The average position of the first maximum in a sample of geometric random variables

Margaret Archibald[1] and Arnold Knopfmacher[2]

[1]Department of Mathematics and Applied Mathematics,
University of Cape Town,
South Africa

[2]School of Mathematics,
University of the Witwatersrand,
Johannesburg,
South Africa

International Conference on Infinity in Logic & Computation,
University of Cape Town, South Africa, 3–5 November 2007

Introduction

Samples: $(\Gamma_1, \Gamma_2, \ldots \Gamma_n)$ where $\mathbb{P}\{\Gamma_j = i\} = pq^{i-1}$, for $1 \leq j \leq n$, with $p + q = 1$.

Example

If $p = q = \frac{1}{2}$, then

- the probability of a 1 is $\frac{1}{2}(\frac{1}{2})^0 = \frac{1}{2}$
- the probability of a 4 is $\frac{1}{2}(\frac{1}{2})^3 = \frac{1}{16}$

I.e., 1 is twice as likely to occur as 2 and so on.

The alphabet is infinite.

Eg. 2113131212121

Previous Work

Szpankowski and Rego (1990): *Maximum order statistic.*

$$\mathbb{E}_n = \log_Q n + \frac{\gamma}{L} + \frac{1}{2} + P_0(\log_Q n) + O\Big(\frac{1}{n}\Big),$$

$$\mathbb{V}_n = \frac{\pi^2}{6L^2} + \frac{1}{12} + P_1(\log_Q n), \qquad \Big(Q = \frac{1}{q}, \quad L = \log Q\Big).$$

Previous Work

Szpankowski and Rego (1990): *Maximum order statistic.*

$$\mathbb{E}_n = \log_Q n + \frac{\gamma}{L} + \frac{1}{2} + P_0(\log_Q n) + O\left(\frac{1}{n}\right),$$

$$\mathbb{V}_n = \frac{\pi^2}{6L^2} + \frac{1}{12} + P_1(\log_Q n), \qquad \left(Q = \frac{1}{q}, \quad L = \log Q\right).$$

Kirschenhofer and Prodinger (1993): *Average dth largest element.*

As $n \to \infty$, $\quad \left(H_n = \sum_{k=1}^{n} \frac{1}{k} \quad \text{and} \quad H_n^{(2)} = \sum_{k=1}^{n} \frac{1}{k^2}\right)$

$$\mathbb{E}_n^{(d)} \sim \log_Q n + \frac{\gamma}{L} + \frac{1}{2} - \frac{1}{L} H_{d-1} + P_2(\log_Q n),$$

$$\mathbb{V}_n^{(d)} \sim \frac{\pi^2}{6L^2} + \frac{1}{12} - \frac{1}{L^2} H_{d-1}^{(2)} - [P_2^2]_0 + P_3(\log_Q n).$$

Note: $[P_2^2]_0$ is the mean of the square of the fluctuations of the expectation.

Previous Work

Problem: Råde (1991); Solution: Griffin and Lossers (1994)

Toss coins until all show heads. ‘Heads’ occurs with probability p.

Previous Work

Problem: Råde (1991); Solution: Griffin and Lossers (1994)

Toss coins until all show heads. 'Heads' occurs with probability p.

Also on this topic...

- Eisenberg, Stengle and Strang (1993)
- Brands, Steutel and Wilms (1994)
- Baryshnikov, Eisenberg and Stengle (1995)

Previous Work

Problem: Råde (1991); Solution: Griffin and Lossers (1994)

Toss coins until all show heads. 'Heads' occurs with probability p.

Also on this topic...

- Eisenberg, Stengle and Strang (1993)
- Brands, Steutel and Wilms (1994)
- Baryshnikov, Eisenberg and Stengle (1995)

Kirschenhofer and Prodinger (1996)

Number of winners.
(Also generalised to d below max.)

Previous Work

Skip lists (see Pugh 1990)

Kirschenhofer and Prodinger (1994): Path length of skip lists.

Prodinger (1996)

Left-to-right maxima (strict and weak, e.g., 11213321).

Louchard and Prodinger (2006)

Found sth factorial moment for the number of elements a below (resp. above/at) the level of the $k+1$th maximum.

Method

Symbolic Expression

$$\{1,2,\ldots,k-1\}^* k \{1,2,\ldots,k\}^*$$

Bivariate generating function

$$F(z,u) := \sum_{k \geq 1} \frac{pq^{k-1}z}{(1-zu(1-q^{k-1}))(1-z(1-q^k))}$$

"Position" $\equiv$ Number of places *before* the first maximum occurs.

Method

Symbolic Expression

$$\{1, 2, \ldots, k-1\}^* k \{1, 2, \ldots, k\}^*$$

Bivariate generating function

$$F(z,u) := \sum_{k \geq 1} \frac{pq^{k-1}z}{(1 - zu(1-q^{k-1}))(1 - z(1-q^k))}$$

"Position" $\equiv$ Number of places *before* the first maximum occurs.

Definition - Mean and variance of PGFs

Given a PGF $P(u) = \sum_{k \geq 0} p_k u^k$ ($p_k = Pr\{X = k\}$) for a random variable X:

- $P'(1)$ gives the expected value of X
- $P''(1) + P'(1) - P'(1)^2$ gives the variance.

Method - The Expected Value

Partial fractions on...

$$\frac{\partial}{\partial u} F(z,u)|_{u=1} = \sum_{k\geq 1} \frac{pq^{k-1}(1-q^{k-1})z^2}{(1-z(1-q^k))(1-z(1-q^{k-1}))^2}$$

...leads to...

$$[z^n]\frac{\partial}{\partial u} F(z,u)|_{u=1}$$
$$= \frac{1}{p} \sum_{k\geq 2} \left[(q^{1-k}-1)(1-q^k)^n - (q^{1-k}+q+n-1)(1-q^{k-1})^n \right].$$

Use Binomial Theorem and get rid of sum on k:

$$[z^n]\frac{\partial}{\partial u}F(z,1) = \sum_{i=2}^{n}\binom{n}{i}(-1)^i\frac{q^{i-1}(q^i - 1 + nq^i - nq)}{(1-q^{i-1})(1-q^i)} + \frac{qn(n-1)}{p}$$

Method - Expectation

Use Binomial Theorem and get rid of sum on k:

$$[z^n]\frac{\partial}{\partial u}F(z,1) = \sum_{i=2}^{n}\binom{n}{i}(-1)^i\frac{q^{i-1}(q^i-1+nq^i-nq)}{(1-q^{i-1})(1-q^i)}+\frac{qn(n-1)}{p}$$

Consequently, $(Q = q^{-1})$

$$[z^n]\frac{\partial}{\partial u}F(z,u)|_{u=1} = -\underbrace{\sum_{i=2}^{n}\binom{n}{i}(-1)^i\frac{1}{Q^{i-1}-1}}_{\alpha}$$

$$-n\underbrace{\sum_{i=2}^{n}\binom{n}{i}(-1)^i\frac{1}{Q^i-1}}_{\beta}+\frac{n(n-1)}{Q-1}.$$

Method - Expectation - Rice's Integrals

Let C be a curve surrounding the points $1, 2, \ldots, n$ in the complex plane, and let $f(z)$ be analytic inside C. Then

$$\sum_{k=1}^{n} \binom{n}{k} (-1)^k f(k) = -\frac{1}{2\pi i} \int_C [n;z] f(z) dz,$$

where

$$[n;z] = \frac{(-1)^{n-1} n!}{z(z-1)\cdots(z-n)} = \frac{\Gamma(n+1)\Gamma(-z)}{\Gamma(n+1-z)}. \qquad \square$$

By extending the contour of integration, we can express (see Flajolet and Sedgewick (1995)) the asymptotic expansion as

$$\sum \mathrm{Res}([n;z]f(z)) + \text{smaller order terms},$$

where the sum is taken over all poles different from $1, \ldots, n$.

As an example, consider $(Q := \frac{1}{q})$

$$\alpha = \sum_{i=2}^{n} \binom{n}{i} (-1)^i \frac{1}{Q^{i-1} - 1}.$$

Integrand in question: $[n; z] f(z) = \frac{\Gamma(n+1)\Gamma(-z)}{\Gamma(n+1-z)} \frac{1}{Q^{z-1}-1}$.
Double pole: $z = 1$;
Simple poles: $z = 0$ and $z = \chi_k + 1$ $\left(\chi_k = \frac{2k\pi i}{\log Q} \text{ for } k \in \mathbb{Z} \backslash \{0\}\right)$.

Result - Expectation

Theorem 1.

The average position E_n of the first (left-most) occurrence of the maximum in a sample of geometric random variables is given by

$$E_n = n\Big(\frac{1}{L} + \frac{1}{1-Q} + \delta_{E1}(\log_Q n)\Big) + \frac{1}{L} + \frac{Q}{1-Q} - \delta_{E2}(\log_Q n) + o(1)$$

where $Q = \frac{1}{q}$; $L = \log Q$, $\chi_k = 2k\pi i/L$,

$$\delta_{E1}(x) := \frac{1}{L}\sum_{k\neq 0} \chi_k \Gamma(-1-\chi_k) e^{2k\pi i x}$$

and

$$\delta_{E2}(x) := \frac{1}{2L}\sum_{k\neq 0} \chi_k (1+\chi_k)(\chi_k - 2)\Gamma(-1-\chi_k) e^{2k\pi i x}.$$

Method - Variance

Differentiating partially twice gives

$$\frac{\partial^2}{\partial u^2} F(z,u) = \sum_{k\geq 1} \frac{2pq^{k-1}z^3(1-q^{k-1})^2}{(1-z(1-q^k))(1-zu(1-q^{k-1}))^3}$$

and hence

$$\begin{aligned}
&\frac{\partial^2}{\partial u^2} F(z,1) \\
&= 2\sum_{k\geq 1}\Bigg[\frac{(q^{k-1}-1)^2}{p^2q^{2k-2}(1-z(1-q^k))} - \frac{1}{(1-z(1-q^{k-1}))^3} \\
&\qquad - \frac{q^{1-k}(1-3q^{k-1}+2q^k)}{p(1-z(1-q^{k-1}))^2} \\
&\qquad - \frac{q^{2-2k}(1+3q^{2k-2}-3q^{k-1}+q^k-3q^{2k-1}+q^{2k})}{p^2(1-z(1-q^{k-1}))}\Bigg].
\end{aligned}$$

$$[z^n]\frac{\partial^2}{\partial u^2}F(z,1) = \underbrace{\frac{2}{(Q-1)^2}\sum_{i=3}^{n}\binom{n}{i}(-1)^i\frac{Q^i-1}{1-Q^{i-2}}}_{(a)}$$

$$+\underbrace{\frac{4Q}{(Q-1)^2}\sum_{i=3}^{n}\binom{n}{i}(-1)^i\frac{Q^i-1}{Q^{i-1}-1}}_{(b)}+\underbrace{\frac{2n}{Q-1}\sum_{i=3}^{n}\binom{n}{i}(-1)^i\frac{Q^i}{1-Q^{i-1}}}_{(c)}$$

$$+\underbrace{\left(\frac{n(3Q-1)}{Q-1}-n^2\right)\sum_{i=3}^{n}\binom{n}{i}(-1)^i\frac{Q^i}{Q^i-1}}_{(d)}+\underbrace{\frac{2Q^2}{(Q-1)^2}-\frac{2nQ(Q+1)}{(Q-1)^2}}_{(e)}$$

$$\underbrace{-\frac{n^2Q(3Q^2-7Q-6)}{2(Q-1)^2(Q+1)}+\frac{n^3Q(2Q^2-2Q-1)}{(Q-1)^2(Q+1)}-\frac{n^4Q^2}{2(Q^2-1)}}_{(e)}.$$

Hence the main terms of the second factorial moment are:

$$n^2 \frac{2L + 3 - 4Q + Q^2}{2L(Q-1)^2} + n \frac{8QL - 2L - 5Q^2 + 4Q + 1}{2L(Q-1)^2} + \frac{2Q^2L - 3Q^2 - 1 + 4Q}{L(Q-1)^2}.$$

We also consider (among other things) the constant terms which arise from squaring the fluctuations from the expectation. Recall

$$\delta_{E1}(x) := \frac{1}{L} \sum_{k \neq 0} \chi_k \Gamma(-1 - \chi_k) e^{2k\pi i x}, \quad \text{where} \quad \chi_k = 2k\pi i / L.$$

Hence, we require

$$\frac{1}{L^2} \sum_{k \neq 0} \chi_k (-\chi_k) \Gamma(-1 - \chi_k) \Gamma(-1 + \chi_k).$$

Method - Variance - Squaring the Fluctuations

To find the 'zeroth' fourier coefficient of the square, we use a method devised by Prodinger in 2004. Consider the function

$$I_1 := \frac{1}{2\pi i}\int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} F(z)dz$$

where

$$F(z) := \frac{-L}{e^{Lz}-1} z^2 \Gamma(-1-z)\Gamma(-1+z).$$

Express I_1 in two ways:

- Shift contour line left and collect residues
- Sum negative residues right of the line $\Re z = \frac{1}{2}$

Method - Variance - Squaring the Fluctuations

The residues for the simple poles at $z = 0$ and $z = \chi_k$, $k \neq 0$ can be calculated in order to write

$$I_1 = \frac{1}{2\pi i} \int_{(-\frac{1}{2})} F(z)dz + 1 + \sum_{k \neq 0} \chi_k(-\chi_k)\Gamma(-1-\chi_k)\Gamma(-1+\chi_k).$$

Now rewrite $-\frac{1}{e^{Lz}-1}$ as $1 + \frac{1}{e^{-Lz}-1}$ and use the change of variable $z := -z$ to get

$$I_1 = LI_2 - I_1 + 1 + \sum_{k \neq 0} \chi_k(-\chi_k)\Gamma(-1-\chi_k)\Gamma(-1+\chi_k),$$

where

$$I_2 = \frac{1}{2\pi i} \int_{(-\frac{1}{2})} z^2\Gamma(-1-z)\Gamma(-1+z)dz \qquad \left(= -\log 2 + \frac{L}{2} \right).$$

Shifting I_1 right gives

$$I_1 = \frac{L}{4(1-Q)} + \frac{QL^2}{2(Q-1)^2} + L\sum_{l\geq 2}\frac{l(-1)^l}{(Q^l-1)(l+1)(l-1)}.$$

Equating the two expressions for I_1 leaves us with

$$\sum_{k\neq 0}\chi_k(-\chi_k)\Gamma(-1-\chi_k)\Gamma(-1+\chi_k)$$
$$= \frac{L}{2(1-Q)} + \frac{QL^2}{(Q-1)^2} + 2L\sum_{l\geq 2}\frac{l(-1)^l}{(Q^l-1)(l+1)(l-1)}$$
$$+ L\log 2 - \frac{L}{2} - 1.$$

which is what we need apart from a factor of L^{-2}, $(L = \log Q)$.

Method - Variance - Fluctuations

It is again possible to find the fluctuations explicitly. For the fluctuations involving the largest term n^2 of the variance, we consider two sources.

- The usual method seen in the expectation applied to the second factorial moment.
- Squaring the expectation fluctuations.

Result - Variance

Theorem 2.

The variance of the position of the first occurrence of the maximum in a sample of geometric random variables is given by

$$Var = n^2\left(\frac{1+Q}{2L(Q-1)} - \frac{1}{L^2}\right) + n\left(\frac{Q}{(Q-1)^2} - \frac{2}{L^2} + \frac{Q+1}{2L(Q-1)}\right) + \frac{Q}{(Q-1)^2} - \frac{1}{L^2} + o(1).$$

There are also negligibly small contributions from the fluctuating terms.

Result - Variance

Theorem 2 continued.

The contributions from fluctuating terms to order n^2 are

$$\frac{Qn^2}{2L(1-Q)} + \frac{Qn^2}{(Q-1)^2} + \frac{2n^2}{L}\sum_{l\geq 2}\frac{l(-1)^l}{(Q^l-1)(l+1)(l-1)} + \frac{\log 2n^2}{L} - \frac{n^2}{L^2}$$

and the fluctuations are

$$\delta_v(x) := \frac{-n^2}{L}\sum_{k\neq 0} V_k e^{2k\pi i x},$$

where V_k is given by

$$\frac{\Gamma(-2-\chi_k)}{L(Q-1)}\big(-L(Q+1) - 4\chi_k(L+Q-1) + \chi_k^2(2-2Q+QL-L)\big)$$
$$-\sum_{l\geq 1}\frac{l(-1)^l}{(l+1)!}(l-\chi_k)\Gamma(l-1-\chi_k)\frac{Q^l+1}{Q^l-1}.$$

Note - Variance

As $Q \longrightarrow 1$

As $q \longrightarrow 1$ samples of geometric variables tend in behaviour to that of a permutation of n numbers.

For permutations, the average number of places before the maximum and the second moment are

$$\frac{1}{n}\sum_{k=1}^{n}(k-1) = \frac{n-1}{2} \quad \text{and} \quad \frac{1}{n}\sum_{k=1}^{n}(k-1)^2 = \frac{n^2}{3} - \frac{n}{2} + \frac{1}{6}$$

respectively.

Thus the variance is $\frac{n^2}{12} - \frac{1}{12}$, which is also obtained by taking the limit as $Q \longrightarrow 1$ of the main term in the variance in Theorem 2.

Thank you.