

Gap-free samples of geometric random variables – Extended Abstract ^{*}

Paweł Hitczenko[†]

Arnold Knopfmacher[‡]

Abstract

We study the probability that a sample of independent, identically distributed random variables with a geometric distribution is *gap-free*, that is, that the sizes of the variables in the sample form an interval. We indicate that this problem is closely related to the asymptotic probability that a random composition of an integer n is likewise gap-free.

1 Introduction.

We study samples of independent, identically distributed random variables with a geometric distribution. Specifically, let $\Gamma_1, \Gamma_2, \Gamma_3, \dots$ be independent identically distributed geometric random variables with parameter p , that is, $\mathbb{P}(\Gamma_1 = j) = pq^{j-1}, j = 1, 2, \dots$, with $p + q = 1$. We will be interested in the probability that a random sample of n such variables is *gap-free*, that is, that the sizes of the variables in the sample form an interval. In addition if the interval starts at 1, the sample is said to be *complete*.

We can restrict our attention to the probability p_n that a sample of length n is complete, since the probability that a geometric sample of length n has no ones is exponentially small.

In the case $p = 1/2$ this probability turns out to be exactly $1/2$. The case of $p \neq \frac{1}{2}$ is more interesting. In fact, in that case the sequence (p_n) does not have a limit, but exhibits small oscillations. An asymptotic expression for p_n in the case of $p \neq \frac{1}{2}$ is derived in Section 2.

Some of the previous studies relating to combinatorics of geometric random variables are as follows. In [12] the number of left-to-right maxima was investigated

in the model of *words* (strings) $a_1 \dots a_n$, where the letters $a_i \in \mathbb{N}$ are independently generated according to the geometric distribution. H.-K. Hwang and his collaborators obtained further results about this limiting behaviour in [2]. The two parameters ‘value’ and ‘position’ of the r th left-to-right maximum for geometric random variables were considered in a subsequent paper [9]. Other combinatorial questions have been considered by Prodinger in e.g. [13, 14].

The combinatorics of geometric random variables has gained importance because of their applications in computer science. We mention just two areas: **skiplists** [3, 11, 15] and **probabilistic counting** [4, 8].

The special case $p = 1/2$ of geometric random variables is closely related to *compositions* of n as shown in [5, 6]. This led us to consider the same question for compositions.

A composition of a natural number n is said to be *gap-free* if the part sizes occurring in it form an interval. In addition if the interval starts at 1, the composition is said to be *complete*.

Example. Of the 32 compositions of $n = 6$, there are 21 gap-free compositions arising from permuting the order of the parts of the partitions

$$6, 3 + 3, 3 + 2 + 1, 2 + 2 + 2, 2 + 2 + 1 + 1, \\ 2 + 1 + 1 + 1 + 1, 1 + 1 + 1 + 1 + 1$$

and 18 complete compositions arising from permuting of the order of the parts in

$$3 + 2 + 1, 2 + 2 + 1 + 1, 2 + 1 + 1 + 1 + 1, \\ 1 + 1 + 1 + 1 + 1 + 1.$$

In the full version of this paper we will show that the proportion of gap-free or of complete compositions of n is

$$1/2 + O\left(\log^{3/2} n / \sqrt{n}\right) \quad \text{as } n \rightarrow \infty,$$

by reducing the compositions problem to the special case $p = 1/2$ of geometric random variables.

2 Geometric samples.

Consider $\Gamma = (\Gamma_1, \Gamma_2, \dots, \Gamma_n)$ a sample of n i.i.d. geometric random variables with parameter p . Let

^{*}The first author is supported in part by the NSA grant MSPF-02G-043. The research was conducted during a visit to the John Knopfmacher Centre for Applicable Analysis and Number Theory at the University of Witwatersrand, Johannesburg, South Africa. He would like to thank the Centre for the invitation and for financial support. The material of the second author is based upon work supported by the National Research Foundation under grant number 2053740.

[†]Department of Mathematics, Drexel University, Philadelphia, PA 19104, U.S.A.

[‡]The John Knopfmacher Centre for Applicable Analysis and Number Theory, Department of Computational and Applied Mathematics, University of the Witwatersrand, P. O. Wits, 2050 Johannesburg, South Africa.

$p_n = \mathbb{P}(\Gamma \in \mathcal{C})$ be the probability that $(\Gamma_1, \dots, \Gamma_n)$ is complete. To obtain a recurrence relation we condition on the number of Γ_j 's that are equal to 1. Since being complete implies that there is at least one 1 among the values of Γ_i 's, by the law of total probability we find that

$$\begin{aligned} p_n &= \sum_{j=1}^n \mathbb{P} \left(\left\{ \Gamma \in \mathcal{C} \right\} \cap \left\{ \sum_{\ell=1}^n I_{\Gamma_\ell=1} = j \right\} \right) \\ &= \sum_{j=1}^n \mathbb{P} \left(\Gamma \in \mathcal{C} \mid \sum_{\ell=1}^n I_{\Gamma_\ell=1} = j \right) \times \\ &\quad \times \mathbb{P} \left(\sum_{\ell=1}^n I_{\Gamma_\ell=1} = j \right). \end{aligned}$$

We now observe that, given that j out of n Γ_k 's are one, the sample is complete if and only if the remaining $n - j$ variables take on all the values between 2 and their maximum. This is the same as to say that the geometric sample $(\Gamma_k - 1)$ of length $n - j$ is complete, given that all $n - j$ of them are at least 2. But, by the memoryless property of geometric random variables, the conditional distribution of $\Gamma_k - 1$ given that $\Gamma_k \geq 2$ is just that of Γ_k . Since those of Γ_k 's that are at least 2 remain independent, this just means that

$$\mathbb{P} \left(\Gamma \in \mathcal{C} \mid \sum_{\ell=1}^n I_{\Gamma_\ell=1} = j \right) = p_{n-j}.$$

Since

$$\mathbb{P} \left(\sum_{\ell=1}^n I_{\Gamma_\ell=1} = j \right) = \binom{n}{j} p^j q^{n-j},$$

substituting these two expressions and changing the order of summation by letting $k = n - j$, we obtain the following recurrence for p_n 's:

$$(2.1) \quad p_n = \begin{cases} 1, & \text{if } n = 0; \\ \sum_{k=0}^{n-1} p_k \binom{n}{k} q^k p^{n-k}, & \text{if } n \geq 1. \end{cases}$$

Before analysing the general case let us note that if $p = 1/2$ then the sequence $p_0 = 1$ and $p_k = 1/2$ for $k \geq 1$ is a solution. Indeed, proceeding inductively we get

$$\begin{aligned} p_n &= \sum_{k=0}^{n-1} p_k \binom{n}{k} q^k p^{n-k} \\ &= \frac{1}{2^n} + \sum_{k=1}^{n-1} \frac{1}{2} \binom{n}{k} \frac{1}{2^n} \\ &= \frac{1}{2} \sum_{k=0}^n \binom{n}{k} \frac{1}{2^n} = \frac{1}{2}. \end{aligned}$$

The case of $p \neq 1/2$ is more interesting. In fact, in that case the sequence (p_n) does not have a limit, but exhibits small oscillations. As illustrated in Figures 1,2,3 below, both the period and amplitude of the oscillations depend on the size of p .

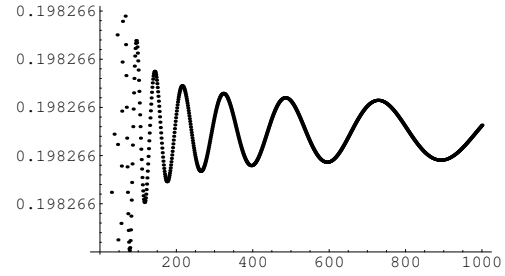


Figure 1: Plot of p_n for $p = 1/3$ and $1 \leq n \leq 1000$.

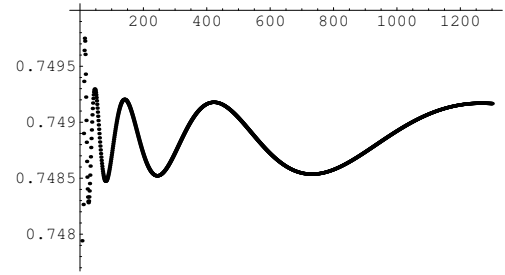


Figure 2: Plot of p_n for $p = 2/3$ and $1 \leq n \leq 1300$.

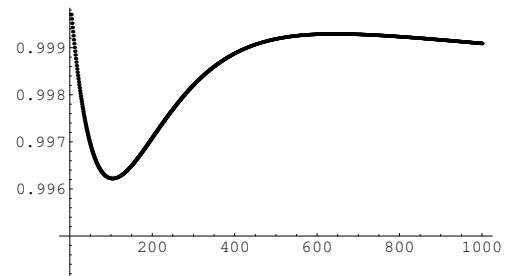


Figure 3: Plot of p_n for $p = 0.99$ and $1 \leq n \leq 1000$.

To treat this case we will follow an approach that became quite common in the analysis of certain algorithms (see e.g. numerous examples in [7, 16]). We Poissonize the problem by considering the Poisson transform of the sequence (p_n) , analyse its asymptotics, and then we de-Poissonize to recover the asymptotics of (p_n) . To carry out this program let $P(z)$, for z complex, be the

Poisson transform of (p_n) . That is

$$\begin{aligned}
P(z) &= \sum_{n=0}^{\infty} p_n \frac{z^n}{n!} e^{-z} \\
&= p_0 e^{-z} + \sum_{n=1}^{\infty} \frac{z^n e^{-z}}{n!} \left\{ \sum_{\ell=0}^{n-1} p_\ell \binom{n}{\ell} q^\ell p^{n-\ell} \right\} \\
&= e^{-z} + e^{-z} \sum_{n=1}^{\infty} \sum_{\ell=0}^{n-1} p_\ell \frac{q^\ell z^\ell p^{n-\ell} z^{n-\ell}}{\ell!(n-\ell)!} \\
&= e^{-z} + e^{-z} \sum_{\ell=0}^{\infty} p_\ell \frac{q^\ell z^\ell}{\ell!} \sum_{n=\ell+1}^{\infty} \frac{p^{n-\ell} z^{n-\ell}}{(n-\ell)!} \\
&= e^{-z} + e^{-z} \sum_{\ell=0}^{\infty} p_\ell \frac{q^\ell z^\ell}{\ell!} \{e^{pz} - 1\} \\
&= e^{-z} + e^{-qz} \sum_{\ell=0}^{\infty} p_\ell \frac{q^\ell z^\ell}{\ell!} \{1 - e^{-pz}\} \\
&= e^{-z} + (1 - e^{-pz})P(qz).
\end{aligned}$$

Hence $P(z)$ satisfies the following functional equation

$$\begin{aligned}
(2.2) \quad P(z) &= P(qz) + e^{-z} - e^{-pz}P(qz) \\
&= P(qz) + e^{-z} - e^{-z} \sum_{n=0}^{\infty} p_n \frac{q^n z^n}{n!} \\
&= P(qz) - T(z),
\end{aligned}$$

where $T(z) = e^{-z} \sum_{n=1}^{\infty} p_n \frac{q^n z^n}{n!}$, and since it is clear from (2.1) and the binomial formula that $0 \leq p_n \leq 1$, the series converges absolutely for every z . Moreover, for $x \in \mathbf{R}$, $T(x) = O(x)$ as $x \rightarrow 0_+$ and $T(x)$ has exponential decay as $x \rightarrow \infty$. Thus the Mellin transform of $T(x)$ exists in the strip $\langle -1, \infty \rangle := \{s \in \mathbf{C} : -1 < \Re(s) < \infty\}$. By direct iteration we obtain for every $m \geq 0$

$$P(z) = P(qz) - T(z) = P(q^{m+1}z) - \sum_{j=0}^m T(q^j z),$$

and by passing to a limit with m ,

$$P(z) = 1 - \sum_{j=0}^{\infty} T(q^j z).$$

Letting $Q(z) = P(z) - 1 = -\sum_{j=0}^{\infty} T(q^j z)$ and taking the Mellin transform we obtain

$$Q^*(s) = -\sum_{j=0}^{\infty} q^{-js} T^*(s) = -\frac{T^*(s)}{1 - q^{-s}} = \frac{T^*(s)}{q^{-s} - 1},$$

provided that series converges. Since this happens for $\Re(s) < 0$, $Q^*(s)$ will exist in a strip $\langle -1, 0 \rangle$. Inverting

the Mellin transform yields

$$Q(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} Q^*(s) x^{-s} ds,$$

for any $-1 < c < 0$. This integral can be evaluated with the aid of the residue theorem. Now, $x^{-s} Q^*(s) = x^{-s} T^*(s)/(q^{-s} - 1)$ has simple poles whenever $q^{-s} = 1$, i.e. at $\chi_k = 2k\pi i / \ln(1/q)$, $k = 0, \pm 1, \pm 2, \dots$, with corresponding residues

$$\left(\frac{q}{x}\right)^{\chi_k} \frac{T^*(\chi_k)}{\ln(1/q)}.$$

The main term comes from $k = 0$ and the remaining residues will contribute oscillatory terms of relatively small amplitude. In order to complete the proof we will need to de-Poissonize this result. Once this is done we will conclude that

$$\begin{aligned}
p_n &\sim P(n) = Q(n) + 1 \\
&\sim 1 - \frac{T^*(0)}{\ln(1/q)} \\
&\quad - \frac{2}{\ln(1/q)} \Re \left(\sum_{k=1}^{\infty} \exp\{\chi_k \ln(q/n)\} T^*(\chi_k) \right).
\end{aligned}$$

The values $T^*(\chi_k)$ are given by

$$\begin{aligned}
T^*(\chi_k) &= \mathcal{M} \left(e^{-z} \sum_{j=1}^{\infty} p_j \frac{q^j z^j}{j!}; \chi_k \right) \\
&= \sum_{j=1}^{\infty} p_j \frac{q^j}{j!} \mathcal{M}(z^j e^{-z}; \chi_k) \\
&= \sum_{j=1}^{\infty} p_j \frac{q^j}{j!} \Gamma(\chi_k + j) \\
&= \sum_{j=1}^{\infty} p_j q^j \frac{\Gamma(\chi_k + j)}{\Gamma(j+1)}.
\end{aligned}$$

Since the series converge geometrically fast, they can be evaluated numerically with the aid of (2.1). For example, setting $k = 0$ gives the main term

$$1 - \frac{T^*(0)}{\ln(1/q)} = 1 - \frac{1}{\ln(1/q)} \sum_{j=1}^{\infty} p_j \frac{q^j}{j}.$$

Its plot as a function of p is given in Figure 4.

To de-Poissonize we use the fact that $P(z)$ satisfies (2.3) which is a special case of [16, Theorem 10.5] (see also [7]) with $\gamma_1(z) \equiv 0$, $\gamma_2(z) \equiv 1$, and $t(z) = -T(z)$. Thus we need to verify conditions (10.28) – (10.32) of

[16]. But this is straightforward: (10.28) holds for any $\beta > 0$, (10.29) holds as well, since

$$\begin{aligned} |t(z)| &= |T(z)| \\ &\leq |e^{-z}| \sum_{n=1}^{\infty} p_n \frac{q^n |z|^n}{n!} \\ &\leq e^{-\Re(z)} \sum_{n=1}^{\infty} \frac{q^n |z|^n}{n!} \\ &\leq e^{-\Re(z)} e^{q|z|}, \end{aligned}$$

which is bounded by 1 provided $\Re(z)/|z| > q$, which holds in a cone

$$\mathcal{S}_\theta := \{z = re^{i\theta} : \cos \theta > q\}.$$

(10.30) is trivial and (10.31) holds outside the cone since for $z \notin \mathcal{S}_\theta$, $\Re(z) < \alpha|z|$ for some $\alpha < 1$. Finally, (10.32) is true since

$$|T(z)|e^{\Re(z)} \leq e^{\Re(z)} |e^{-z}| \sum_{n=1}^{\infty} p_n \frac{q^n |z|^n}{n!} \leq e^{q|z|} \leq \frac{1}{3} e^{\alpha|z|},$$

as long as $q < \alpha < 1$ and $|z|$ is large enough.

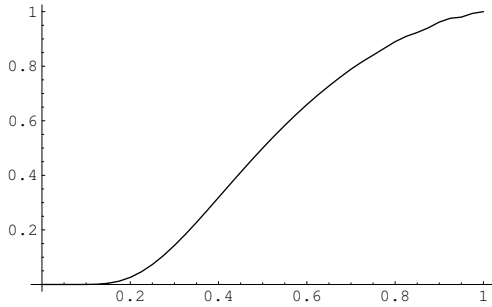


Figure 4: Plot of the non-oscillating limit term for p_n for $0 \leq p \leq 1$.

3 Further Remarks.

As shown above the special case $p = 1/2$ does not have oscillations. This also follows from the asymptotic formula for p_n in view of the fact that for $p = 1/2$ and $\chi_k = 2k\pi i / \ln 2$,

$$\begin{aligned} T^*(\chi_k) &= \sum_{j=1}^{\infty} \frac{1}{2^{j+1}} \frac{\Gamma(\chi_k + j)}{\Gamma(j+1)} \\ &= \frac{1}{2} \left(-1 + 2^{\frac{2ik\pi}{\ln 2}} \right) \Gamma\left(\frac{2ik\pi}{\ln 2}\right) \\ &= 0, \end{aligned}$$

for $k = \pm 1, \pm 2, \pm 3, \dots$

It is interesting to investigate the amplitude of the oscillations on either side of the critical value $p = 1/2$. As shown in Figure 5 these become very small near to the critical value.

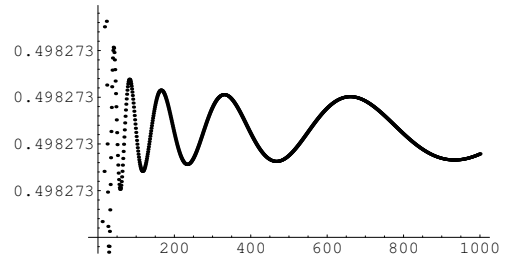


Figure 5: Plot of p_n for $p = 0.499$ and $1 \leq n \leq 1000$.

For $0 \leq p \leq 1/2$ the amplitude of the fluctuations increase steadily up until around $p = 0.48$ and then decrease rapidly to zero.

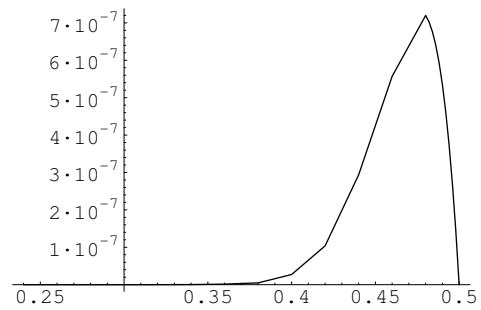


Figure 6: Plot of the amplitude of the fluctuations for $p \leq 0.5$.

For $p \geq 1/2$ the amplitude of the fluctuations increase steadily and in general are orders of magnitude larger than for $p < 1/2$.

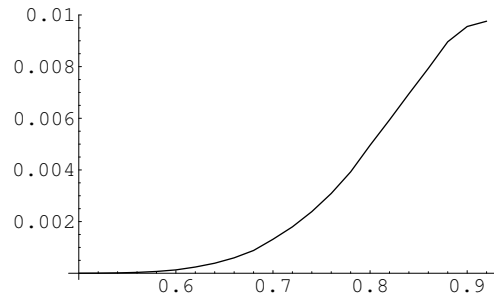


Figure 7: Plot of the amplitude of the fluctuations for $p \geq 0.5$.

References

- [1] G. E. Andrews., *The Theory of Partitions*, Addison – Wesley, Reading, MA, 1976.
- [2] Z.-D. Bai, H.-K. Hwang, and W.-Q. Liang, *Normal approximations of the number of records in geometrically distributed random variables*, *Random Struct. Alg.*, 13 (1998), pp. 319–334.
- [3] L. Devroye, *A limit theory for random skip lists*, *Ann. Appl. Probab.*, 2 (1992), pp. 597–609.
- [4] P. Flajolet and G. N. Martin, *Probabilistic Counting Algorithms for Data Base Applications*, *J. Comp. Syst. Sci.*, 31 (1985), pp. 182–209.
- [5] P. Hitczenko, C. D. Savage, *On the multiplicity of parts in a random composition of a large integer*, *SIAM J. Discrete Math.*, to appear, 2004.
- [6] P. Hitczenko, G. Louchard, *Distinctness of compositions of an integer: a probabilistic analysis*, *Random Struct. Alg.*, 19 (2001), pp. 407–437.
- [7] P. Jacquet, W. Szpankowski, *Analytical de-Poissonization and its applications*, *Theoret. Comput. Sci.*, 201 (1998), pp. 1–62.
- [8] P. Kirschenhofer and H. Prodinger, *On the Analysis of Probabilistic Counting*, in *Lecture Notes in Mathematics*, (E.Hlawka, R.F.Tichy eds.) 1452, pp. 117–120, 1990.
- [9] A. Knopfmacher and H. Prodinger. *Combinatorics of geometrically distributed random variables: Value and position of the r th left-to-right maximum*, *Discrete Math.*, 226 (2001), pp. 255–267.
- [10] D. Knuth, *The Art of Computer Programming, vol. 3: Sorting and Searching*, Addison-Wesley, Reading, MA, 1973.
- [11] T. Papadakis, I. Munro and P. Poblete, *Average search and update costs in skip lists*, *BIT*, 32 (1992), pp. 316–332.
- [12] H. Prodinger, *Combinatorics of geometrically distributed random variables: Left-to-right maxima*, *Discrete Math.*, 153 (1996), pp. 253–270.
- [13] H. Prodinger, *Combinatorics of geometrically distributed random variables: Inversions and a parameter of Knuth*, *Annals of Combinatorics*, 5 (2001), 241–150.
- [14] H. Prodinger, *Combinatorics of geometrically distributed random variables: Lengths of ascending runs*, *LATIN2000*, *Lecture Notes in Computer Science* 1776, pp. 473–482, 2000.
- [15] W. Pugh, *Skip lists: a probabilistic alternative to balanced trees*, *Comm. ACM*, 33 (1990), pp. 668–676.
- [16] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley, 2001.