Combinatorics, Probability and Computing (2014) 23, 670–685.
doi:10.1017/S096354831400011X

The Largest Missing Value in a Sample of Geometric Random Variables

MARGARET ARCHIBALD[1†] and ARNOLD KNOPFMACHER[2‡]

[1]The John Knopfmacher Centre for Applicable Analysis and Number Theory, University of the Witwatersrand, PO Wits, 2050 Johannesburg, South Africa
(e-mail: zigarch@gmail.com)

[2]The John Knopfmacher Centre for Applicable Analysis and Number Theory,
University of the Witwatersrand,
PO Wits, 2050 Johannesburg, South Africa
(e-mail: arnold.knopfmacher@wits.ac.za)

Received 27 August 2012; revised 10 June 2013; first published online 22 May 2014

We consider samples of n geometric random variables with parameter $0 < p < 1$, and study the largest missing value, that is, the highest value of such a random variable, less than the maximum, that does not appear in the sample. Asymptotic expressions for the mean and variance for this quantity are presented. We also consider samples with the property that the largest missing value and the largest value which does appear differ by exactly one, and call this the LMV property. We find the probability that a sample of n variables has the LMV property, as well as the mean for the average largest value in samples with this property. The simpler special case of $p = 1/2$ has previously been studied, and verifying that the results of the present paper coincide with those previously found for $p = 1/2$ leads to some interesting identities.

2010 *Mathematics subject classification*: Primary 60C05
Secondary 05A16

1. Introduction

In 1985, Flajolet and Martin wrote a classic paper on probabilistic counting (see [5]). They estimated the number of distinct elements in a large multiset by studying the size of least missing value in a geometric sample, when $p = 1/2$. In this paper we approximate the size of the largest missing value in a geometric sample for arbitrary $0 < p < 1$, which

† This work is based upon research supported by the National Research Foundation.
‡ This material is based upon work supported by the National Research Foundation under grant number 2053740.

could in turn also be used as a way to estimate the number of distinct elements in a large multiset.

Over the past few years, there have been various papers on the notion of 'gaps' and 'missing values' in geometrically distributed sequences of random variables (see [8], [10] and [7]). Initially, researchers were interested in counting 'gap-free' samples, and this sparked an interest in the number of gaps or missing values in a sample. Aside from its intrinsic interest, gaps have also found applications in the analysis of algorithms: see [4] and [11]. The maximum value of samples of geometric random variables has also attracted the attention of researchers, and has led to publications such as [15], [9], [13], and [1].

In this paper we continue in this vein, and investigate firstly the mean and variance of the *largest missing value* in a geometrically distributed sequence. Let the sample of geometric random variables be given by $(\Gamma_1, \Gamma_2, \ldots \Gamma_n)$, where $\mathbb{P}\{\Gamma_j = i\} = pq^{i-1}$, for $1 \leqslant j \leqslant n$, $i \in \mathbb{N}$ (0 is not included), and with $p + q = 1$. Then we want to find the largest value which does not appear in the sequence $\Gamma_1, \Gamma_2, \ldots \Gamma_n$ as long as there is at least one larger value which does appear. If the sample is 'complete' (has all values from 1 up to the maximum occurring in the sample) the largest missing value is taken to be 0.

It frequently happens, as we shall show, that when the largest missing value is non-zero, it is then exactly one smaller than than the largest value which appears in the sample. In such a case we say that the sample has the LMV property (the 'largest missing value' property).

We will determine the probability that a non-complete geometric sample has the LMV property, as well as the size of the largest value of such a sample.

Previously, these questions have been studied in the special case of $p = 1/2$ in [2], and subsequently the $p = 1/2$ case was revisited in [12]. However, neither of the methods of [2] or [12] apply to the more general case of $0 < p < 1$ as studied in the present paper.

The methods used in [12] would work for Theorems 5.1 and 6.1. However, for Theorems 3.1 and 4.1, the analysis cannot be extended, as there are no suitable expansions involving the function $\nu(k)$, the number of ones in the binary representation of integer k.

2. Notation

In this paper we use the following notation:

$$
\begin{aligned}
Q &:= 1/q, \\
L &:= \log Q, \\
\chi_k &:= \frac{2k\pi \mathrm{i}}{L} \quad \text{(where } k \in \mathbb{Z}, k \neq 0 \text{ and i denotes } \sqrt{-1}\text{)}, \\
H_k &:= \sum_{i=1}^{k} \frac{1}{i} \quad \text{(denotes the } k\text{th harmonic number)}, \\
\gamma &:= 0.577\ldots \quad \text{(denotes Euler's constant)}.
\end{aligned}
$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

2.1. Some known results

Let p_n denote the probability that a sample of n geometric random variables is complete. Then, as shown in [8], when $p = q = 1/2$,

$$p_n = \frac{1}{2} \quad \text{for every } n \geqslant 1$$

and otherwise (for $0 < q < 1$, as $n \to \infty$),

$$p_n \sim 1 - \frac{1}{L} \sum_{j \geqslant 1} p_j \frac{q^j}{j} - \delta_P(\log_Q n), \tag{2.1}$$

where

$$\delta_P(x) := \frac{1}{L} \sum_{k \neq 0} \sum_{j \geqslant 1} \frac{p_j q^j}{j!} \Gamma(\chi_k + j) e^{-2k\pi \mathrm{i} x}. \tag{2.2}$$

For convenience set

$$P(n) := \frac{1}{L} \sum_{j \geqslant 1} p_j \frac{q^j}{j} + \delta_P(\log_Q n). \tag{2.3}$$

We note that a further interesting expression for p_n in terms of certain integrals was obtained in [10], but it will not be used in this paper.

3. Largest missing value: mean

We begin by studying the expectation of the largest missing value. Since this quantity is defined to be zero for complete samples, we consider only non-complete samples below.

Theorem 3.1. *The expectation for the largest missing value in a non-complete sample of geometrically distributed variables is asymptotic to*

$$\log_Q n + 1 - \frac{\frac{1}{L^2} \sum_{i \geqslant 1} \frac{q^i p_i}{i!} \Gamma'(i) + \delta_A(\log_Q n)}{P(n)}$$

as $n \to \infty$, where the fluctuating terms are given by

$$\delta_A(x) := \frac{1}{L^2} \sum_{k \neq 0} \sum_{i \geqslant 1} \frac{q^i p_i}{i!} \Gamma'(i + \chi_k) e^{-2k\pi \mathrm{i} x}$$

and $P(n)$ is given by (2.3).

Note. Even though the numerator and denominator depend on p_i, the initial values can be computed exactly from the recursion in [8], and using these terms in the sums in Theorem 3.1 we can obtain high precision because of the geometric rate of convergence. This is also true for Theorem 4.1.

Proof. Let $C(z)$ be the exponential generating function (hereafter EGF) of complete geometrically distributed samples. That is, the coefficient of $z^n/n!$ in the expansion of

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

$C(z)$ is the probability p_n that a geometric sample of length n is complete. The EGF of geometrically distributed samples where the largest value to occur is k is then

$$T_k(z) = \prod_{i=1}^{k-1} e^{zpq^{i-1}} \big(C(zq^k) - 1\big). \tag{3.1}$$

Here $\prod_{i=1}^{k-1} e^{zpq^{i-1}}$ represents all the values smaller than k, which are allowed to occur any number of times (including 0), and $C(zq^k) - 1$ represents the values larger than k. That is, we want to have a (non-empty) complete sample of geometric random variables, but where the first or smallest value is now not 1 but $k+1$, which corresponds to the substitutions of zq^k for z in $C(z) - 1$. So, if

$$T(z) := \sum_{k \geqslant 1} k T_k(z),$$

then the average value for the largest missing value in a sample of geometric random variables is

$$\begin{aligned} t_n := n![z^n]T(z) &= n![z^n] \sum_{k \geqslant 1} k \prod_{i=1}^{k-1} e^{zpq^{i-1}} \big(C(zq^k) - 1\big) \\ &= n![z^n] \sum_{k \geqslant 1} k e^{z(1-q^{k-1})} \big(C(zq^k) - 1\big). \end{aligned}$$

Except for the case $q = 1/2$ (see [8]), there is no explicit expression for the generating function $C(z)$. Instead we will use Poissonization and Mellin transforms to find $n![z^n]T(z)$.

We start by Poissonizing and define

$$\begin{aligned} \hat{T}(z) := T(z)e^{-z} &= e^{-z} \sum_{k \geqslant 1} k e^{z(1-q^{k-1})} \big(C(zq^k) - 1\big) \\ &= \sum_{k \geqslant 1} k e^{-zq^{k-1}} \big(C(zq^k) - 1\big). \end{aligned} \tag{3.2}$$

Now, let

$$p_n := \mathbb{P}(\Gamma \in \mathcal{C})$$

be the probability that $\Gamma = \Gamma_1, \ldots, \Gamma_n$ is complete. Then from [8] we have the recurrence

$$p_n = \begin{cases} 1 & \text{if } n = 0, \\ \sum_{k=0}^{n-1} p_k \binom{n}{k} q^k p^{n-k} & \text{if } n \geqslant 1. \end{cases}$$

Then if $\hat{C}(z)$ is the Poisson transform of (p_n), we have

$$\hat{C}(z) = \sum_{n \geqslant 0} p_n \frac{z^n}{n!} e^{-z}.$$

For convenience, let

$$Q(z) = \hat{C}(z) - 1 = e^{-z} \sum_{n \geqslant 0} p_n \frac{z^n}{n!} - 1.$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

So we have (from (3.2))

$$\hat{T}(z) := T(z)e^{-z} = \sum_{k\geqslant 1} k\left(e^{-zpq^{k-1}}(1+Q(zq^k)) - e^{-zq^{k-1}}\right). \tag{3.3}$$

Now, let

$$\tilde{Q}(z) := (e^{-z})^{p/q}Q(z).$$

Then

$$\sum_{k\geqslant 1} ke^{-zpq^{k-1}}Q(zq^k) = \sum_{k\geqslant 1} k\tilde{Q}(zq^k),$$

and the Mellin transform of (3.3) is

$$\begin{aligned}
&\sum_{k\geqslant 1} k\left((pq^{k-1})^{-s}\Gamma(s) + q^{-ks}\tilde{Q}^*(s) - (q^{k-1})^{-s}\Gamma(s)\right)\\
&\quad = \sum_{k\geqslant 1} k\Bigg((pq^{k-1})^{-s}\Gamma(s)\\
&\qquad + q^{-ks}\left(\sum_{i\geqslant 1}\frac{p_i}{i!}\Gamma(i+s)q^{i+s} + \Gamma(s)(q^s - q^s p^{-s})\right) - (q^{k-1})^{-s}\Gamma(s)\Bigg)\\
&\quad = \frac{q^{-s}}{(1-q^{-s})^2}\sum_{i\geqslant 1}\frac{p_i}{i!}\Gamma(i+s)q^{i+s},
\end{aligned} \tag{3.4}$$

and exists in the strip $\langle -1, 0\rangle$. By inverting the Mellin transform, we obtain an integral which we approximate by moving the contour of integration to the right. We encounter a double pole at $s = 0$, leaving a negative residue of

$$\frac{1}{L^2}\sum_{i\geqslant 1}\frac{q^i p_i}{i!}\Gamma(i)\log z + \frac{1}{L^2}\sum_{i\geqslant 1}\frac{q^i p_i}{i!}\Gamma(i)\left(L - \frac{\Gamma'(i)}{\Gamma(i)}\right).$$

De-Poissonization (as explained in [14]) implies that $t_n \sim \hat{T}(n)$, so for the main asymptotic term of t_n we have

$$\frac{1}{L^2}\sum_{i\geqslant 1}\frac{q^i p_i}{i}\log n + \frac{1}{L^2}\sum_{i\geqslant 1}\frac{q^i p_i}{i!}\Gamma(i)\left(L - \frac{\Gamma'(i)}{\Gamma(i)}\right). \tag{3.5}$$

For the fluctuations we find the negative residue of (3.4) at $s = \chi_k$ for $k \neq 0$, namely

$$\frac{1}{L^2}\sum_{k\neq 0}\sum_{i\geqslant 1}\frac{q^i z^{-\chi_k} p_i}{i!}\Gamma(i+\chi_k)\left(L + \log z - \frac{\Gamma'(i+\chi_k)}{\Gamma(i+\chi_k)}\right).$$

This implies that the fluctuating contributions to the asymptotic expansion of t_n are

$$\begin{aligned}
&\frac{\log n}{L^2}\sum_{k\neq 0}\sum_{i\geqslant 1}\frac{q^i p_i}{i!}\Gamma(i+\chi_k)e^{-2k\pi \mathrm{i}\log_Q n}\\
&\qquad + \frac{1}{L^2}\sum_{k\neq 0}\sum_{i\geqslant 1}\frac{q^i p_i}{i!}\Gamma(i+\chi_k)\left(L - \frac{\Gamma'(i+\chi_k)}{\Gamma(i+\chi_k)}\right)e^{-2k\pi \mathrm{i}\log_Q n}\\
&\quad = (\log_Q n + 1)\delta_P(\log_Q n) - \delta_A(\log_Q n).
\end{aligned} \tag{3.6}$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

The result of Theorem 3.1 follows after dividing the sum of (3.5) and (3.6) by the probability $1-p_n$ that the sample is non-complete, where by (2.1),

$$1-p_n \sim P(n) = \frac{1}{L}\sum_{j\geqslant 1} p_j \frac{q^j}{j} + \delta_P(\log_Q n). \tag{3.7}$$

This gives

$$\begin{aligned}
&\frac{\log_Q n\left(\frac{1}{L}\sum_{i\geqslant 1}\frac{q^i p_i}{i} + \delta_P(\log_Q n)\right)}{P(n)} \\
&\quad + \frac{\frac{1}{L^2}\sum_{i\geqslant 1}\frac{q^i p_i}{i}\left(L - \frac{\Gamma'(i)}{\Gamma(i)}\right) + \delta_P(\log_Q n) - \delta_A(\log_Q n)}{P(n)} \\
&= \log_Q n + 1 - \frac{\frac{1}{L^2}\sum_{i\geqslant 1}\frac{q^i p_i}{i}\frac{\Gamma'(i)}{\Gamma(i)} + \delta_A(\log_Q n)}{P(n)},
\end{aligned}$$

as claimed. □

3.1. The case $p = 1/2$

By substituting $p = 1/2$ in Theorem 3.1, we can check this against the corresponding result in [2]. That result was derived by simpler means, by using the explicit expression for $C(z)$ known for the case $p = 1/2$.

We first simplify the expression in Theorem 3.1 to (here $L = \log 2$)

$$\log_2 n + 1 - \frac{\frac{1}{L^2}\sum_{i\geqslant 1}\frac{2^{-i-1}}{i!}\Gamma'(i) + \frac{1}{L^2}\sum_{k\neq 0}\sum_{i\geqslant 1}\frac{2^{-i-1}}{i!}\Gamma'(i+\chi_k)e^{-2k\pi \mathrm{i}\log_2 n}}{\frac{1}{L}\sum_{j\geqslant 1}\frac{2^{-j-1}}{j} + \frac{1}{L}\sum_{k\neq 0}\sum_{j\geqslant 1}2^{-j-1}\frac{\Gamma(\chi_k+j)}{j!}e^{\chi_k \log(n)}}. \tag{3.8}$$

Now, we have that

$$\frac{1}{L^2}\sum_{i\geqslant 1}\frac{2^{-i-1}}{i!}\Gamma'(i) = \frac{1}{4} - \frac{\gamma}{2L}, \tag{3.9}$$

$$\begin{aligned}
\frac{1}{L^2}\sum_{k\neq 0}\sum_{i\geqslant 1}\frac{2^{-i-1}}{i!}\Gamma'(i+\chi_k)e^{-2k\pi \mathrm{i}\log_2 n} &= \sum_{k\neq 0}\frac{\Gamma(1+\chi_k)}{4\pi \mathrm{i}}e^{-2k\pi \mathrm{i}\log_2 n} \\
&= \sum_{k\neq 0}\frac{\Gamma(1+\chi_k)}{2\chi_k L}e^{-2k\pi \mathrm{i}\log_2 n} \\
&= \frac{1}{2L}\sum_{k\neq 0}\Gamma(-\chi_k)e^{2k\pi \mathrm{i}\log_2 n},
\end{aligned} \tag{3.10}$$

$$\frac{1}{L}\sum_{j\geqslant 1}\frac{2^{-j-1}}{j} = \frac{1}{2}, \tag{3.11}$$

and

$$\frac{1}{L}\sum_{k\neq 0}\sum_{j\geqslant 1}2^{-j-1}\frac{\Gamma(\chi_k+j)}{j!}e^{\chi_k \log(n)} = 0, \tag{3.12}$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

as

$$\sum_{j\geqslant 1} 2^{-j-1}\frac{\Gamma(\chi_k+j)}{j!}=0,\quad \text{for all } k\neq 0.$$

Substituting the formulae (3.9), (3.10), (3.11) and (3.12) into (3.8) gives

$$\log_2 n+1+\frac{\frac{\gamma}{2L}-\frac{1}{4}-\frac{1}{2L}\sum_{k\neq 0}\Gamma(-\chi_k)e^{2k\pi \mathrm{i}\log_2 n}}{1/2}$$
$$=\log_2 n+\frac{\gamma}{L}+\frac{1}{2}-\frac{1}{L}\sum_{k\neq 0}\Gamma(-\chi_k)e^{2k\pi \mathrm{i}\log_2 n},$$

as in [2].

4. Largest missing value: variance

The calculation of the variance is more involved than that of the mean, hence in this section we will not explicitly compute the fluctuating terms which arise. The calculation of these can be done in principle but is very tedious and the contributions, typically of order 10^{-6}, are of little numerical significance.

Theorem 4.1. *The variance for the largest missing value in a non-complete sample of geometrically distributed variables is asymptotic to* (*excluding small fluctuations of mean zero in both numerator and denominator*)

$$\frac{1}{L^2}\frac{\sum_{i\geqslant 1}\frac{1}{i!}q^ip_i\Gamma''(i)}{\sum_{i\geqslant 1}\frac{1}{i}q^ip_i}-\frac{1}{L^2}\frac{(\sum_{i\geqslant 1}\frac{1}{i!}q^ip_i\Gamma'(i))^2}{(\sum_{i\geqslant 1}\frac{1}{i}q^ip_i)^2}-[\delta_E]_0^2$$

as $n\to\infty$, *where* $[\delta_E]_0^2$ *denotes a tiny constant arising from the square of the fluctuating term of the mean value.*

Proof. Again we make use of the EGF of geometrically distributed samples for general p where the largest value to occur is k (see (3.1)):

$$T_k(z)=\prod_{i=1}^{k-1}e^{zpq^{i-1}}\big(C(zq^k)-1\big).$$

Here we want the second moment, and thus we want to find the coefficient of $n!z^n$ in

$$W(z):=\sum_{k\geqslant 1}k^2T_k(z)=\sum_{k\geqslant 1}k^2e^{z(1-q^{k-1})}\big(C(zq^k)-1\big).$$

Poissonizing $W(z)$ gives

$$\hat{W}(z):=W(z)e^{-z}=\sum_{k\geqslant 1}k^2e^{-zq^{k-1}}\big(C(zq^k)-1\big)$$
$$=\sum_{k\geqslant 1}k^2\big(e^{-zpq^{k-1}}(1+Q(zq^k))-e^{-zq^{k-1}}\big),\tag{4.1}$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

where

$$Q(z) = \hat{C}(z) - 1 = e^{-z} \sum_{n \geqslant 0} p_n \frac{z^n}{n!} - 1.$$

as before. The Mellin transform of this is

$$\sum_{k \geqslant 1} k^2 q^{-ks} \left((pq^{-1})^{-s} \Gamma(s) + \sum_{i \geqslant 1} \frac{p_i}{i!} \Gamma(i+s) q^{i+s} + \Gamma(s)(q^s - q^s p^{-s}) - (q^{-1})^{-s} \Gamma(s) \right)$$
$$= \frac{q^s(1+q^s)}{-(1-q^s)^3} \sum_{i \geqslant 1} \frac{p_i}{i!} \Gamma(i+s) q^{i+s}, \tag{4.2}$$

and exists in the strip $\langle -1, 0 \rangle$. We now invert the Mellin transform to get an integral which we approximate by considering the negative residues of the triple pole at $s = 0$. This gives

$$\frac{1}{L^3} \sum_{i \geqslant 1} \frac{q^i p_i}{i} (\log z)^2 + \frac{2}{L^2} \sum_{i \geqslant 1} \frac{q^i p_i}{i} \log z - \frac{2}{L^3} \sum_{i \geqslant 1} \frac{q^i p_i}{i!} \Gamma'(i) \log z$$
$$+ \frac{1}{L} \sum_{i \geqslant 1} \frac{q^i p_i}{i} - \frac{2}{L^2} \sum_{i \geqslant 1} \frac{q^i p_i}{i!} \Gamma'(i) + \frac{1}{L^3} \sum_{i \geqslant 1} \frac{q^i p_i}{i!} \Gamma''(i).$$

De-Poissonizing this leaves us with

$$\frac{1}{L^3} \sum_{i \geqslant 1} \frac{q^i p_i}{i} (\log n)^2 + \frac{2}{L^2} \sum_{i \geqslant 1} \frac{q^i p_i}{i} \log n - \frac{2}{L^3} \sum_{i \geqslant 1} \frac{q^i p_i}{i!} \Gamma'(i) \log n$$
$$+ \frac{1}{L} \sum_{i \geqslant 1} \frac{q^i p_i}{i} - \frac{2}{L^2} \sum_{i \geqslant 1} \frac{q^i p_i}{i!} \Gamma'(i) + \frac{1}{L^3} \sum_{i \geqslant 1} \frac{q^i p_i}{i!} \Gamma''(i).$$

This we divide by the probability that the sample is not complete ($1 - p_n$, but excluding the fluctuations) and subtract the square of the expectation from Theorem 3.1 to get

$$\frac{1}{L^2} \frac{\sum_{i \geqslant 1} \frac{1}{i!} q^i p_i \Gamma''(i)}{\sum_{i \geqslant 1} \frac{1}{i} q^i p_i} - \frac{1}{L^2} \frac{(\sum_{i \geqslant 1} \frac{1}{i!} q^i p_i \Gamma'(i))^2}{(\sum_{i \geqslant 1} \frac{1}{i} q^i p_i)^2} - [\delta_E]_0^2 \tag{4.3}$$

as in Theorem 4.1. The square of the fluctuating term of the mean value leads to the very small additional non-zero contribution $[\delta_E]_0^2$. □

Remark. In the simpler case of $p = 1/2$, the tiny additional term $[\delta_E]_0^2$ was computed explicitly in [2] and was shown to be of magnitude 10^{-12}. By including this known tiny term from the $p = \frac{1}{2}$ case, we have an identity (proved in the subsection below) to show that this result corresponds to the variance in [2].

4.1. Identity for $p = 1/2$

Here we prove that for $p = 1/2$ the result in Theorem 4.1 corresponds to the equivalent result in [2]. That is, we want to simplify the expression in (4.3), after substituting in

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

$p = \frac{1}{2}$, which is

$$\frac{1}{L^2}\frac{\sum_{i\geqslant 1}\frac{1}{i!}2^{-i-1}\Gamma''(i)}{\sum_{i\geqslant 1}\frac{1}{i}2^{-i-1}} - \frac{1}{L^2}\frac{(\sum_{i\geqslant 1}\frac{1}{i!}2^{-i-1}\Gamma'(i))^2}{(\sum_{i\geqslant 1}\frac{1}{i}2^{-i-1})^2} - [\delta_E]_0^2, \tag{4.4}$$

and prove that it equals

$$1 + \frac{2}{L}\sum_{h\geqslant 1}\frac{(-1)^{h-1}}{h(2^h-1)}$$

as in Theorem 2 of [2]. First note that

$$\sum_{i\geqslant 1}\frac{2^{-i-1}}{i} = \frac{L}{2}. \tag{4.5}$$

Now, for the first sum in (4.4) we have that

$$\begin{aligned}\Gamma''(i) &= \Gamma'(i)(-\gamma + H_{i-1}) + \Gamma'(i)\frac{d}{dx}\left(\sum_{k=1}^{\infty}\left(\frac{1}{k} - \frac{1}{x+k-1}\right)\right)\Big|_{x=i}\\ &= \Gamma(i)(-\gamma + H_{i-1})^2 + \Gamma(i)\sum_{k=1}^{\infty}\frac{1}{(i+k-1)^2}\\ &= \Gamma(i)\left(\gamma^2 - 2\gamma H_{i-1} + H_{i-1}^2 + \sum_{k=1}^{\infty}\frac{1}{(i+k-1)^2}\right),\end{aligned}$$

and consequently

$$\begin{aligned}\sum_{i\geqslant 1}\frac{1}{i!}2^{-i-1}\Gamma''(i) &= \sum_{i\geqslant 1}2^{-i-1}\frac{1}{i}\left(\gamma^2 - 2\gamma H_{i-1} + H_{i-1}^2 + \sum_{k=1}^{\infty}\frac{1}{(i+k-1)^2}\right)\\ &= \frac{\gamma^2}{2}\log 2 - \frac{\gamma}{2}\log^2 2 + \frac{1}{2}\sum_{i\geqslant 1}\frac{1}{i2^i}H_{i-1}^2 + \frac{1}{2}\sum_{i\geqslant 1}\frac{1}{i2^i}\sum_{j=i}^{\infty}\frac{1}{j^2}.\end{aligned}$$

Now,

$$H_{m-1}^2 = \sum_{i=1}^{m-1}\frac{1}{i}\sum_{j=1}^{m-1}\frac{1}{j} = \sum_{i=1}^{m-1}\frac{1}{i^2} + 2\sum_{j=1}^{m-1}\sum_{i=1}^{j-1}\frac{1}{ij},$$

so

$$\begin{aligned}&\sum_{m\geqslant 1}\frac{1}{m2^m}H_{m-1}^2\\ &\quad= \sum_{m\geqslant 1}\frac{1}{m2^m}\left(\sum_{i=1}^{m-1}\frac{1}{i^2} + 2\sum_{j=1}^{m-1}\sum_{i=1}^{j-1}\frac{1}{ij}\right)\\ &\quad= \sum_{i\geqslant 1}\frac{1}{i^2}\sum_{m\geqslant i+1}\int_0^{\frac{1}{2}}t^{m-1}dt + 2\sum_{j\geqslant 1}\frac{H_{j-1}}{j}\sum_{m\geqslant j+1}\int_0^{\frac{1}{2}}t^{m-1}dt\\ &\quad= \int_0^{\frac{1}{2}}\frac{1}{1-t}\sum_{i\geqslant 1}\frac{t^i}{i^2}dt + 2\int_0^{\frac{1}{2}}\frac{1}{1-t}\sum_{j\geqslant 1}\frac{t^jH_{j-1}}{j}dt\end{aligned}$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

$$= \left[\sum_{i\geqslant 1}\frac{t^i}{i^2}\log\left(\frac{1}{1-t}\right)\right]_0^{\frac{1}{2}} - \int_0^{\frac{1}{2}}\frac{\log^2(1-t)}{t}dt + 2\int_0^{\frac{1}{2}}\frac{1}{1-t}\left(\log\frac{1}{1-t}\right)^2 dt$$
$$= \frac{1}{12}\left(\pi^2 L + 2L^3 - 3\zeta(3)\right).$$

Next

$$\sum_{i\geqslant 1}\frac{1}{i2^i}\sum_{j=i}^{\infty}\frac{1}{j^2} = \sum_{j=1}^{\infty}\frac{1}{j^2}\sum_{i=1}^{j}\frac{1}{i2^i}$$
$$= \sum_{j=1}^{\infty}\frac{1}{j^2}\int_0^{\frac{1}{2}}\frac{1-t^j}{1-t}dt$$
$$= \sum_{j=1}^{\infty}\frac{1}{j^2}\int_0^{\frac{1}{2}}\frac{1}{1-t}dt - \sum_{j=1}^{\infty}\frac{1}{j^2}\int_0^{\frac{1}{2}}\frac{t^j}{1-t}dt$$
$$= \frac{1}{12}\left(\pi^2 L + 2L^3 + 3\zeta(3)\right).$$

For the second sum in (4.4),

$$\sum_{i\geqslant 1}\frac{1}{i!}2^{-i-1}\Gamma'(i) = \frac{1}{2}\sum_{i\geqslant 1}\frac{(1/2)^i}{i}(-\gamma + H_{i-1}) = -\frac{\gamma L}{2} + \frac{1}{2}\sum_{i\geqslant 1}\frac{(1/2)^i}{i}H_{i-1}.$$

Now

$$\sum_{j\geqslant 1}x^j H_j = \frac{1}{1-x}\log\frac{1}{1-x},$$

so

$$\sum_{j\geqslant 1}x^j H_{j-1} = \frac{x}{1-x}\log\frac{1}{1-x}.$$

Therefore

$$\sum_{j\geqslant 1}\frac{x^j}{j}H_{j-1} = \sum_{j\geqslant 1}H_{j-1}\int_0^x t^{j-1}dt$$
$$= \int_0^x \frac{1}{1-t}\log\frac{1}{1-t}dt$$
$$= \frac{1}{2}\left(\log\frac{1}{1-x}\right)^2.$$

Setting $x = \frac{1}{2}$,

$$\sum_{j\geqslant 1}\frac{(1/2)^j}{j}H_{j-1} = \frac{L^2}{2}.$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

Consequently, for the first two terms in (4.4) we have

$$\begin{aligned}
&\frac{1}{L^2}\frac{\sum_{i\geqslant 1}\frac{1}{i!}2^{-i-1}\Gamma''(i)}{\sum_{i\geqslant 1}\frac{1}{i}2^{-i-1}} - \frac{1}{L^2}\frac{(\sum_{i\geqslant 1}\frac{1}{i!}2^{-i-1}\Gamma'(i))^2}{(\sum_{i\geqslant 1}\frac{1}{i}2^{-i-1})^2}\\
&\quad = \frac{1}{L^2}\frac{2}{L}\left[\frac{\gamma^2 L}{2} - \frac{\gamma L^2}{2} + \frac{1}{2}\frac{1}{12}\left(\pi^2 L + 2L^3 - 3\zeta(3)\right) + \frac{1}{2}\frac{1}{12}\left(\pi^2 L + 2L^3 + 3\zeta(3)\right)\right]\\
&\qquad - \frac{1}{L^2}\left(\frac{2}{L}\right)^2\left[-\frac{\gamma L}{2} + \frac{1}{2}\frac{L^2}{2}\right]^2\\
&\quad = \frac{\pi^2}{6L^2} + \frac{1}{12}.
\end{aligned}$$

From this we must subtract

$$[\delta_E]_0^2 = \frac{\pi^2}{6L^2} - \frac{11}{12} - \frac{2}{L}\sum_{h\geqslant 1}\frac{(-1)^{h-1}}{h(2^h-1)}$$

(see [2, p. 727]), so we get the main term of the variance to be

$$\frac{\pi^2}{6L^2} + \frac{1}{12} - \left(\frac{\pi^2}{6L^2} - \frac{11}{12} - \frac{2}{L}\sum_{h\geqslant 1}\frac{(-1)^{h-1}}{h(2^h-1)}\right) = 1 + \frac{2}{L}\sum_{h\geqslant 1}\frac{(-1)^{h-1}}{h(2^h-1)},$$

as required.

5. Probability that a sample has the LMV property

Now we consider the probability that the largest missing value is one less than the largest part for general p. As observed in the Introduction, the majority of non-complete samples have this LMV property.

Theorem 5.1. *The probability that a non-complete geometric sample of length n has the LMV property is*

$$\frac{2 - \log_Q(Q^2 - Q + 1)}{P(n)} + \frac{\delta_1(\log_Q n)}{P(n)},$$

where

$$\delta_1(x) := \frac{1}{L}\sum_{k\neq 0}\left((Q^2 - Q + 1)^{\chi_k} - 1\right)\Gamma(-\chi_k)e^{2k\pi \mathrm{i}x}.$$

Proof. The generating function for all samples where the largest missing value is k and the largest part is $k+1$ is

$$\begin{aligned}
F_k(z) &= \prod_{i=1}^{k-1} e^{zpq^{i-1}}(e^{zpq^k} - 1) \qquad (5.1)\\
&= e^{zpq^k + z(1-q^{k-1})} - e^{z(1-q^{k-1})}.
\end{aligned}$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

Let

$$F(z) := \sum_{k\geqslant 1} F_k(z).$$

So the probability that the largest missing value and the largest part differ by one is

$$\begin{aligned} n![z^n]F(z) &= n![z^n]\sum_{k\geqslant 1}\left(e^{zpq^k+z(1-q^{k-1})} - e^{z(1-q^{k-1})}\right) \\ &= n![z^n]\sum_{k\geqslant 1}\sum_{j\geqslant 0}\frac{1}{j!}\left((zpq^k + z - zq^{k-1})^j - (z - zq^{k-1})^j\right) \\ &= \sum_{k\geqslant 1}\left((pq^k + 1 - q^{k-1})^n - (1-q^{k-1})^n\right) \\ &= \sum_{r=0}^{n}\binom{n}{r}(-1)^r\left((1-pq)^r - 1\right)\frac{1}{1-q^r} \end{aligned}$$

using the Binomial Theorem. This can be approximated using 'Rice's method'. This technique is briefly explained in the following lemma (see [6], [13], [14]).

Lemma 5.2. *Let $\mathcal{C}$ be a curve surrounding the points $1, 2, \ldots, n$ in the complex plane, and let $f(z)$ be analytic inside $\mathcal{C}$. Then*

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

where the kernel

$$[n;z] = \frac{(-1)^{n-1}n!}{z(z-1)\cdots(z-n)} = \frac{\Gamma(n+1)\Gamma(-z)}{\Gamma(n+1-z)}. \tag{5.2}$$

By extending the contour of integration, it turns out that under suitable growth conditions (see [6]) the asymptotic expansion of our alternating sum is given by

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

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

We use the function

$$f(z) := \frac{(1-pq)^z - 1}{1-q^z},$$

so there is only a simple pole (from the kernel) at $z = 0$ and also simple poles at $z = \chi_k$ (from $f(z)$). Expanding f gives

$$f(z) \sim \frac{1 + z\log(1-pq) - 1}{1 - (1 + z\log q)} \sim \frac{\log(1-pq)}{-\log q} = \log_Q(Q^2 - Q + 1) - 2.$$

Also,

$$[n;z] \sim -\frac{1}{z},$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

so that the residue becomes

$$2 - \log_Q(Q^2 - Q + 1). \tag{5.3}$$

For the fluctuations, we look at the simple poles at $z = \chi_k$. Let $\varepsilon = z + \chi_k$. Then

$$f(z) = \frac{(1-pq)^{\chi_k}(1-pq)^{\varepsilon} - 1}{1 - q^{\chi_k} q^{\varepsilon}} \sim \frac{(1-pq)^{\chi_k} - 1}{-\varepsilon \log q} = \frac{(Q^2 - Q + 1)^{\chi_k} - 1}{\varepsilon \log Q}$$

since $Q^{\chi_k} = 1$, and

$$[n; z] \sim n^{\chi_k} \Gamma(-\chi_k).$$

This means the fluctuations in this case are

$$\frac{1}{L} \sum_{k \neq 0} \big((Q^2 - Q + 1)^{\chi_k} - 1\big) \Gamma(-\chi_k) n^{\chi_k}. \tag{5.4}$$

The formulae in (5.3) and (5.4) give us the result that the probability of a geometric random sample having the largest part only one value away from the largest missing part is

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

We need to divide this by $1 - p_n$ (see formula (3.7)), the probability that the sample is non-complete. This concludes the proof of Theorem 5.1. □

We note that for the case $p = \frac{1}{2}$ this result agrees with the result in [2].

6. The average largest value under the LMV property

In this section we find the average largest value for samples of geometric random variables which have the LMV property. That is, if a non-complete sample has a largest part and a largest missing value differing by one, what is the average of the largest part? We find the average largest missing value in this case, and then add one to the result. The method is similar to that of the previous section, used in finding the probability that a sample has the LMV property. We consider the set of non-complete samples only.

Theorem 6.1. *The average largest value for geometric samples which have the LMV property is*

$$\log_Q n + 1 + \frac{\frac{2\gamma}{L} + \big(1 - \frac{\gamma}{L}\big) \log_Q(Q^2 - Q + 1) - \frac{1}{2} \log_Q^2(Q^2 - Q + 1) + \delta_2(\log_Q n)}{2 - \log_Q(Q^2 - Q + 1) + \delta_1(\log_Q n)},$$

where the fluctuating terms are given by $\delta_1(x)$ (*defined in Theorem 5.1*), *and*

$$\delta_2(x) := -\sum_{k \neq 0} \frac{\Gamma'(-\chi_k)}{L^2} ((Q^2 - Q + 1)^{\chi_k} - 1) e^{2k\pi \mathrm{i} x} + \sum_{k \neq 0} \frac{\Gamma(-\chi_k)}{L} \big((Q^2 - Q + 1)^{\chi_k} (\log_Q(Q^2 - Q + 1) - 1) - 1\big) e^{2k\pi \mathrm{i} x}.$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

Proof. The generating function for all samples whose largest missing value is k and whose largest part is $k+1$ is

$$F_k(z) := \prod_{i=1}^{k-1} e^{zpq^{i-1}}(e^{zpq^k}-1) = e^{zpq^k+z(1-q^{k-1})} - e^{z(1-q^{k-1})}$$

as in (5.1). If we define the function

$$G(z) := \sum_{k\geqslant 1} kF_k(z),$$

then the average largest missing value is given by

$$\begin{aligned} n![z^n]G(z) &= n![z^n]\sum_{k\geqslant 1} k\left(e^{zpq^k+z(1-q^{k-1})} - e^{z(1-q^{k-1})}\right) \\ &= \sum_{k\geqslant 1} k\left((pq^k+1-q^{k-1})^n - (1-q^{k-1})^n\right) \\ &= \sum_{r=0}^{n}\binom{n}{r}(-1)^r\left((1-pq)^r-1\right)\sum_{k\geqslant 1} kq^{r(k-1)} \\ &= \sum_{r=0}^{n}\binom{n}{r}(-1)^r\frac{(1-pq)^r-1}{(1-q^r)^2}. \end{aligned}$$

This alternating sum is again a candidate for Rice's method, and the function in question is

$$f(z) := \frac{(1-pq)^z-1}{(1-q^z)^2}.$$

The expression $[n;z]f(z)$ has a double pole at $z=0$, and a double pole at $z=\chi_k$. By expanding $[n;z]$ and $f(z)$ to two terms around $z=0$, we get

$$f(z) = \frac{e^{z\log(1-pq)}-1}{(1-e^{z\log q})^2} \sim \frac{\log(1-pq)(1+\frac{z\log(1-pq)}{2})(1-z\log q)}{zL^2},$$

and with the harmonic number $H_n = \sum_{j=1}^{n}\frac{1}{j}$,

$$[n;z] \sim -\frac{1}{z}(1+zH_n).$$

Thus the residue for $z=0$ is

$$\begin{aligned} [z^{-1}][n;z]f(z) &= \frac{-\log(1-pq)}{L^2}\left(H_n + \frac{\log(1-pq)}{2} + L\right) \\ &\sim \frac{1}{L}(2-\log_Q(Q^2-Q+1))\left(\log n + \gamma + \frac{\log(Q^2-Q+1)-\log(Q^2)}{2} + L\right), \end{aligned}$$

as $n\to\infty$. We can express this as

$$\log_Q n(2-\log_Q(Q^2-Q+1)) + \frac{2\gamma}{L} + \log_Q(Q^2-Q+1)\left(1-\frac{\gamma}{L}\right) - \frac{1}{2}\log_Q^2(Q^2-Q+1). \tag{6.1}$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

For the double pole at $z = \chi_k$, let $\varepsilon = z - \chi_k$ to get

$$f(z) := \frac{(1-pq)^{\chi_k}(1-pq)^{\varepsilon}-1}{(1-q^{\chi_k}q^{\varepsilon})^2} \sim \frac{\left((1-pq)^{\chi_k}(1+\varepsilon\log(1-pq)+\frac{\varepsilon^2\log^2(1-pq)}{2!})-1\right)(1-\varepsilon\log q)}{\varepsilon^2 L^2}$$

and (see [3])

$$[n;z] \sim \Gamma(-\chi_k)n^{\chi_k}\left[1-\psi(-\chi_k)\varepsilon+\varepsilon\log n\right].$$

So the residue is

$$\begin{aligned}[\varepsilon^{-1}][n;z]f(z) &= \frac{\Gamma(-\chi_k)n^{\chi_k}}{L^2}\big((-\psi(-\chi_k)+\log n)((1-pq)^{\chi_k}-1) \\ &\quad + (1-pq)^{\chi_k}\log(1-pq)+((1-pq)^{\chi_k}-1)L\big) \\ &= \log n\frac{\Gamma(-\chi_k)}{L^2}((Q^2-Q+1)^{\chi_k}-1)n^{\chi_k} - \frac{\Gamma'(-\chi_k)}{L^2}((Q^2-Q+1)^{\chi_k}-1)n^{\chi_k} \\ &\quad + \frac{\Gamma(-\chi_k)}{L}\big((Q^2-Q+1)^{\chi_k}\log_Q(Q^2-Q+1)-1-(Q^2-Q+1)^{\chi_k}\big)n^{\chi_k},\end{aligned} \tag{6.2}$$

since $Q^{\chi_k} = 1$. Adding (6.1) to (6.2) (summed on all non-zero k) gives

$$\begin{aligned}&\log_Q n(2-\log_Q(Q^2-Q+1)) + \frac{2\gamma}{L} \\ &+\log_Q(Q^2-Q+1)\left(1-\frac{\gamma}{L}\right) - \frac{1}{2}\log_Q^2(Q^2-Q+1) \\ &+\log n\sum_{k\neq 0}\frac{\Gamma(-\chi_k)}{L^2}((Q^2-Q+1)^{\chi_k}-1)n^{\chi_k} - \sum_{k\neq 0}\frac{\Gamma'(-\chi_k)}{L^2}((Q^2-Q+1)^{\chi_k}-1)n^{\chi_k} \\ &+\sum_{k\neq 0}\frac{\Gamma(-\chi_k)}{L}\big((Q^2-Q+1)^{\chi_k}\log_Q(Q^2-Q+1)-1-(Q^2-Q+1)^{\chi_k}\big)n^{\chi_k}.\end{aligned} \tag{6.3}$$

By conditional probability it is now necessary to divide (6.3) by $\tilde{p}_n(1-p_n)$ where (see Theorem 5.1)

$$\tilde{p}_n(1-p_n) := 2-\log_Q(Q^2-Q+1)+\delta_1(\log_Q n).$$

This gives the average largest missing value, so we add 1 to get to the result in Theorem 6.1. □

7. Concluding remarks

It is interesting to compare the average largest value for all geometric random samples with that of samples with the LMV property. If we ignore the tiny fluctuations, we have the following.

The average largest part for all samples is (see [15])

$$\log_Q n + \frac{\gamma}{L} + \frac{1}{2}.$$

https://doi.org/10.1017/S096354831400011X Published online by Cambridge University Press

The average largest part for samples with the LMV property simplifies (see Theorem 6.1) to

$$\log_Q n + 1 + \frac{\gamma}{L} + \frac{1}{2}\log_Q(Q^2 - Q + 1).$$

The difference of $\frac{1}{2} + \frac{1}{2}\log_Q(Q^2 - Q + 1)$ increases monotonically from 1 to $3/2$ as Q goes from 1 to ∞.

References

[1] Archibald, M. and Knopfmacher, A. (2007) The average position of the first maximum in a sample of geometric random variables. In *Discrete Mathematics and Theoretical Computer Science*, DMTCS Proceedings Vol. AH, pp. 269–278.

[2] Archibald, M. and Knopfmacher, A. (2011) The largest missing value in a composition of an integer. *Discrete Math.* **311** 723–731.

[3] Archibald, M., Knopfmacher, A. and Prodinger, H. (2006) The number of distinct values in a geometrically distributed sample. *Europ. J. Combin.* **27** 1059–1081.

[4] Bondesson, L., Nilsson, T. and Wikstrand, G. (2007) Probability calculus for silent elimination: A method for medium access control. Research report in mathematical statistics 3, Department of Mathematics and Mathematical Statistics, Umeå University.

[5] Flajolet, P. and Martin, G. N. (1985) Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci.* **31** 182–209.

[6] Flajolet, P. and Sedgewick, R. (1995) Mellin transforms and asymptotics: Finite differences and Rice's integrals. *Theoret. Comput. Sci.* **144** 101–124.

[7] Goh, W. and Hitczenko, P. (2007) Gaps in samples of geometric random variables. *Discrete Math.* **307** 2871–2890.

[8] Hitczenko, P. and Knopfmacher, A. (2005) Gap-free compositions and gap-free samples of geometric random variables. *Discrete Math.* **294** 225–239.

[9] Kirschenhofer, P. and Prodinger, H. (1996) The number of winners in a discrete geometrically distributed sample. *Ann. Appl. Probab.* **6** 687–694.

[10] Louchard, G. and Prodinger, H. (2008) On gaps and unoccupied urns in sequences of geometrically distributed random variables. *Discrete Math.* **308** 1538–1562.

[11] Louchard, G. and Prodinger, H. (2010) Asymptotic results for silent elimination. In *Discrete Mathematics and Theoretical Computer Science*, DMTCS Proceedings Vol. 12, pp. 185–196.

[12] Louchard, G. and Prodinger, H. (2013) The largest missing value in a composition of an integer and some Allouche–Shallit-like identities. *J. Integer Sequences* **16** #13.2.2.

[13] Prodinger, H. (1996) Combinatorics of geometrically distributed random variables: Left-to-right maxima. *Discrete Math.* **153** 253–270.

[14] Szpankowski, W. (2001) *Average Case Analysis of Algorithms on Sequences*, Wiley.

[15] Szpankowski, W. and Rego, V. (1990) Yet another application of a binomial recurrence: Order statistics. *Computing* **43** 401–410.