tails. *Advances in Applied Probability*, **28** 463–480.

[11] Gourdon, X. (1996) *Combinatoire, Algorithmique et Géométrie des Polynômes*. Ph.D. Thesis, Ecole Polytechnique.

[12] Grübel, R. (1998) Hoare's selection algorithm: a Markov chain approach. *Journal of Applied Probability*. **35** 36–45.

[13] Grübel, R. (1999) On the median-of-k version of Hoare's selection algorithm. *Theoretical Informatics and Applications*, **33** 177–192.

[14] Grübel, R. and Rösler, U. (1996) Asymptotic distribution theory for Hoare's selection algorithm. *Advances in Applied Probability*, **28** 252–269.

[15] Guibas, L. J. (1975) A principle of independence for binary tree searching. *Acta Informatica*, **4** 293–298.

[16] Hensley, D. (1986) The convolution powers of the Dickman function. *Journal of the London Mathematical Society*, **33** 395–406.

[17] Hildebrand, A. and Tenenbaum, G. (1993) Integers without large prime factors. *Journal de Théorie des Nombres de Bordeaux*, **5** 411–484.

[18] Hirth, U. M. (1997) Probabilistic number theory, the GEM/Poisson-Dirichlet distribution and the arc-sine law. *Combinatorics, Probability and Computing*, **6** 57–77.

[19] Hoare, C. A. R. (1961) Algorithm 65, FIND. *Communications of the ACM*, **4** 321–332.

[20] Kirschenhofer, P., Martínez C. and Prodinger, H. (1997) Analysis of Hoare's find algorithm with median-of-three partition. *Random Structures and Algorithms*, **10** 143-156.

[21] Kirschenhofer, P. and Prodinger, H. (1998) Comparisons in Hoare's Find algorithm. *Combinatorics, Probability and Computing*, **7** 111–120.

[22] Knopfmacher, A. and Manstavicius, E. (1997) On the largest degree of an irreducible factor of a polynomial in $F_q[X]$. *Lietuvos Matematikos Rinkinys*, **37** (1997), 50–60. Translation in *Lithuanian Mathematical Journal*, **37** 38–45.

[23] Knuth, D. E. (1972) Mathematical analysis of algorithms. In *Information Processing 71* (Proc. IFIP Congress, Ljubljana, 1971), Vol. 1: Foundations and systems, pp. 19–27. North-Holland, Amsterdam, 1972.

[24] Kodaj, B. and Móri, T. F. (1997) On the number of comparisons in Hoare's algorithm "FIND". *Studia Scientiarum Mathematicarum Hungarica*, **33** 185–207.

[25] Loève, M. (1977) *Probability Theory. I*, Fourth Edition, Springer-Verlag, New York.

[26] Mahmoud, H. M. (2000) Sorting. A Distribution Theory. Wiley-Interscience, New York.

[27] Mahmoud, H. M., Modarres, R. and Smythe, R. T. (1995) Analysis of QUICKSELECT: an algorithm for order statistics. *RAIRO Informatique Théorique et Applications*, **29** 255–276.

[28] Mahmoud, H. M. and Smythe, R. T. (1998) Probabilistic analysis of Multiple Quick Select. *Algorithmica*, **22** 569–584.

[29] Moree, P. (1993) *Psixylogy and Diophantine Equations*. Dissertation, Rijksuniversiteit te Leiden, Leiden, 1993.

[30] Neininger, R. (1999) *Limit Laws for Random Recursive Structures and Algorithms*. Ph. D. Dissertation, Universität Freiburg.

[31] Paulsen, V. (1997) The moments of FIND. *Journal of Applied Probability*. **34** 1079–1082.

[32] Panholzer, A. and Prodinger, H. (1998) A generating functions approach for the analysis of grand averages for multiple QUICKSELECT. *Random Structures and Algorithms*, **13** 189–209.

[33] Sedgewick, R. (1980) *Quicksort*. Gurland Publishing, NY.

[34] Shepp, L. A. and Lloyd, S. P. (1966) Ordered cycle lengths in a random permutation. *Transactions of the American Mathematical Society*, **121** 340–357.

[35] Tenenbaum, G. (1995) *Introduction to Analytic and Probabilistic Number Theory*. Translated from the second French edition (1995) by C. B. Thomas. Cambridge University Press, Cambridge.

[36] Watterson, G. A. (1976) The stationary distribution of the infinitely-many neutral alleles diffusion model. *Journal of Applied Probability*, **13** 639–651; Correction: *ibid*, **14** 897.