Probability in the Engineering and Informational Sciences, **26**, 2012, 561–568.
doi:10.1017/S0269964812000186

ANALYTIC METHODS FOR SELECT SETS

J. GAITHER
Department of Mathematics
Purdue University
West Lafayette
IN 47907
E-mail: jgaither@math.purdue.edu

M.D. WARD
Department of Statistics
Purdue University
West Lafayette
IN 47907
E-mail: mdw@purdue.edu

We analyze the asymptotic number of items chosen in a selection procedure. The procedure selects items whose rank among all previous applicants is within the best $100p$ percent of the number of previously selected items. We use analytic methods to obtain a succinct formula for the first-order asymptotic growth of the expected number of items chosen by the procedure.

1. INTRODUCTION

This study responds to Krieger, Pollak, and Samuel-Cahn [2], which analyzes a selection rule in which a number of items are sequentially observed. Some of the items are retained; the others are permanently discarded. None are revisited. The values of the first n items are random variables $X_1, X_2, \ldots, X_n$, such that the $n!$ orderings are equally likely (no ties allowed). The selection procedure only utilizes the relative rank of the random variables. The random variable of interest is L_n, the number of the first n items that are retained.

 0269-9648/12 $25.00

As in [2], "'better' is equivalent to 'smaller'". Inheriting their notation, we let R_i^n be the rank of the ith item among the first n items, that is,

$$R_i^n := \sum_{j=1}^{n} I\{X_j \le X_i\} = \#\{j \mid X_j \le X_i, \text{with } 1 \le j \le n\},$$

where $I\{A\}$ is an indicator for event A. The first item is always retained, so $L_1 = 1$. For $n \ge 2$, the nth item is retained if its rank among the first n applicants is within the best $100p$ percent of L_{n-1}, that is, if $R_n^n \le \lceil pL_{n-1} \rceil$. (The value $0 < p \le 1$ is fixed.) We illustrate the first few cases

1. Since $L_1 = 1$, and $\lceil pL_1 \rceil = 1$, item 2 is retained iff $R_2^2 = 1$, that is, when $X_2 < X_1$. So $P(L_2 = 2) = P(L_2 = 1) = 1/2$.

2a. If $L_2 = 1$, we have $\lceil pL_2 \rceil = 1$, so item 3 is retained iff $R_3^3 = 1$, that is, if $X_3 < \min\{X_1, X_2\}$. So $P(L_3 = 2 \mid L_2 = 1) = 1/3$, and $P(L_3 = 1 \mid L_2 = 1) = 2/3$.

2b. If $L_2 = 2$:
   (a) For $0 < p \le 1/2$, we have $\lceil pL_3 \rceil = 1$, so item 3 is retained iff $R_3^3 = 1$, that is, if $X_3 < \min\{X_1, X_2\}$. So $P(L_3 = 3 \mid L_2 = 2) = 1/3$ and $P(L_3 = 2 \mid L_2 = 2) = 2/3$.
   (b) For $1/2 < p \le 1$, we have $\lceil pL_3 \rceil = 2$, so item 3 is retained iff R_3^3 is 1 or 2, that is, if $X_3 \not> \max\{X_1, X_2\}$. So $P(L_3 = 3 \mid L_2 = 2) = 2/3$ and $P(L_3 = 2 \mid L_2 = 2) = 1/3$.

Another way to view a recursive definition of the L_n's is given in (2) of Section 4.

2. MOTIVATION

The first main result proved by Krieger et al. [2] is that, for $0 < p \le 1$, there exists a constant $c_p > 0$ such that $E(L_n)/n^p \to c_p$ as $n \to \infty$ (Theorem 4.1 of [2]). Krieger et al. only state $c_1 = 1/2$; they do not give any other values of c_p. Furthermore, they state, on page 366 of [2], that "It seems impossible to determine c_p analytically, except for $p = 1$." In this study, however, we accomplish this task: We use analytic methods to reveal the values c_p for all p.

Krieger et al. also used the simulation to estimate the values of c_p, but several of these estimations were inaccurate; we give precise values for all c_p in this report. When p is rational, we can use symbolic algebra to evaluate the c_p.

3. MAIN RESULTS

THEOREM 1: *As $n \to \infty$, we have $E(L_n)/n^p \to c_p$, where*

$$c_p = \frac{1 + \sum_{k \ge 1} \frac{\lceil pk \rceil - pk}{\lceil pk \rceil} \prod_{j=1}^{k} \frac{1}{1 + \frac{p}{\lceil pj \rceil}}}{(p+1)\Gamma(p+1)}. \quad \textbf{(1)}$$

TABLE 1. Some representative values of c_p

p	c_p
1	$\frac{1}{2}$
1/2	$\frac{2\sqrt{\pi}}{3}$
1/3	$\frac{\pi^2}{3(\Gamma(2/3))^2}$
2/3	$\frac{2^{1/3}\pi\sqrt{3}}{5\Gamma(2/3)}$
1/4	$\frac{\sqrt{2}\pi^3}{10(\Gamma(3/4))^3}$
3/4	$\frac{4\pi 3^{1/4}\sqrt{2}}{21\Gamma(3/4)}$
1/5	$\frac{16\pi^4}{375(\Gamma(4/5))^4(3-\sqrt{5})}$
2/5	$\frac{4\pi^{3/2}(\sqrt{5}+1)2^{3/5}\Gamma(7/10)}{7(\sqrt{5}-1)(5+\sqrt{5})(\Gamma(4/5))^2}$
3/5	$\frac{5\pi 3^{3/10}\Gamma(3/5)}{12\Gamma(8/15)\Gamma(2/3)}$
4/5	$\frac{5\Gamma(1/5)2^{2/5}}{36}$
1/6	$\frac{4\pi^5}{189(\Gamma(5/6))^5}$
5/6	$\frac{12\pi 5^{1/6}}{55\Gamma(5/6)}$

When $p \in \mathbb{Q}$, for example, $p = r/s$, then c_p has a form we can symbolically evaluate:

$$c_p = \frac{1 + \sum_{\ell \geq 0} \left(\prod_{\sigma=1}^{\ell} \mu_{r,s}(\sigma) \right) \left(\sum_{b=1}^{s-1} \nu_{r,s}(\ell, b) \right)}{(p+1)\Gamma(p+1)},$$

where $\mu_{r,s}(\sigma) = \prod_{j=1}^{s} \frac{1}{1+\frac{p}{(\sigma-1)r+\lceil pj \rceil}}$ *and* $\nu_{r,s}(\ell, b) = \frac{\lceil pb \rceil - pb}{r\ell + \lceil pb \rceil} \prod_{i=1}^{b} \frac{1}{1+\frac{p}{r\ell+\lceil pi \rceil}}$.

This theorem yields succinct values of c_p. To demonstrate the intimate relation to the Gamma function, we list several c_p's in Table 1.

In Table 2, we improve upon the values from Table 1 of Krieger et al. [2]. (Their values cover the case $n = 10{,}000$, and our values correspond to the asymptotic case,

TABLE 2. Values of c_p (compare with Table 1 of [2], which lists values for $n = 10{,}000$)

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
c_p	5.803	2.961	2.193	1.671	1.182	1.202	1.048	0.841	0.693	0.500

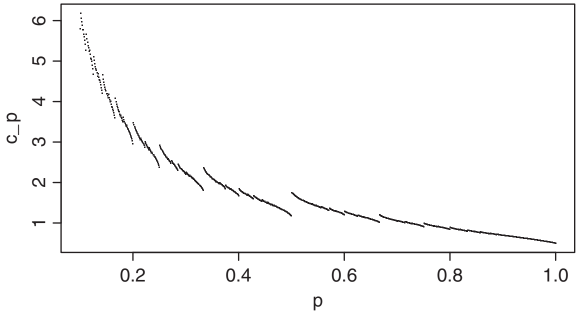

FIGURE 1. Values of c_p for $p = j/1,000$, where $100 \le j \le 1,000$.

that is, as $n \to \infty$.) In Figure 1 we graph c_p. [We conjecture c_p is continuous at each irrational p but only left-continuous (not right-continuous) at each rational p.]

4. LEMMAS AND PROOFS

The L_n's are defined recursively, as in Lemma 2.1(i) of [2]:

$$L_1 = 1 \quad \text{and} \quad L_{n+1} = \begin{cases} L_n + 1 & \text{with probability } \lceil pL_n \rceil/(n+1), \\ L_n & \text{otherwise.} \end{cases} \tag{2}$$

In particular, L_n is an integer-valued random variable with mass on $[1, n]$.

For succinctness, we use the notation

$$P_{n,k} := P(L_n = k).$$

We use generating functions as a key tool in the proofs. Thus, we define

$$g(z) = \sum_{n \ge 1} E(L_n) z^n \quad \text{and} \quad f(z) = \sum_{n \ge 1} E(\lceil pL_n \rceil - pL_n) z^n.$$

The fundamental recurrence is that $L_1 = 1$ and, for $n > 1$,

$$P_{n+1,k+1} = \frac{\lceil pk \rceil}{n+1} P_{n,k} + \left(1 - \frac{\lceil p(k+1) \rceil}{n+1}\right) P_{n,k+1}. \tag{3}$$

LEMMA 2: *For each $n \geq 1$,*

$$E(L_{n+1}) - E(L_n) = \frac{pE(L_n) + E(\lceil pL_n \rceil - pL_n)}{n+1}.$$

PROOF OF LEMMA 2: The lemma basically follows from the fundamental recurrence given in (3). The recurrence gives

$$E(L_{n+1}) - E(L_n) = \sum_k (kP_{n+1,k} - kP_{n,k}) = \frac{\sum k \left(\lceil p(k-1) \rceil P_{n,k-1} - \lceil pk \rceil P_{n,k}\right)}{n+1}.$$

We can shift the values of k by 1 in the first part, to obtain

$$E(L_{n+1}) - E(L_n) = \frac{\sum_k \left((k+1)\lceil pk \rceil P_{n,k} - k\lceil pk \rceil P_{n,k}\right)}{n+1} = \frac{\sum_k \lceil pk \rceil P_{n,k}}{n+1}.$$

The numerator is $E(\lceil pL_n \rceil)$, so the lemma follows. ■

We turn Lemma 2 into a differential equation, using generating functions. Multiplying by z^{n+1}, summing over $n \geq 1$, and differentiating yields

$$(1-z)g'(z) - 1 = (p+1)g(z) + f(z).$$

Noting that $g(0) = 0$, this differential equation has solution

$$g(z) = \frac{\int_0^z (1+f(t))(1-t)^p \, dt}{(1-z)^{p+1}}.$$

We handle $g(z)$ with analytic methods, that is, with $z \in \mathbb{C}$, as espoused in [1,3]. Since $f(t)$ has real-valued coefficients between 0 and 1, then $\int_0^z (1+f(t))(1-t)^p dt$ does not have singularities that are strictly inside the unit circle in $\mathbb{C}$. Also, $\int_0^1 (1+f(t))(1-t)^p \, dt$ is a constant (to be determined below). Thus, the singularity of $g(z)$ at $z = 1$ is a pole of order $p+1$; any other singularity located directly on the boundary of the unit circle could only be a pole of order 1 or less. Thus, the singularity at $z = 1$ completely determines the first-order asymptotic growth of the coefficients in the Maclaurin representation of $g(z)$. This is the result of Krieger et al., namely

$$E(L_n)/n^p \sim c_p,$$

but we have the additional fact that

$$c_p = \frac{\int_0^1 (1+f(t))(1-t)^p\,dt}{\Gamma(p+1)}.$$

Of course $\int_0^1 (1-t)^p\,dt = \frac{1}{p+1}$, so $c_p = \frac{1}{(p+1)\Gamma(p+1)} + \frac{\int_0^1 f(t)\,(1-t)^p\,dt}{\Gamma(p+1)}$. The Maclaurin series of $(1-t)^p$ is $(1-t)^p = \sum_{n\geq 0} \frac{\Gamma(n-p)}{\Gamma(-p)\Gamma(n+1)} t^n$, and thus

$$f(t)(1-t)^p = \sum_{k\geq 1}(\lceil pk\rceil - pk)\sum_{m\geq 1} P_{m,k} \sum_{n\geq 0} \frac{\Gamma(n-p)}{\Gamma(-p)\Gamma(n+1)} t^{n+m}.$$

Next we evaluate the corresponding definite integral

$$\int_0^1 f(t)(1-t)^p\,dt = \sum_{k\geq 1}(\lceil pk\rceil - pk)\sum_{m\geq 1} P_{m,k} \sum_{n\geq 0} \frac{\Gamma(n-p)}{\Gamma(-p)\Gamma(n+1)} \frac{1}{n+m+1}.$$

To simplify, we note

$$\sum_{n\geq 0} \frac{\Gamma(n-p)}{\Gamma(-p)\Gamma(n+1)} \frac{1}{n+m+1} = \frac{m!\Gamma(p+1)}{\Gamma(m+p+2)},$$

and thus

$$c_p = \frac{1}{(p+1)\Gamma(p+1)} + \sum_{k\geq 1}(\lceil pk\rceil - pk)\sum_{m\geq 1} \frac{m! P_{m,k}}{\Gamma(m+p+2)}. \tag{4}$$

LEMMA 3: *For $k > 1$,*

$$P_{m,k} = \lceil p(k-1)\rceil \sum_{n<m} \frac{P_{n,k-1}}{n+1} \prod_{\ell=n+2}^{m} \left(1 - \frac{\lceil pk\rceil}{\ell}\right).$$

PROOF OF LEMMA 3: If $L_m = k$, there must be a *largest* value $n < m$ such that $L_n = k-1$. Since n is the largest such value, $L_\ell = k$ for $n < \ell \leq m$. Thus

$$\begin{aligned} P_{m,k} &= P(L_m = k) \\ &= \sum_{n<m} P(L_n = k-1) P(L_{n+1} = L_{n+2} = \cdots = L_m = k \mid L_n = k-1) \\ &= \sum_{n<m} P_{n,k-1} \frac{\lceil p(k-1)\rceil}{n+1} \prod_{\ell=n+2}^{m} \left(1 - \frac{\lceil pk\rceil}{\ell}\right). \end{aligned}$$

Factoring out $\lceil p(k-1)\rceil$ completes the proof of the lemma. ∎

COROLLARY 4: *For $k > 1$, we have*

$$\sum_{m\geq 1}\frac{m!P_{m,k}}{\Gamma(m+p+2)} = \frac{\lceil p(k-1)\rceil}{\lceil pk\rceil + p}\sum_{n\geq 1}\frac{n!P_{n,k-1}}{\Gamma(n+p+2)}.$$

PROOF OF COROLLARY 4: By Lemma 3,

$$\begin{aligned}\sum_{m\geq 1}\frac{m!P_{m,k}}{\Gamma(m+p+2)} &= \sum_{m\geq 1}\frac{m!\lceil p(k-1)\rceil\sum_{n<m}\frac{P_{n,k-1}}{n+1}\prod_{\ell=n+2}^{m}\left(1-\frac{\lceil pk\rceil}{\ell}\right)}{\Gamma(m+p+2)}\\ &= \lceil p(k-1)\rceil\sum_{n\geq 1}\frac{P_{n,k-1}}{n+1}\sum_{m>n}\frac{m!\prod_{\ell=n+2}^{m}\left(1-\frac{\lceil pk\rceil}{\ell}\right)}{\Gamma(m+p+2)}\\ &= \lceil p(k-1)\rceil\sum_{n\geq 1}\frac{P_{n,k-1}}{n+1}\frac{(n+1)!}{(\lceil pk\rceil+p)\Gamma(n+p+2)}.\end{aligned}$$

This completes the proof of the corollary. ■

Applying Corollary 4, a total of $k-1$ times to (4) yields

$$c_p = \frac{1}{(p+1)\Gamma(p+1)} + \sum_{k\geq 1}(\lceil pk\rceil - pk)\left(\prod_{j=2}^{k}\frac{\lceil p(j-1)\rceil}{\lceil pj\rceil + p}\right)\sum_{m\geq 1}\frac{m!P_{m,1}}{\Gamma(m+p+2)}. \quad \textbf{(5)}$$

Simplifying, we have

$$\prod_{j=2}^{k}\frac{\lceil p(j-1)\rceil}{\lceil pj\rceil + p} = \frac{1+p}{\lceil pk\rceil}\prod_{j=1}^{k}\frac{\lceil pj\rceil}{\lceil pj\rceil + p} = \frac{1+p}{\lceil pk\rceil}\prod_{j=1}^{k}\frac{1}{1+\frac{p}{\lceil pj\rceil}}. \quad \textbf{(6)}$$

Also $L_m = 1$ iff the 2nd, 3rd, …, mth items are not retained, so

$$P_{m,1} = \prod_{j=2}^{m}\left(1-\frac{\lceil p\rceil}{j}\right) = \prod_{j=2}^{m}\left(1-\frac{1}{j}\right) = 1/m.$$

So

$$\sum_{m\geq 1}\frac{m!P_{m,1}}{\Gamma(m+p+2)} = \sum_{m\geq 1}\frac{(m-1)!}{\Gamma(m+p+2)} = \frac{1}{(p+1)^2\Gamma(p+1)}. \quad \textbf{(7)}$$

Substituting (6) and (7) into (5) gives (1), the main equation of the theorem. Finally, in the rational case, $p = r/s$, so we simplify (1) by grouping the numerator's terms according to the value of k mod s. Writing $k = \ell s + b$ yields

$$\lceil pk \rceil - pk = \lceil p(\ell s + b) \rceil - p(\ell s + b) = r\ell + \lceil pb \rceil - r\ell - pb = \lceil pb \rceil - pb.$$

So

$$\sum_{k \geq 1} \frac{\lceil pk \rceil - pk}{\lceil pk \rceil} \prod_{j=1}^{k} \frac{1}{1 + \frac{p}{\lceil pj \rceil}} = \sum_{b=1}^{s-1} \sum_{\ell \geq 0} \frac{\lceil pb \rceil - pb}{r\ell + \lceil pb \rceil} \prod_{j=1}^{\ell s + b} \frac{1}{1 + \frac{p}{\lceil pj \rceil}}$$

and

$$\begin{aligned} \prod_{j=1}^{\ell s + b} \frac{1}{1 + \frac{p}{\lceil pj \rceil}} &= \left(\prod_{j=1}^{\ell s} \frac{1}{1 + \frac{p}{\lceil pj \rceil}} \right) \left(\prod_{i=\ell s+1}^{\ell s + b} \frac{1}{1 + \frac{p}{\lceil pi \rceil}} \right) \\ &= \left(\prod_{\sigma=1}^{\ell} \prod_{j=1}^{s} \frac{1}{1 + \frac{p}{(\sigma-1)r + \lceil pj \rceil}} \right) \left(\prod_{i=1}^{b} \frac{1}{1 + \frac{p}{r\ell + \lceil pi \rceil}} \right). \end{aligned}$$

Defining $\mu_{r,s}(\sigma)$ and $\nu_{r,s}(\ell, b)$ as in the theorem statement, and substituting, yields Theorem 1.

Acknowledgements

M.D. Ward's research is supported by NSF Science & Technology Center for Science of Information Grant CCF-0939370. We sincerely thank A.M. Krieger, M. Pollak, and E. Samuel-Cahn, for their original study. We are indebted to the editor, Sheldon Ross, for his gracious guidance and feedback; we also thank one anonymous reviewer. We thank E.H. Goins, G. Louchard, H.M. Mahmoud, and H. Prodinger for brief advice, and A. Davidson and M. Gopaladesikan, for early discussions.

References

1. Flajolet, P. & Sedgewick R. (2009). *Analytic combinatorics*. Cambridge: Cambridge University Press.
2. Krieger, A.M., Pollak, M., & Samuel-Cahn, E. (2007). Select sets: Rank and file. *Annals of Applied Probability*, 17: 360–385.
3. Szpankowski, W. (2001). *Average case analysis of algorithms on sequences*. New York: Wiley.