Explicit and asymptotic formulae for the expected values of the order statistics of the Cantor distribution

Arnold Knopfmacher[a], Helmut Prodinger[b]

[a]Department of Computational & Applied Mathematics, Witwatersrand University, 2050, South Africa

[b]Department of Algebra and Discrete Mathematics, Technical University of Vienna, A-1040 Vienna, Austria

Abstract

We derive the exact solution to a recurrence relation obtained by Hosking for the expected value of the minimum order statistic of the Cantor distribution. In addition, we indicate how an asymptotic estimate can be derived for this and similar sums involving binomial coefficients and Bernoulli numbers.

Key words: order statistics, alternating sums, Bernoulli numbers, Rice's method

1. Introduction

In a recent article Hosking (1994) studied the order statistics of the Cantor distribution. In particular he derived a recurrence relation from which the average values of the order statistics could be computed step by step.

In this paper we derive an explicit solution to this recurence relation as well as a simple asymptotic approximation to the exact result. In doing this we indicate a general approach that could be used to solve other recurrence relations of a similar nature.

Rather than speaking of the Cantor distribution we will interpret the recursion in terms of *random strings* $c_1\, c_2\, c_3 \ldots$ where the

c_i's $\in \{0,1\}$ are independently obtained by flips of a fair coin. Given a parameter ϕ with $0 < \phi \le \frac{1}{2}$ we then consider the random variable T that maps $c_1\, c_2\, c_3 \ldots$ to the real number

$$T(c_1\, c_2\, c_3 \ldots) = \frac{1-\phi}{\phi} \sum_{i\ge 1} c_i \phi^i \ \in [0,1]\ .$$

The strings now have a natural order from the usual ordering of the real numbers. This is easily seen to be equivalent to the lexicographic ordering of strings, i.e. $c_1\, c_2\, c_3 \ldots <$ $d_1\, d_2\, d_3 \ldots$ iff there is a k such that $c_i = d_i$ for $i = 1, \ldots, k-1$ and $c_k = 0$ while $d_k = 1$. It thus makes sense to speak of order statistics for strings. Suppose that n independent random strings $w_1, \ldots, w_n$ are produced by this process. We denote by a_n the average value of the minimum of the n real numbers $T(w_1), \ldots, T(w_n)$. Similarly let $a_n^{(k)}$ denote the average value of the kth smallest number, $1 \le k \le n$, with $a_n^{(1)} \equiv a_n$. The particular choice $\phi = \frac{1}{3}$ produces real numbers $T(w_i)$ that belong to the Cantor set, hence the choice of name for the distributions.

Hosking (1994, equation 5) derives the following recursion for the expected minimum

$$(2^n - 2\phi)a_n = 1 - \phi + \phi \sum_{k=1}^{n-1} \binom{n}{k} a_k \ , \quad n \ge 1 \ . \tag{1.1}$$

It is convenient to set $a_0 = 0$.

To solve this recursion exactly, we follow the approach of Knuth (1973). Let $A(z) = \sum_{n\ge 0} a_n z^n/n!$ be the exponential generating function of the a_n's. Then (1.1) translates to the functional equation

$$A(2z) - 2\phi A(z) = (1-\phi)(e^z - 1) + \phi(e^z A(z) - A(z))$$

or

$$A(2z) = \phi(1+e^z)A(z) + (1-\phi)(e^z - 1) \ .$$

Now introduce

$$\hat{A}(z) = \frac{1}{e^z - 1}\, A(z) = \sum_{n\ge 0} \hat{a}_n\, z^n/n!$$

where we obtain the simpler equation

$$\hat{A}(2z) = \phi \hat{A}(z) + \frac{1-\phi}{e^z + 1} \ .$$

Equating coefficients of $\frac{z^n}{n!}$ on both sides leads to (Abramowitz and Stegun (1972, pp. 804-5))

$$2^n \hat{a}_n = \phi \hat{a}_n - (1-\phi)\, \frac{B_{n+1}}{n+1}\, (2^{n+1} - 1)$$

where B_n denotes a Bernoulli number defined by

$$\sum_{n\ge 0} B_n\, z^n/n! = \frac{z}{e^z - 1} \ .$$

Therefore

$$\hat{a}_n = -(1-\phi)\, \frac{B_{n+1}}{n+1}\, \frac{2^{n+1} - 1}{2^n - \phi} \ .$$

Now from the relation $A(z) = (e^z - 1)\hat{A}(z)$ we obtain

$$a_n = \sum_{k=0}^{n-1} \binom{n}{k} \hat{a}_k \ , \quad n \geq 1 \ ,$$

which yields the explicit formula

$$a_n = -(1-\phi) \sum_{k=0}^{n-1} \binom{n}{k} \frac{B_{k+1}}{k+1} \frac{2^{k+1}-1}{2^k - \phi} \ .$$

It is convenient to rewrite this solution to Hosking's recurrence in the form

$$a_n = -\frac{1-\phi}{n+1} \sum_{k=1}^{n} \binom{n+1}{k} B_k \frac{2^k - 1}{2^{k-1} - \phi} \ . \tag{1.2}$$

It is interesting to note that in the limiting case $\phi = \frac{1}{2}$ this simplifies to

$$a_n = -\frac{1}{n+1} \sum_{k=1}^{n} \binom{n+1}{k} B_k = \frac{1}{n+1} \ , \quad n \geq 1 \ ,$$

using the recurrence for the Bernoulli numbers

$$B_{n+1} = \sum_{k=0}^{n+1} \binom{n+1}{k} B_k \ , \quad B_0 = 1 \ .$$

In the other limiting case $\phi = 0$ we obtain from (1.2) (or immediately from (1.1)) the solution

$$a_n = \frac{1}{2^n} \ , \quad n \geq 1 \ .$$

2. Asymptotics

For the general case $0 < \phi < \frac{1}{2}$, (1.2) does not simplify further. Instead we show how to obtain an asymptotic expression for a_n. Consider the sum

$$S_n = \sum_{k=1}^{n-1} \binom{n}{k} B_k \frac{2^k - 1}{2^{k-1} - \phi}$$

and note that $a_n = -\frac{1-\phi}{n+1} S_{n+1}$. We express this sum as a contour integral (see e.g. Flajolet and Sedgewick (1994))

$$S_n = \frac{1}{2\pi i} \int_{\mathcal{C}} \frac{(-1)^n n!}{z(z-1)\cdots(z-n)} \left(-z\zeta(1-z)\right) \frac{2^z - 1}{2^{z-1} - \phi} \, dz$$

where $\zeta(s)$ denotes Riemann's zeta function and where $\mathcal{C}$ denotes a positively oriented curve that encircles the poles $1, 2, \ldots, n-1$ of the integrand and no other poles. The choice of integrand here is motivated in part by the result $B_k = -k\zeta(1-k)(-1)^k$, $\quad k \geq 1$. The name commonly used for this approach is "Rice's method". Szpankowski (1988) noted that one can change the path of integration above and in addition use the relation

$$\begin{aligned} \frac{(-1)^{n-1} n!}{z(z-1)\cdots(z-n)} &= \frac{\Gamma(n+1)\Gamma(-z)}{\Gamma(n+1-z)} \\ &= n^z \Gamma(-z) \left(1 + O\left(\frac{1}{n}\right)\right) \end{aligned}$$

to produce the alternative representation,

$$S_n = \frac{1}{2\pi i} \int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} n^z \Gamma(1-z)\zeta(1-z) \, \frac{2^z - 1}{2^{z-1} - \phi} \, dz \cdot \left(1 + O\left(\frac{1}{n}\right)\right) .$$

Shifting the line of integration to the left and collecting the residues will yield the asymptotic expansion. The pole of the zeta function at $z = 0$ is cancelled by the factor $2^z - 1$, so only the solutions of the equation

$$2^{z-1} - \phi = 0$$

are candidates for poles. They are

$$z = 1 + \log_2 \phi + \chi_k , \quad k \in \boldsymbol{Z} \quad \text{with} \quad \chi_k = \frac{2k\pi i}{\log 2} .$$

The residue at such a point is

$$n^{1+\log_2 \phi + \chi_k} \Gamma(-\log_2 \phi + \chi_k)\zeta(-\log_2 \phi + \chi_k) \, \frac{2\phi - 1}{\phi \log 2} .$$

Now note that

$$n^{\chi_k} = e^{2k\pi i \log_2 n}$$

and define

$$\delta(x) = \sum_{k \neq 0} \Gamma(-\log_2 \phi + \chi_k)\zeta(-\log_2 \phi + \chi_k) e^{2k\pi i x} .$$

This is a periodic function of x with period 1, mean 0 and a very small amplitude provided that ϕ is not too close to 0. Then for our sum S_n we have the asymptotic estimate

$$S_n = n^{1+\log_2 \phi} \, \frac{2\phi - 1}{\phi \log 2} \left\{ \Gamma(-\log_2 \phi)\zeta(-\log_2 \phi) + \delta(\log_2 n) \right\} \cdot \left(1 + O\left(\frac{1}{n}\right)\right)$$

and consequently

$$a_n = n^{\log_2 \phi}\ \frac{(1-\phi)(1-2\phi)}{\phi \log 2}\Big\{\Gamma(-\log_2 \phi)\zeta(-\log_2 \phi) + \delta(\log_2 n)\Big\}\left(1 + O\left(\frac{1}{n}\right)\right)\ . \quad (2.1)$$

If we disregard the numerically small fluctuating term we can write

$$a_n \sim n^{\log_2 \phi}\ \frac{(1-\phi)(1-2\phi)}{\phi \log 2}\ \Gamma(-\log_2 \phi)\zeta(-\log_2 \phi)\ . \quad (2.2)$$

In the table below we compare the approximation (2.2) to the exact value for a_n obtained from the recursion (1.1) in the 'classical' case $\phi = \frac{1}{3}$ for which $a_n \sim \frac{c}{n^{\log_2 3}}$ where $c = 2\frac{\Gamma(\log_2 3)\zeta(\log_2 3)}{3 \log 2} = 1.99670497\ldots$.

n	a_n	$a_n(n+1)^{\log_2 3}$
20	0.0156734	1.95358
40	0.0054856	1.97436
60	0.00293339	1.98176
80	0.00187479	1.98532
100	0.001312296	1.98756
120	0.000994329	1.98915
140	0.000780667	1.9902
160	0.000632884	1.99095
180	0.000525843	1.99155
200	0.00044548	1.99208

Figure 1: Graph of a_n versus n in decreasing orders of magnitude for $\phi = 0.5,\ 0.4,\ 0.3,$ 0.2 and 0.1.

Another (possibly more complicated) method of producing this result can be found in the classical book of Knuth (1973). In addition by manipulating the explicit solution (1.2) for a_n we can rewrite it in the form

$$a_n = \frac{1-\phi}{n+1}\left\{2 + \frac{1-2\phi}{\phi}\sum_{k=1}^{n}\binom{n+1}{k}\frac{B_k}{\phi^{-1}2^{k-1}-1}\right\}$$

to which the general result of Prodinger (1993, Theorem 20) is applicable and again yields (2.2). However this particular theorem unlike the method above would not necessarily apply to other similar recursions of this type.

It is worthwhile to note that explicit formulas for $a_n^{(k)}$ for k fixed can now be deduced from (1.2) with the aid of the recurrence Hosking (1994, equation 6)

$$n a_{n-1}^{(k)} = (n-k)a_n^{(k)} + k a_n^{(k+1)}\ .$$

For example we have

$$
\begin{aligned}
a_n^{(2)} &= a_n + n(a_{n-1} - a_n) = n a_{n-1} - (n+1)a_n + 2a_n \\
&= -(1-\phi)\sum_{k=1}^{n-1}\binom{n}{k} B_k \,\frac{2^k-1}{2^{k-1}-\phi} + (1-\phi)\sum_{k=1}^{n}\binom{n+1}{k} B_k \,\frac{2^k-1}{2^{k-1}-\phi} + 2a_n \;.
\end{aligned}
\tag{2.3}
$$

More generally David (1981, p.48) provides the exact formula

$$
a_n^{(k)} = \sum_{j=0}^{k-1}\binom{n-j-1}{n-k}\binom{n}{j}(-1)^{k-j-1}a_{n-j} \;. \tag{2.4}
$$

Using (2.4), it is possible to obtain asymptotics for $a_n(k)$, for every fixed k, in the same manner as for a_n. We obtain (apart from the fluctuations)

$$
a_n^{(k)} \sim c(k)a_n \;.
$$

where

$$
c(k) = \prod_{m=1}^{k-1}\left(1 - \frac{\log_2\phi}{m}\right) \;.
$$

In the limiting case $\phi = \frac{1}{2}$, we find by (2.4) and *Vandermonde's formula* that

$$
a_n^{(k)} = \frac{k}{n+1} \;,
$$

which is intuitively clear, since $\phi = \frac{1}{2}$ means that we have the uniform distribution over $[0,1]$, and a random sample of n points will typically dissect the interval $[0,1]$ into $n+1$ subintervals of length $\frac{1}{n+1}$.

The other limiting case $\phi = 0$ means that we may choose only the points 0 or 1, both with probability $\frac{1}{2}$. We find

$$
a_n^{(k)} = 2^{-n}\sum_{i=0}^{k-1}\binom{n}{i} \;,
$$

which can easily be proved from (2.4) by induction and has also an obvious interpretation.

Acknowledgement

The second author wishes to thank the Centre for Applicable Analysis and Number Theory, University of Witwatersrand, for the opportunity to visit in September 1994. The research for this article was carried out during this period.

References

Abramowitz, M. and I.A. Stegun (1972), *Handbook of Mathematical Functions*, (Dover, New York).

David, H.A. (1981), *Order Statistics* (Wiley, New York, 2nd ed.).

Flajolet, P. and R. Sedgewick (1994), Mellin transforms and asymptotics: Finite differences and Rice's integral, *Theoret. Comput. Sci.,* to appear.

Hosking, J.R.M. (1994), Moments of order statistics of the Cantor distribution, *Statist. Prob. Lett.* **19**, 161-165.

Knuth, D.E. (1973), *The Art of Computer Programming*, Vol III (Addison-Wesley, Reading MA, 2nd ed.).

Prodinger, H. (1993), How to select a loser, *Discrete Math.* **120**, 149-159.

Szpankowski, W. (1988), The evaluation of an alternative sum with applications to the analysis of some data structures, *Inform. Proc. Lett.* **28**, 13-19.