J. Appl. Prob. **46,** 1038–1051 (2009)
Printed in England
© *Applied Probability Trust* 2009

GAPS IN DISCRETE RANDOM SAMPLES

RUDOLF GRÜBEL,* *Leibniz Universität Hannover*
PAWEŁ HITCZENKO,** *Drexel University*

Abstract

Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with values in the set $\mathbb{N}_0$ of nonnegative integers. Motivated by applications in enumerative combinatorics and analysis of algorithms we investigate the number of gaps and the length of the longest gap in the set $\{X_1, \dots, X_n\}$ of the first n values. We obtain necessary and sufficient conditions in terms of the tail sequence $(q_k)_{k\in\mathbb{N}_0}$, $q_k = \mathrm{P}(X_1 \geq k)$, for the gaps to vanish asymptotically as $n \to \infty$: these are $\sum_{k=0}^{\infty} q_{k+1}/q_k < \infty$ and $\lim_{k\to\infty} q_{k+1}/q_k = 0$ for convergence almost surely and convergence in probability, respectively. We further show that the length of the longest gap tends to ∞ in probability if $q_{k+1}/q_k \to 1$. For the family of geometric distributions, which can be regarded as the borderline case between the light-tailed and the heavy-tailed situations and which is also of particular interest in applications, we study the distribution of the length of the longest gap, using a construction based on the Sukhatme–Rényi representation of exponential order statistics to resolve the asymptotic distributional periodicities.

Keywords: Geometric distribution; heavy tail; light tail; periodicities; Sukhatme–Rényi representation

2000 Mathematics Subject Classification: Primary 60C05
Secondary 60F99

1. Introduction

Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with values in $\mathbb{N}_0$, the set of nonnegative integers. Standard examples are the geometric and the Poisson distributions. We consider the first n values of the sequence as a random set,

$$A_n := \{X_1, \dots, X_n\}.$$

Obviously,

$$A_n \subset \{m_n, \dots, M_n\} \quad \text{with} \quad M_n := \max_{1\leq i\leq n} X_i \quad \text{and} \quad m_n := \min_{1\leq i\leq n} X_i.$$

By a *gap* we mean a contiguous and nonempty subset $\{j, \dots, j+l-1\}$ of the complement $\{m_n, \dots, M_n\} \setminus A_n$ of A_n in the sample range that is maximal in the sense that both $j-1$, $j+l \in A_n$. We then call l the *length* of the gap. We are interested in the total number, Y_n, of gaps and the length, L_n, of the longest gap among the first n sample values.

Such quantities are of interest in enumerative combinatorics, in particular in connection with compositions of integers, and analysis of algorithms, in particular approximate counting, and

* Postal address: Institut für Mathematische Stochastik, Leibniz Universität Hannover, Postfach 6009, D-30060 Hannover, Germany. Email address: rgrubel@stochastik.uni-hannover.de
** Postal address: Departments of Mathematics and Computer Science, Drexel University, 3141 Chestnut Str., Philadelphia, PA 19104, USA. Email address: phitczenko@math.drexel.edu
Supported in part by the NSA grant #H98230-09-1-0062

elsewhere; see [6], [4], and the references therein. A related concept, weak gaps, essentially the size of $\{0, \ldots, M_n\} \setminus A_n$, has been investigated in [9]. In all three references the geometric distribution plays a central role, as does the approach to such problems by methods from complex analysis, using, for example, Mellin transforms, analytic de-Poissonization, and singularity analysis.

In the present paper we investigate the gaps for general discrete distributions, with the aim of classifying these distributions with respect to the asymptotic behavior of the number of gaps or the length of the longest gap as the sample size increases to infinity. The results show that the geometric case can be seen as a 'borderline' between $L_n \to 0$ and $L_n \to \infty$. A second aim of the present paper is the study of the distributional asymptotics of L_n as $n \to \infty$ for geometric random samples. We show that the asymptotic distributional periodicities can be resolved in terms of a suitable background construction.

Our methods are probabilistic. For example, in connection with almost-sure convergence for distributions with light tails we regard the sequence of sample maxima as a Markov chain; we use the Sukhatme–Rényi representation in connection with the geometric case, and we use this representation together with the quantile transformation in the heavy-tailed case.

Apart from being connected to combinatorics and theoretical computer science, the setup studied in this paper is also related to infinite urn models, where urns are numbered $0, 1, 2 \ldots$ and balls are independently put into urn k with probability p_k. The classical models have a finite number of urns and have been extensively studied, but the infinite case has already been considered in [8]. These models have received some attention recently; see [1] and [7] for example. The latter gives a local limit theorem for the number of occupied urns, which is the cardinality of A_n in our notation. The survey [3] also points to other applications of infinite urn models. Still, by far, the most heavily studied model concerns the geometric probabilities p_k, and aside from the papers treating this case, we are not aware of any results on the structure of gaps in a general setting.

The results are given in the next section, together with some related remarks and examples. Proofs are collected in Section 3.

2. Results

2.1. Light and heavy tails

Our first two results deal with the extreme case that the gaps will eventually vanish. Let $(\Omega, \mathcal{A}, \mathrm{P})$ be the basic probability space on which the variables $(X_i)_{i\in\mathbb{N}}$ are defined. Set $p_k = \mathrm{P}(X_1 = k)$ for $k \in \mathbb{N}_0$. To avoid the trivial situation that gaps may be caused by some of the p_ks being 0, we assume throughout that

$$p_k > 0 \quad \text{for all } k \in \mathbb{N}_0. \tag{1}$$

In view of $L_n \in \mathbb{N}_0$ the almost-sure convergence of L_n to 0 as $n \to \infty$ is equivalent to the property that $L_n(\omega) = 0$ from some $n = n(\omega)$ onwards for P-almost all ω. Of course, at this end of the spectrum the number of gaps and the length of the longest gap become asymptotically indistinguishable in view of $\{Y_n = 0\} = \{L_n = 0\}$, so that it is enough to consider one of these variables. Let $(q_k)_{k\in\mathbb{N}_0}$,

$$q_k := \sum_{j=k}^{\infty} p_j = \mathrm{P}(X_i \geq k) \quad \text{for all } k \in \mathbb{N}_0,$$

be the tail sequence associated with $(p_k)_{k\in\mathbb{N}_0}$.

Theorem 1. *The sequence $(L_n)_{n\in\mathbb{N}}$ converges to 0 with probability 1 as $n \to \infty$ if and only if*

$$\sum_{k=0}^{\infty} \frac{q_{k+1}}{q_k} < \infty. \tag{2}$$

For the weaker convergence in probability, we again adapt the convergence to the fact that L_n and Y_n are nonnegative and integer valued. Convergence of L_n to 0 in probability is equivalent to $\lim_{n\to\infty} \mathrm{P}(L_n = 0) = 1$, and similarly for Y_n. In the proof we will see that if L_n does not converge to 0 in probability then we even have $\mathrm{P}(\limsup_{n\to\infty} L_n \geq 1) = 1$, which of course is not surprising in view of Kolmogorov's 0–1 law for terminal events.

Theorem 2. *Let Z_n be L_n or Y_n. Either of the conditions (3) or (4) below is necessary and sufficient for the convergence in probability of Z_n to 0 as $n \to \infty$:*

$$\lim_{k\to\infty} \frac{q_{k+1}}{q_k} = 0, \tag{3}$$

$$\lim_{n\to\infty} \mathrm{E}\, Z_n = 0. \tag{4}$$

Remark 1. Conditions (2) and (3) can be rewritten in terms of the individual probabilities p_k in (1) as $\sum_{k=0}^{\infty} p_{k+1}/p_k < \infty$ and $\lim_{k\to\infty} p_{k+1}/p_k = 0$, respectively; see Lemma 1, below.

Example 1. The Poisson distribution with mean λ is an example that satisfies (3), but not (2) as

$$\frac{p_{k+1}}{p_k} = \frac{\lambda}{k+1}.$$

More broadly, suppose that $p_k \propto (c/k^\alpha)^k$ for some constants $c > 0$ and $\alpha > 0$. We then have $p_{k+1}/p_k \propto k^{-\alpha}$. Hence, convergence in probability to 0 of the longest gap (or the number of gaps) holds for the full family, but almost-sure convergence requires that $\alpha > 1$.

At the other end of the spectrum of tail behavior we obtain a sufficient condition for the longest gap to converge to ∞ in probability.

Theorem 3. *If*

$$\lim_{k\to\infty} \frac{q_{k+1}}{q_k} = 1 \tag{5}$$

then, for all $l \in \mathbb{N}$,

$$\lim_{n\to\infty} \mathrm{P}(L_n \geq l) = 1.$$

The methods that we will use in the proofs can also be used to obtain more specific results on the asymptotic behavior of L_n or Y_n as $n \to \infty$ under specific assumptions on the asymptotics of the individual probabilities p_k as $k \to \infty$.

Theorem 4. *Suppose that $p_k \propto 1/k^\alpha$ for some constant $\alpha > 1$. Then*

$$\mathrm{E}\, Y_n \propto n^{1/\alpha}. \tag{6}$$

2.2. Geometric case

We now consider the special case that the X-sequence is from a geometric distribution: for some p, $0 < p < 1$, and all $i \in \mathbb{N}$,

$$\mathrm{P}(X_i = k) = (1-p)^k p \quad \text{for all } k \in \mathbb{N}_0. \tag{7}$$

This case plays a central role in the application in enumerative combinatorics and analysis of algorithms that we mentioned in the introduction.

We write $\mathcal{L}(Y)$ for the distribution of a random quantity Y. Our main result below implies that the family $\{\mathcal{L}(L_n) : n \in \mathbb{N}\}$ is tight and that L_{n_m} converges in distribution along subsequences $(n_m)_{m\in\mathbb{N}}$ of a specific type determined by p. This is a familiar phenomenon in the analysis of random discrete structures and often appears in connection with problems in enumerative combinatorics or analysis of algorithms. In the present context it has already been noted in [4], [6], and [9].

Remark 2. (a) We mention in passing that (7) is the 'number of failures' version of the geometric distribution. With this version we have support $\mathbb{N}_0$ as required in (1). Trivial modifications lead to a variant for the geometric distribution that arises as the time of the first success, and indeed, a similar comment applies to our results in connection with more general integer shifts of arbitrary discrete distributions.

(b) We expect that the results in this section can be extended from the geometric case to a more general class of distributions with tail ratios converging to a limit, i.e. with

$$\lim_{k\to\infty} \frac{q_{k+1}}{q_k} = \eta \in (0, 1), \tag{8}$$

possibly under additional conditions on the rate of convergence in (8).

Our aim now is to give a probabilistic construction that leads to a representation of the whole family of potential limit distributions along subsequences as deterministic transformations of one single distribution; see [5] for more on this approach and some related examples. A similar construction has also been used in [2] in connection with the analysis of an election algorithm. Such a construction can be used to handle simultaneously a variety of random variables related to gaps. Below we only deal with L_n, but the method can also be used for Y_n. Indeed, the geometric case can be seen as a borderline between the distributions that have an asymptotically contiguous sample range and those where the gaps (number, maximal length) grow beyond all bounds. For example, large geometric samples will have one long contiguous part starting at 0, and our method can be used to analyze the distributional asymptotics of the size

$$S_n := \max\{k \in \mathbb{N}_0 : \{0, 1, \ldots, k\} \subset \{X_1, \ldots, X_n\}\}$$

of this block as $n \to \infty$, or of the difference $M_n - S_n$.

The starting point for the construction is a sequence $(V_i)_{i\in\mathbb{N}}$ of independent random variables where, for each $i \in \mathbb{N}$, V_i has an exponential distribution with mean $1/i$. Then

$$M'_n := \max\{V_i : i = 1, \ldots, n-1\} \uparrow M_\infty := \sup\{V_i : i \in \mathbb{N}\}$$

with

$$\mathrm{P}(M_\infty \le x) = \prod_{k=1}^{\infty} (1 - \mathrm{e}^{-kx}) \quad \text{for all } x \ge 0. \tag{9}$$

In particular, the maximum of the V-variables is finite with probability 1. Let

$$W_{l,n} := \sum_{i=l}^{n} V_i \quad \text{for } l \le n.$$

It is easy to check that, for all $l \in \mathbb{N}$,

$$Z_{l,n} := W_{l,n} - \log n \;\to\; Z_{l,\infty} \quad \text{as } n \to \infty$$

almost surely and in quadratic mean for some finite random variable $Z_{l,\infty}$; see, e.g. the martingale argument given in [5]. Finally, we define the functions $\phi_p\colon [0,\infty) \to \mathbb{N}$ and $\psi_p\colon [0,\infty) \to [0,1)$ by

$$\phi_p(x) := \lfloor c(p)^{-1}x \rfloor \quad \text{and} \quad \psi_p(x) := \{c(p)^{-1}x\}, \tag{10}$$

with

$$c(p) := -\log(1-p).$$

Here $\{x\}$ denotes the fractional part of x; it should be clear from the context whether the curly brackets refer to this function or whether they are used to denote a set.

We can now state our next result. Note that the lower bound in (11), below, does not depend on n, which implies that $\{\mathcal{L}(L_n)\colon n \in \mathbb{N}\}$ is tight.

Theorem 5. *With the notation introduced above,*

$$\mathrm{P}(M_\infty \le c(p)(l-1)) \le \mathrm{P}(L_n \le l) \le \mathrm{P}(M_n' \le c(p)(l+1)) \tag{11}$$

for all $n, l \in \mathbb{N}$. Furthermore, if $(n_m)_{m\in\mathbb{N}}$ is such that $n_m \to \infty$ and $\psi_p(\log n_m) \to \eta$ for some $\eta \in [0,1]$ as $m \to \infty$, then L_{n_m} converges in distribution to $L_\infty(\eta)$ as $m \to \infty$, with

$$L_\infty(\eta) := \max_{l\in\mathbb{N}} (c(p)^{-1}V_l + \psi_p(Z_{l+1,\infty} + c(p)\eta) - \psi_p(Z_{l,\infty} + c(p)\eta)). \tag{12}$$

Finally, for all $\eta \in [0,1]$ and $l \in \mathbb{N}$,

$$\mathrm{P}(M_\infty \le c(p)(l-1)) \le \mathrm{P}(L_\infty(\eta) \le l) \le \mathrm{P}(M_\infty \le c(p)(l+1)). \tag{13}$$

It may not be apparent that the maximum in (12) is taken over a set of integer values, but we will see in the proof that

$$c(p)^{-1}V_l + \psi_p(Z_{l+1,\infty} + c(p)\eta) - \psi_p(Z_{l,\infty} + c(p)\eta) \in \{\lfloor c(p)^{-1}V_l \rfloor, \lceil c(p)^{-1}V_l \rceil\}.$$

Theorem 5 can be used to obtain information about the family of limit distributions. For example, it follows from (12) that, for all $\eta \in [0,1]$,

$$|L_\infty(\eta) - c(p)^{-1}M_\infty| \le 1.$$

Furthermore (note that we have suppressed the dependence on p in (13)), we see that $pL_\infty(\eta)$ converges in distribution to M_∞ as $p \to 0$, whatever η, which means that, for small success probabilities, the periodicity will become negligible, and which also gives the order of growth of the longest gap. The last statement of Theorem 5 (see (13)) provides upper and lower bounds for the distribution function of $L_\infty(\eta)$ that arise from shifting the continuous distribution function

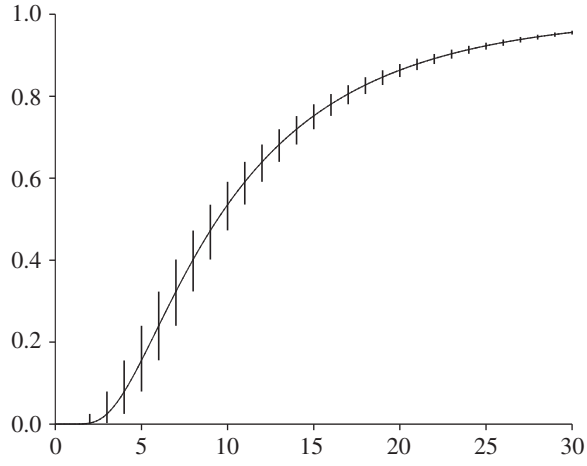

FIGURE 1: Continuous limit and discrete bounds (see text).

of M_∞, which is given explicitly (and in a numerically accessible form) in (9). Figure 1 shows the distribution function of $c(p)^{-1}M_\infty$ and bounds for $\mathrm{P}(L_\infty(\eta) \le l)$, $0 \le \eta \le 1$, for $p = \frac{1}{10}$. As $L_\infty(\eta)$ is an integer-valued random variable, bounds for its distribution function can be specified by intervals for the values in $l \in \mathbb{N}_0$; in the figure, these intervals are visualized by vertical bars.

Constructions of the above type can be used to obtain an intuitive understanding of the structure of gaps (in the geometric case, but also more generally). Clearly, as n increases, either a new gap may appear at the right end of the sample due to a jump in the largest value, or nothing may happen at all if the next sample value is already occupied, or an existing gap may shrink or be divided into two smaller gaps. For small p, the limit model may serve as an approximation if, for example, interest is in the probability that the largest gap is the one at the right end. The next result shows that this happens with probability slightly bigger than $\frac{1}{2}$.

Theorem 6. *Let V_i, $i \in \mathbb{N}$, be as above, and let $M_{l,\infty} := \max_{i \ge l} V_i$. Then, as $p \to 0$, the limiting probability that the longest gap occurs as the difference between the two largest sample values converges to*

$$\mathrm{P}(V_1 > M_{2,\infty}) = \int_0^\infty \mathrm{e}^{-x} \prod_{k=2}^{\infty} (1 - \mathrm{e}^{-kx})\,\mathrm{d}x = \mathrm{E}\,M_{1,\infty} - \mathrm{E}\,M_{2,\infty} \approx 0.516.$$

3. Proofs

We will prove the results for the light-tailed case first, then deal with the geometric case, and finally give the proofs for heavy-tailed distributions as these use the constructions introduced for the geometric case. Let

$$h_k := \mathrm{P}(X_1 \ge k+1 \mid X_1 \ge k) = \frac{q_{k+1}}{q_k}, \qquad k \in \mathbb{N}_0,$$

be the tail ratios. We first substantiate the remark following Theorem 2.

Lemma 1. *We have*

$$\lim_{k\to\infty} h_k = 0 \quad\Longleftrightarrow\quad \lim_{k\to\infty} \frac{p_{k+1}}{p_k} = 0, \tag{14}$$

$$\sum_{k=0}^{\infty} h_k < \infty \quad\Longleftrightarrow\quad \sum_{k=0}^{\infty} \frac{p_{k+1}}{p_k} < \infty. \tag{15}$$

Remark 3. For distributions with heavy tails, (5) is implied by $\lim_{k\to\infty} p_{k+1}/p_k = 1$, but the converse is not true.

Proof of Lemma 1. We first show that both parts of (14) are equivalent to

$$\lim_{k\to\infty} r_k = 0 \quad \text{with} \quad r_k := \frac{q_{k+1}}{p_k}. \tag{16}$$

The sufficiency of the latter for the right-hand side of (14) is clear. The necessity follows from the observation that if $p_{k+1}/p_k \to 0$ as $k \to \infty$ then, for a fixed $\delta \in (0, 1)$, there exists a k_0 such that, for all $k \geq k_0$,

$$\frac{p_{k+1}}{p_k} \leq \delta.$$

Therefore, for such ks and $m \geq 1$, we have

$$p_{k+m} = \frac{p_{k+m}}{p_{k+m-1}} \frac{p_{k+m-1}}{p_{k+m-2}} \cdots \frac{p_{k+2}}{p_{k+1}} p_{k+1} \leq \delta^{m-1} p_{k+1}. \tag{17}$$

Hence, whenever $k \geq k_0$, the numerator of (16) is bounded by $p_{k+1}/(1-\delta)$ and (16) follows. The equivalence of $\lim_{k\to\infty} h_k = 0$ and (16) follows immediately from

$$h_k = 1 - \frac{1}{1+r_k}, \qquad r_k = \frac{1}{1-h_k} - 1.$$

For the second statement of the lemma, we first show that both parts of (15) are equivalent to $\sum_{k=0}^{\infty} r_k < \infty$, using similar arguments as in the proof of the first statement. If the sequence $(p_{k+1}/p_k)_{k\in\mathbb{N}}$ is summable then $p_{k+1}/p_k \to 0$, and we can use the bound (17) to obtain summability of $(r_k)_{k\in\mathbb{N}}$. To obtain summability of the r-sequence from the summability of the h-sequence we use the fact that

$$\frac{1}{1-x} - 1 \leq 2x \quad \text{for } 0 \leq x \leq \frac{1}{2},$$

which implies that $r_k \leq 2h_k$ for all sufficiently large k.

3.1. Proof of Theorem 1

We first show that condition (2) implies almost-sure convergence.

Because of (1) we have $M_n \uparrow \infty$ with probability 1, i.e. $M_n(\omega) \uparrow \infty$ for all $\omega \in A$, with some $A \in \mathcal{A}$ such that $\mathrm{P}(A) = 1$. Let $(U_n)_{n\in\mathbb{N}}$ be the subsequence of strictly increasing values. Formally, we set

$$U_1(\omega) := M_1(\omega), \qquad U_{n+1}(\omega) = \min\{M_j(\omega) \colon M_j(\omega) > U_n(\omega)\} \quad \text{for all } n \in \mathbb{N}$$

for all $\omega \in A$; on the null set $\Omega \setminus A$ we may assign some arbitrary value to the sequence. Then $(U_n)_{n\in\mathbb{N}}$ is a Markov chain with state space $\mathbb{N}_0$ and transition probabilities

$$p_{j,j+l} = \mathrm{P}(X_1 = j + l \mid X_1 > j) \quad \text{for all } j \in \mathbb{N}_0,\ l \in \mathbb{N}.$$

Let B be the event that infinitely many $j \in \mathbb{N}_0$ do not appear in the U-sequence and let B_j be the event that j does, but $j+1$ does not appear. Clearly, using $M_n \uparrow \infty$ again, $B \cap A = \limsup_{j\to\infty} B_j \cap A$ and

$$\begin{aligned}
\mathrm{P}(B_j) &= \sum_{n=1}^{\infty} \mathrm{P}(U_n = j,\ U_{n+1} > j+1) \\
&= \sum_{n=1}^{\infty} \mathrm{P}(U_{n+1} > j+1 \mid U_n = j)\,\mathrm{P}(U_n = j) \\
&= \left(\sum_{l=2}^{\infty} p_{j,j+l}\right)\left(\sum_{n=1}^{\infty} \mathrm{P}(U_n = j)\right) \\
&\le \mathrm{P}(X_1 \ge j+2 \mid X_1 \ge j+1) \\
&= h_{j+1},
\end{aligned}$$

where in the penultimate step we used the fact that the events $\{U_n = j\}$, $n \in \mathbb{N}$, are disjoint. The Borel–Cantelli lemma now gives $\mathrm{P}(B) = 0$, which means that

$$\eta := \inf\{j \in \mathbb{N}\colon \{k \in \mathbb{N}\colon k \ge j\} \subset \{U_n\colon n \in \mathbb{N}\}\}$$

is finite with probability 1. We further define

$$\begin{aligned}
\rho(\omega) &:= \inf\{n \in \mathbb{N}\colon M_n(\omega) \ge \eta(\omega)\}, \\
\tau_j(\omega) &:= \inf\{n \in \mathbb{N}\colon X_n(\omega) = j\}, \qquad j = 0, 1, \ldots.
\end{aligned}$$

Again, on some set C of probability 1, all these random variables are finite. Finally, for all $\omega \in C$, we have $L_n(\omega) = 0$ for all $n \ge \max\{\tau_0(\omega), \ldots, \tau_{\eta(\omega)}(\omega)\}$. This proves that L_n converges to 0 with probability 1 as $n \to \infty$.

We now show that condition (2) is also necessary for almost-sure convergence. In the Markov chain framework let A_k be the event that $\{j \in \mathbb{N}\colon j \ge k\}$ is a subset of the range $\{U_n\colon n \in \mathbb{N}\}$ of the process of successive maxima. Let $k \in \mathbb{N}$ be given, and let $\tau := \inf\{n \in \mathbb{N}\colon U_n \ge k\}$. Using the strong Markov property and the fact that $U_\tau = k$ on A_k, we obtain

$$\begin{aligned}
\mathrm{P}(A_k) &= \sum_{l=1}^{\infty} \mathrm{P}(U_{l+j} = k+j \text{ for all } j \in \mathbb{N} \mid U_l = k)\,\mathrm{P}(\tau = l) \\
&= \sum_{l=1}^{\infty}\left(\prod_{j=0}^{\infty} \mathrm{P}(U_{l+j+1} = k+j+1 \mid U_{l+j} = k+j)\right)\mathrm{P}(\tau = l) \\
&= \prod_{j=0}^{\infty}(1 - h_{k+j+1}) \\
&\le \exp\left(-\sum_{j=k+1}^{\infty} h_j\right).
\end{aligned}$$

Hence, if $\sum_{k=0}^{\infty} h_k = \infty$ then $\mathrm{P}(A_k) = 0$ for all $k \in \mathbb{N}$, and the statement follows by noting that, for $\omega \notin \liminf_{k\to\infty} A_k$, we have $\limsup_{n\to\infty} L_n(\omega) \ge 1$.

3.2. Proof of Theorem 2

Using Chebyshev's inequality and a comment preceding the statement of Theorem 2, we see that it is enough to prove that (3) implies (4) and that (18) implies (3), with

$$\lim_{n\to\infty} \mathrm{P}(Z_n = 0) = 1. \tag{18}$$

We will use the alternative version of (3) given in Lemma 1. Let

$$A_n(j) = \bigcup_{k=1}^{n} \{X_k = j\}$$

be the event that the value j appears among the first n random variables.

For the proof of the first implication, define T_n to be the number of pairs (j, m), $j, m \in \mathbb{N}_0$, $j < m$, such that j does not appear among $X_1, \dots, X_n$ but m does. Note that $Z_n \le T_n$. Following the custom of identifying sets and their indicators we therefore have

$$Z_n \le T_n = \sum_{j<m} A_n^{\mathrm{c}}(j) \cap A_n(m),$$

which in view of the fact that

$$\mathrm{P}(A_n^{\mathrm{c}}(j) \cap A_n(m)) = \mathrm{P}\biggl(\bigcup_{r=1}^{n} \{X_r = m\} \cap \bigcap_{k\ne r} \{X_k \ne j\}\biggr) \le np_m(1-p_j)^{n-1}$$

leads to the upper bound

$$\mathrm{E}\, T_{n+1} \le (n+1) \sum_{j<m} p_m (1-p_j)^n \le \frac{n+1}{n} \sum_{j=0}^{\infty} n\mathrm{e}^{-np_j} \sum_{m>j} p_m. \tag{19}$$

We need to show that the right-hand side of (19) converges to 0 as $n \to \infty$. Let $\varepsilon > 0$ be given. From (3) and (14), there is a $j_0 \in \mathbb{N}$ such that $p_{j+1}/p_j \le \varepsilon$ for all $j \ge j_0$. Since the p_js are positive, each of the terms $n\mathrm{e}^{-np_j}$ has limit 0 as $n \to \infty$. Thus, we can further choose n_0 in dependence of ε and j_0 such that, for all $n \ge n_0$,

$$\sum_{j=0}^{j_0} n\mathrm{e}^{-np_j} \sum_{m>j} p_m \le \sum_{j=0}^{j_0} n\mathrm{e}^{-np_j} \le \varepsilon.$$

To bound the rest of the sum, note that if $m > j > j_0$ then by (17) we have $p_m \le \varepsilon^{m-j-1} p_{j+1}$, so that replacing $\sum_{m>j} p_m$ by $p_{j+1}/(1-\varepsilon)$ will increase its value. As $j \mapsto p_j$ is decreasing on $\{j \ge j_0\}$, we can next define j_n by

$$j_n := \inf\{j \ge j_0 \colon np_j \le 1\},$$

and, neglecting the unimportant multiplicative factor $1/(1-\varepsilon)$, we split the remaining sum as

$$\sum_{j_0<j<j_n} np_{j+1}\mathrm{e}^{-np_j} + \sum_{j\ge j_n} np_{j+1}\mathrm{e}^{-np_j}.$$

In view of $j_n \geq j_0$ we can bound the second sum by

$$\varepsilon \sum_{j \geq j_n} np_j \mathrm{e}^{-np_j} \leq \varepsilon n \sum_{j \geq j_n} p_j \leq \varepsilon n p_{j_n}(1 + \varepsilon + \varepsilon^2 + \cdots) \leq \frac{\varepsilon}{1 - \varepsilon}.$$

We now consider the range $j_0 < j < j_n$. Again, by (17) we have

$$p_{j_n - 1} \leq \varepsilon^{j_n - 1 - j} p_j,$$

so that

$$np_j \geq \varepsilon^{j - (j_n - 1)} np_{j_n - 1} > 1,$$

where the last inequality follows from the definition of j_n and the fact that the exponent is nonpositive in our range of js. Since the function $x\mathrm{e}^{-x}$ is decreasing for $x > 1$, we have

$$np_j \exp(-np_j) \leq \varepsilon^{j - (j_n - 1)} np_{j_n - 1} \exp(-\varepsilon^{j - (j_n - 1)} np_{j_n - 1}),$$

and, therefore,

$$\begin{aligned}
\sum_{j_0 < j < j_n} np_{j+1} \exp(-np_j) &\leq \varepsilon \sum_{j_0 < j < j_n} np_j \exp(-np_j) \\
&\leq \varepsilon \sum_{j_0 < j < j_n} \varepsilon^{j - (j_n - 1)} np_{j_n - 1} \exp(-\varepsilon^{j - (j_n - 1)} np_{j_n - 1}) \\
&\leq \varepsilon \sum_{k \geq 0} \varepsilon^{-k} np_{j_n - 1} \exp(-\varepsilon^{-k} np_{j_n - 1}) \\
&\leq \varepsilon \left(\exp(-1) + \int_0^\infty \varepsilon^{-x} np_{j_n - 1} \exp(-\varepsilon^{-x} np_{j_n - 1}) \, \mathrm{d}x \right).
\end{aligned}$$

Changing variables to $y = \varepsilon^{-x} np_{j_n - 1}$ leads to the value $\exp(-np_{j_n - 1})/\log(1/\varepsilon)$ for the integral. This completes the proof that (3) implies (4).

For the proof that (18) implies (3), we first note that $Z_n \geq 1$ on

$$A_n^{\mathrm{c}}(j) \cap A_n(j+1) \cap \{m_n < j\} \supset A_n^{\mathrm{c}}(j) \cap A_n(j+1) \cap A_n(0), \qquad j \geq 1,$$

so that

$$\begin{aligned}
\mathrm{P}(Z_n \geq 1) &\geq \mathrm{P}(A_n^{\mathrm{c}}(j) \cap A_n(j+1)) - \mathrm{P}(A_n^{\mathrm{c}}(0)) \\
&= \mathrm{P}(A_n^{\mathrm{c}}(j)) - \mathrm{P}(A_n^{\mathrm{c}}(j) \cap A_n^{\mathrm{c}}(j+1)) - (1 - p_0)^n \\
&= (1 - p_j)^n - (1 - p_j - p_{j+1})^n - (1 - p_0)^n.
\end{aligned}$$

Suppose now that (3) does not hold. Then, by Lemma 1, p_{j+1}/p_j does not converge to 0 as $j \to \infty$, so we can find a $\delta > 0$ and an increasing sequence $(j_k)_{k \in \mathbb{N}} \subset \mathbb{N}$ such that

$$\frac{p_{j_k + 1}}{p_{j_k}} \geq \delta \quad \text{for all } k \in \mathbb{N},$$

and with $n_k := \lceil 1/p_{j_k} \rceil$ we would obtain

$$\begin{aligned}
\limsup_{n \to \infty} \mathrm{P}(Z_n \geq 1) &\geq \liminf_{k \to \infty} ((1 - p_{j_k})^{n_k} - (1 - p_{j_k} - p_{j_k + 1})^{n_k} - (1 - p_0)^{n_k}) \\
&\geq \mathrm{e}^{-1} - \mathrm{e}^{-1 - \delta} \\
&> 0,
\end{aligned}$$

which contradicts (18).

3.3. Proof of Theorem 5

Let $(E_i)_{i\in\mathbb{N}}$ be a sequence of independent, standard exponential random variables. It is well known that geometric random variables can be obtained from exponentially distributed random variables by discretization. The sequence $(X_i)_{i\in\mathbb{N}}$ we are interested in is equal in distribution to the sequence $(\phi_p(E_i))_{i\in\mathbb{N}}$ (see (10)). Next, let $E_{(n:i)}$, $1 \le i \le n$, be the (ascending) order statistics associated with the first n of the E-variables, i.e.

$$E_{(n:1)} < E_{(n:2)} < \cdots < E_{(n:n)}, \qquad \{E_{(n:i)} : i = 1, \dots, n\} = \{E_i : i = 1, \dots, n\}.$$

By the Sukhatme–Rényi representation (see, e.g. [10, p. 721]),

$$\mathcal{L}((E_{(n:1)}, \dots, E_{(n:n)})) = \mathcal{L}((V_n, V_n + V_{n-1}, \dots, V_n + \cdots + V_1))$$

for all $n \in \mathbb{N}$, with $(V_i)_{i\in\mathbb{N}}$ as in Subsection 2.2. Applying ϕ_p to the components of these vectors we obtain a representation for the order statistics associated with the first n X-variables. These in turn give the elements of A_n in increasing order, after an obvious reduction step that does not change the gaps. With $W_{l,n}$ as defined in Subsection 2.2 we therefore have

$$\mathcal{L}((X_{(n:1)}, \dots, X_{(n:n)})) = \mathcal{L}((\phi_p(W_{n,n}), \dots, \phi_p(W_{1,n}))),$$

which implies that the variable L_n in the theorem has the same distribution as

$$L'_n := \max\{\phi_p(W_{l,n}) - \phi_p(W_{l+1,n}) : l = 1, \dots, n-1\}. \tag{20}$$

It should be noted that this representation refers to the individual random variables only and not to any joint distributions of more than one of the L_ns.

From (10), it follows that

$$x - c(p) \le c(p)\phi_p(x) \le x,$$

so that

$$V_l - c(p) \le c(p)(\phi_p(W_{l,n}) - \phi_p(W_{l+1,n})) \le V_l + c(p).$$

Using (20), we now obtain (11).

For the proof of the second part of the theorem, we first note that (see (10))

$$c(p)(\phi_p(x) + \psi_p(x)) = x \quad \text{for all } x \in \mathbb{R},$$

which gives

$$c(p)(\phi_p(W_{l,n}) - \phi_p(W_{l+1,n})) = V_l + c(p)(\psi_p(W_{l+1,n}) - \psi_p(W_{l,n})).$$

Suppose now that $\psi_p(\log n_m) \to \eta$. The limiting random variables $Z_{l,\infty}$, $l \in \mathbb{N}$, have continuous distribution functions. Since $\psi_p(x + c(p)k) = \psi_p(x)$ for all $k \in \mathbb{Z}$ and as both functions are continuous outside the countable set $c(p)\mathbb{Z}$, we obtain, with probability 1,

$$\psi_p(W_{l,n_m}) = \psi_p(Z_{l,n_m} + c(p)\psi_p(\log n_m)) \to \psi_p(Z_{l,\infty} + c(p)\eta)$$

as $m \to \infty$. Together with an elementary analytic argument about maxima and limits this gives the second assertion of the theorem.

Finally, we note that the maximum in (12) is taken over quantities of the form

$$a + \{b\} - \{a + b\},$$

which is equal to either $\lfloor a \rfloor$ or $\lceil a \rceil$. This substantiates the remark following Theorem 5 and also leads to the upper bound in (13). The lower bound in (13) follows immediately from the lower bound in (11) and the weak convergence.

3.4. Proof of Theorem 6

As in the proof of (12), we find that the limiting probability that the longest gap arises as the difference between the two largest sample values is equal to the probability of the event

$$A_{p,\eta} := \{R_{1,p,\eta} \geq R_{l,p,\eta} \text{ for all } l \geq 2\}$$

with

$$R_{l,p,\eta} := c(p)^{-1} V_l + \psi_p(Z_{l+1,\infty} + c(p)\eta) - \psi_p(Z_{l,\infty} + c(p)\eta).$$

Once again, using the continuity of the respective distribution functions, this leads to

$$\begin{aligned}
\lim_{p\to 0} \mathrm{P}(A_{p,\eta}) &= \mathrm{P}(V_1 > V_l \text{ for all } l \geq 2) \\
&= \mathrm{P}(V_1 > M_{2,\infty}) \\
&= \int_0^\infty \mathrm{e}^{-x} \prod_{k=2}^{\infty} (1 - \mathrm{e}^{-kx})\,\mathrm{d}x \\
&= \int_0^\infty (\mathrm{P}(M_{2,\infty} \leq x) - \mathrm{P}(M_{1,\infty} \leq x))\,\mathrm{d}x \\
&= \int_0^\infty (\mathrm{P}(M_{1,\infty} \geq x) - \mathrm{P}(M_{2,\infty} \geq x))\,\mathrm{d}x \\
&= \mathrm{E}\, M_{1,\infty} - \mathrm{E}\, M_{2,\infty}.
\end{aligned}$$

The numerical evaluation of the integral in the third line is straightforward.

3.5. Proof of Theorem 3

We recall the definition of the quantile function F^{-1} associated with a distribution function F,

$$F^{-1}(y) = \inf\{x \in \mathbb{R} \colon F(x) \geq y\}, \qquad 0 < y < 1. \tag{21}$$

It is well known that the random variable $Y = F^{-1}(U)$ has distribution function F if U is uniformly distributed on the unit interval. Similarly, $Y = \Psi(V)$, with

$$\Psi(y) := F^{-1}(1 - \mathrm{e}^{-y}), \qquad y > 0,$$

has distribution F if V is exponentially distributed with mean 1. We need an auxiliary result.

Lemma 2. *If (5) holds then*

$$\lim_{w\to\infty} (\Psi(w + v) - \Psi(w)) = \infty \quad \textit{for all } v > 0. \tag{22}$$

Proof. We have

$$\Psi(w + v) - \Psi(w) = F^{-1}(1 - \mathrm{e}^{-v}\mathrm{e}^{-w}) - F^{-1}(1 - \mathrm{e}^{-w}),$$

which means that (22) follows once we have shown that, for all $\eta < 1$ and $a > 0$, there exists a $y_0 > 0$ such that, for all $y \leq y_0$,

$$F^{-1}(1 - \eta y) \geq F^{-1}(1 - y) + a. \tag{23}$$

Now suppose that η and a are given. Choose $\delta = \delta(\eta, a) > 0$ such that

$$\left| \frac{\log \eta}{\log(1-\delta)} \right| > a + 1. \tag{24}$$

Because of $q_{k+1}/q_k \to 1$ as $k \to \infty$, we can further choose $k_0 = k_0(\delta)$ such that

$$\frac{q_{k+1}}{q_k} \geq 1 - \delta \quad \text{for all } k \geq k_0.$$

Now set $y_0 := q_{k_0+1}$. We claim that, with these choices,

$$\inf\{k : q_{k+1} \leq \eta y\} \geq \inf\{k : q_{k+1} \leq y\} + a \quad \text{for all } y \leq y_0. \tag{25}$$

By the definition of the quantile function, (21), this would imply (23).

For the proof of (25), we set $k_1 = k_1(y) := \inf\{k : q_{k+1} \leq y\}$. Clearly, $q_{k_1} > y$ and $k_1 \geq k_0$ in view of $q_{k_1+1} \leq y \leq y_0 = q_{k_0+1}$. Hence, for all $l \in \mathbb{N}$,

$$q_{k_1+l+1} = q_{k_1} \prod_{j=0}^{l} \frac{q_{k_1+j+1}}{q_{k_1+j}} \geq y(1-\delta)^{l+1},$$

so that, for q_{k_1+l} not to exceed ηy, we need $(1-\delta)^{l+1} \leq \eta$. From this, (25) follows by using (24).

With the exponential quantile function Ψ and the Sukhatme–Rényi representation (see Subsection 3.3), we find that the gap between the maximum and the second largest of the first n of the X-variables, and, hence, the length of the longest gap, is bounded from below by

$$\Psi(W_{2,n} + V_1) - \Psi(W_{2,n}) - 1,$$

with $W_{2,n}$ and V_1 as defined in Section 2. In the representation we have $V_1 > 0$ and $W_{2,n} \to \infty$ as $n \to \infty$, both with probability 1. This, together with Lemma 2, yields the assertion of the theorem.

3.6. Proof of Theorem 4

We first note that Y_n is 1 less than

$$\sum_{j=0}^{\infty} A_n^{\mathrm{c}}(j) \cap A_n(j+1)$$

(the extra 1 being for the smallest value in the sample), and the probability of the latter event is

$$(1-p_j)^n - (1-p_j-p_{j+1})^n.$$

Furthermore,

$$\mathrm{e}^{-np_j} - (1-p_j)^n \leq \frac{\mathrm{e}}{2} p_j^2 n \mathrm{e}^{-np_j} \leq \frac{p_j}{2},$$

as $x\mathrm{e}^{-x} \leq \mathrm{e}^{-1}$ for $x \geq 0$. It follows that the difference between

$$\sum_{j=0}^{\infty} ((1-p_j)^n - (1-p_j-p_{j+1})^n) \quad \text{and} \quad \sum_{j=0}^{\infty} \mathrm{e}^{-p_j n}(1-\mathrm{e}^{-p_{j+1} n})$$

is $O(1)$, so it suffices to approximate the latter sum. Under our assumptions, it is of the same order as

$$\sum_{j=1}^{\infty} \mathrm{e}^{-n/j^\alpha}(1-\mathrm{e}^{-n/(j+1)^\alpha}) = \sum_{j=1}^{\infty} \mathrm{e}^{-n/j^\alpha}(1-\mathrm{e}^{-n/j^\alpha}) + \sum_{j=1}^{\infty}(\mathrm{e}^{-n/j^\alpha}-\mathrm{e}^{-n/(j+1)^\alpha})\mathrm{e}^{-n/j^\alpha}.$$

The second sum is (termwise) nonpositive and bounded below by

$$\sum_{j=1}^{\infty}(\mathrm{e}^{-n/j^\alpha}-\mathrm{e}^{-n/(j+1)^\alpha}) \geq -1,$$

by telescoping. Regarding the first sum as a Riemann approximation to the integral

$$\int_1^\infty \mathrm{e}^{-n/x^\alpha}(1-\mathrm{e}^{-n/x^\alpha})\,\mathrm{d}x$$

and changing variables to $y = n/x^\alpha$, we see that it is asymptotic to

$$\frac{n^{1/\alpha}}{\alpha}\int_0^\infty \frac{\mathrm{e}^{-y}}{y^{1+1/\alpha}}(1-\mathrm{e}^{-y})\,\mathrm{d}y$$

as $n \to \infty$. Replacing the Riemann sum by the integral introduces an $O(1)$ error since $\mathrm{e}^{-y}(1-\mathrm{e}^{-y})$ is bounded for $y \geq 0$, increasing up to $y_0 = \ln 2$ and decreasing afterwards. This gives (6).

Acknowledgements

The first author would like to thank Stephan Klawunn for his pertinent remarks on the first version of the paper. We address our special thanks to the anonymous referees, whose detailed and thorough comments significantly improved the paper.

References

[1] Bogachev, L. V., Gnedin, A. V. and Yakubovich, Y. V. (2008). On the variance of the number of occupied boxes. *Adv. Appl. Math.* **40,** 401–432.
[2] Bruss, F. T. and Grübel, R. (2003). On the multiplicity of the maximum in a discrete random sample. *Ann. Appl. Prob.* **13,** 1252–1263.
[3] Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Prob. Surveys* **4,** 146–171.
[4] Goh, W. M. Y. and Hitczenko, P. (2007). Gaps in samples of geometrically distributed random variables. *Discrete Math.* **307,** 2871–2890.
[5] Grübel, R. (2007). Distributional asymptotics in the analysis of algorithms: Periodicities and discretization. *Discrete Math. Theoret. Comput. Sci.* **AH,** 451–460.
[6] Hitczenko, P. and Knopfmacher, A. (2005). Gap-free compositions and gap-free samples of geometric random variables. *Discrete Math.* **294,** 225–239.
[7] Hwang, H.-K. and Janson, S. (2008). Local limit theorems for finite and infinite urn models. *Ann. Prob.* **36,** 992–1022.
[8] Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17,** 373–401.
[9] Louchard, G. and Prodinger, H. (2008). On gaps and unoccupied urns in sequences of geometrically distributed random variables. *Discrete Math.* **308,** 1538–1562.
[10] Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. John Wiley, New York.