Available online at www.sciencedirect.com
ScienceDirect
Discrete Mathematics 308 (2008) 1538 – 1562
DISCRETE MATHEMATICS
www.elsevier.com/locate/disc

# On gaps and unoccupied urns in sequences of geometrically distributed random variables

Guy Louchard[a], Helmut Prodinger[b]

[a] Université Libre de Bruxelles, Département d'Informatique, CP 212, Boulevard du Triomphe, B-1050 Bruxelles, Belgium
[b] University of Stellenbosch, Mathematics Department, 7602 Stellenbosch, South Africa

Received 27 April 2005; received in revised form 2 April 2007; accepted 3 April 2007
Available online 11 April 2007

**Abstract**

This paper continues the study of gaps in sequences of $n$ geometrically distributed random variables, as started by Hitczenko and Knopfmacher [Gap-free samples of geometric random variables, Discrete Math. 294 (2005) 225–239], who concentrated on sequences which were gap-free. Now we allow gaps, and count some related parameters.

Our terminology of gaps just means empty "urns" (within the range of occupied urns), if we think about an urn model. This might be called *weak* gaps, as opposed to *maximal* gaps, as in Hitczenko and Knopfmacher [Gap-free samples of geometric random variables, Discrete Math. 294 (2005) 225–239]. If one considers only "gap-free" sequences, both notions coincide asymptotically, as $n \to \infty$.

First, the probability $p_n(r)$ that a sequence of length $n$ has a fixed number $r$ of empty urns is studied; this probability is asymptotically given by a constant $p^*(r)$ (depending on $r$) plus some small oscillations. When $p = q = \frac{1}{2}$, everything simplifies drastically; there are no oscillations.

Then, the random variable 'number of empty urns' is studied; all moments are evaluated asymptotically. Furthermore, samples that have $r$ empty urns, in particular the random variable 'largest non-empty urn' are studied. All moments of this distribution are evaluated asymptotically.

The behavior of the quantities obtained in our asymptotic formulæ is also studied for $p \to 0$ resp. $p \to 1$, through a variety of analytic techniques.

The last section discusses the concept called 'super-gap-free.' A sample is super-gap-free, if $r = 0$ and each non-empty urn contains at least 2 items (and $d$-super-gap-free, if they contain $\geqslant d$ items). For the instance $p = q = \frac{1}{2}$, we sketch how the asymptotic probability (apart from small oscillations) that a sample is $d$-super-gap-free can be computed.


*Keywords:* Gaps; Geometric random variables; Urns; Moments; Mellin transform; Gumbel distribution

## 1. Introduction

Let us consider a sequence of $n$ random variables (RV), $Y_1, \ldots, Y_n$, distributed (independently) according to the geometric distribution $\mathrm{Geom}(p)$. Set $q := 1 - p$, then $\mathbb{P}(Y = j) = pq^{j-1}$. If we neglect the order in which the $n$ items arrive, we can think about an urn model, with urns labelled $1, 2, \ldots$, the probability of each ball falling into urn $j$ being given by $pq^{j-1}$.

*E-mail addresses:* louchard@ulb.ac.be (G. Louchard), hproding@sun.ac.za (H. Prodinger).

0012-365X/$ - see front matter 
doi:10.1016/j.disc.2007.04.012

Set the indicator RV (in the sequel we drop the $n$-specification to simplify the notations):[1]

$$X_i := [\![\text{value } i \text{ appears among the } n \text{ RVs}]\!],$$

i.e. urn $i$ is not empty.

A comment on terminology: In [11], Hitczenko and Knopfmacher consider gap-free distributions, i.e., the indices $a, a+1, \ldots, b$ of the non-empty urns form an interval. Without loss of generality one may assume that $a=1$, since there is only an exponentially small probability for the first urn to be empty. So, the extra notation ("complete") for this instance can be ignored.

Gaps themselves are not explicitly mentioned, but it is understood that a gap is a (maximal) sequence of empty urns between non-empty ones.

Our point of view is different here: We say that an urn $b$ is a gap, if it is empty and comes before the last non-empty urn. To distinguish clearly from the maximal gaps mentioned before, we could call this "gaps in the weak sense"; of course, for the terminology "gap-free" both versions amount asymptotically to the same.

In this paper, we study these *gaps in the weak sense.* However, since the name "gaps" should be reserved for the version studied by Hitczenko and Knopfmacher [11], we will use terminology such as "the number of unoccupied urns," or similar ones.

Hitczenko and Knopfmacher analyze the quantity

$$p_n(0) := \mathbb{P}[\text{All urns are occupied up to the maximal non-empty urn}].$$

Recently, Goh and Hitczenko [9] have continued the study of gaps in the "maximal" sense, as described before.

In our paper, we analyze the probability $p_n(r)$ of having $r$ empty urns, the moments of the total number of empty urns and some other parameters.

As a link to more practically oriented research, we mention *probabilistic counting* [7], which can be seen in the context of our empty urn discussion.

The case $p = \frac{1}{2}$ has a particular interest: it is related to the compositions of integers, see [12].

It is intuitively not at all clear, but nevertheless true, that the quantities that we analyze for general $p$, simplify for the special choice of $p = \frac{1}{2}$. This produces identities, since a complicated expression simplifies for a special choice of the parameter. Now, one gets these identities "for free," since two different approaches must eventually lead to the same result. Nevertheless, we believe that there is a genuine interest in producing independent (analytic) proofs for these simplifications. To give a flavor of such identities,

$$\int_0^\infty \mathrm{e}^{-y} \left[ \sum_{j=0}^{\infty} (-1)^{v(j)} \mathrm{e}^{-jy} \right] \frac{\mathrm{d}y}{y} = \frac{\ln 2}{2},$$

where $v(j)$ is the number of ones in the binary representation of $j$.

For those readers who know the literature on analysis of algorithms well, we provide the following *informal* comment that might be helpful and might be compared with the one described in [20,21]: since a variance cannot be negative and the main term fluctuates around zero, the fluctuation must be identical to zero and the Fourier coefficients must be equal to zero. Now, such a combinatorial argument is nice and sweet when it applies! But there are situations as well, when one has to *compute* the Fourier coefficients, and that is at the same level of complexity as to prove in the other instances that they are zero.

The probabilities and moments we compute will have an asymptotic expansion of the form

$$D \sim D^* + D_P(\log n),$$

where the dominant term $D^*$ is independent of $n$ and $D_P$ is a periodic function, with 0 mean, of period 1 and usually of small amplitude, of order $10^{-5}$.

We also provide some asymptotics for $p$ going to 0 and 1.

[1] Here we use the indicator function ('Iverson's notation') proposed by Graham et al. [10].