*Bernoulli* **23**(1), 2017, 249–287
DOI: 10.3150/15-BEJ743

# Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications

ANNA BEN-HAMOU[1], STÉPHANE BOUCHERON[1,2] and MESROB I. OHANNESSIAN[3]

[1]*LPMA UMR 7599 CNRS & Université Paris-Diderot, 5, rue Thomas Mann 75013, Paris, France.*
*E-mail: anna.benhamou@univ-paris-diderot.fr*
[2]*DMA ENS Ulm, 45, rue d'Ulm 75005, Paris, France.*
*E-mail: stephane.boucheron@univ-paris-diderot.fr; url: stephane-v-boucheron.fr*
[3]*University of California, San Diego, Atkinson Hall, 9500 Gilman Dr, La Jolla, CA 92093, USA.*
*E-mail: mesrob@gmail.com; url: sites.google.com/site/mesrob*

An infinite urn scheme is defined by a probability mass function $(p_j)_{j\geq 1}$ over positive integers. A random allocation consists of a sample of $N$ independent drawings according to this probability distribution where $N$ may be deterministic or Poisson-distributed. This paper is concerned with occupancy counts, that is with the number of symbols with $r$ or at least $r$ occurrences in the sample, and with the missing mass that is the total probability of all symbols that do not occur in the sample. Without any further assumption on the sampling distribution, these random quantities are shown to satisfy Bernstein-type concentration inequalities. The variance factors in these concentration inequalities are shown to be tight if the sampling distribution satisfies a regular variation property. This regular variation property reads as follows. Let the number of symbols with probability larger than $x$ be $\vec{\nu}(x) = |\{j : p_j \geq x\}|$. In a regularly varying urn scheme, $\vec{\nu}$ satisfies $\lim_{\tau\to 0} \vec{\nu}(\tau x)/\vec{\nu}(\tau) = x^{-\alpha}$ for $\alpha \in [0, 1]$ and the variance of the number of distinct symbols in a sample tends to infinity as the sample size tends to infinity. Among other applications, these concentration inequalities allow us to derive tight confidence intervals for the Good–Turing estimator of the missing mass.

*Keywords:* concentration; missing mass; occupancy; rare species; regular variation

## 1. Introduction

From the 20th century to the 21st, various disciplines have tried to infer something about scarcely observed events: zoologists about species, cryptologists about cyphers, linguists about vocabularies, and data scientists about almost everything. These problems are all about 'small data' within possibly much bigger data. Can we make such inference?

*Problem setting*. To move into a concrete setting, let $U_1, U_2, \ldots, U_n$ be i.i.d. observations from a fixed but unknown distribution $(p_j)_{j=1}^{\infty}$ over a discrete set of symbols $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$. We consider each $j$ in $\mathbb{N}_+$ as a discrete *symbol* devoid of numerical significance. The terminology of 'infinite urn scheme' comes from the analogy to $n$ independent throws of balls over an infinity of urns, $p_j$ being the probability of a ball falling into urn $j$, at any $i$th throw. We alternatively adhere to the symbols or the urns perspective, based on which carries the intuition best. Species, cyphers, and vocabularies all being discrete, are well modeled as such. The sample size $n$ may be fixed in advance; we call this the *binomial setting*. It may be randomly set by the duration of an experiment; this gives rise to the *Poisson setting*. More precisely, in the latter case we write it as $N$, a Poisson random variable independent of $(U_i)$ and with expectation $t$. We index all Poisson-setting quantities by $t$ and write them with functional notations, instead of subscripts used for the fixed-$n$ scheme.

For each $j, n \in \mathbb{N}_+$, let $X_{n,j} = \sum_{i=1}^{n} \mathbb{I}_{\{U_i=j\}}$ be the number of times symbol $j$ occurs in a sample of size $n$, and $X_j(t) = \sum_{i=1}^{N(t)} \mathbb{I}_{\{U_i=j\}}$ the Poisson version. In questions of underrepresented data, the central objects are sets of symbols that are repeated a small number $r$ of times. The central quantities are the *occupancy counts* $K_{n,r}$ [respectively, $K_r(t)$ for the Poisson setting], defined as the number of symbols that appear exactly $r$ times in a sample of size $n$:

$$K_{n,r} = \big|\{j, X_{n,j} = r\}\big| = \sum_{j=1}^{\infty} \mathbb{I}_{\{X_{n,j}=r\}}.$$

The collection $(K_{n,r})_{r\geq 1}$ [resp. $(K_r(t))_{r\geq 1}$] has been given many names, such as the "profile" (in information theory (Orlitsky *et al.* [36])) or the "fingerprint" (in theoretical computer science (Batu *et al.* [6], Valiant and Valiant [40])) of the probability distribution $(p_j)_{j\in\mathbb{N}_+}$. Here we refer to them by *occupancy counts* individually, and *occupancy process* all together.

The occupancy counts then combine to yield the cumulated occupancy counts $K_{n,\overline{r}}$ [respectively $K_{\overline{r}}(t)$] and the total number of distinct symbols in the sample, or the total number of occupied urns, often called the *coverage* and denoted by $K_n$ [respectively $K(t)$]:

$$K_{n,\overline{r}} = \big|\{j, X_{n,j} \geq r\}\big| = \sum_{j=1}^{\infty} \mathbb{I}_{\{X_{n,j}\geq r\}} = \sum_{s\geq r} K_{n,s}$$

and

$$K_n = \big|\{j, X_{n,j} > 0\}\big| = \sum_{j=1}^{\infty} \mathbb{I}_{\{X_{n,j}>0\}} = \sum_{r\geq 1} K_{n,r}.$$

In addition to the occupancy numbers and the number of distinct symbols, we also address the *rare* (or small-count) *probabilities* $M_{n,r}$ [respectively, $M_r(t)$], defined as the probability mass corresponding to all symbols that appear exactly $r$ times:

$$M_{n,r} = \mathbb{P}\big(\{j, X_{n,j} = r\}\big) = \sum_{j=1}^{\infty} p_j \mathbb{I}_{\{X_{n,j}=r\}}.$$

In particular, we focus on $M_{n,0} = \sum_{j=1}^{\infty} p_j \mathbb{I}_{\{X_{n,j}=0\}}$, which is called the missing mass, and which corresponds to the probability of all the unseen symbols.

Explicit formulas for the moments of the occupancy counts and masses can be derived in the binomial and Poisson settings. The occupancy counts' expectations are given by

$$\mathbb{E}K_n = \sum_{j=1}^{\infty}\bigl(1-(1-p_j)^n\bigr), \qquad \mathbb{E}K(t) = \sum_{j=1}^{\infty}\bigl(1-\mathrm{e}^{-tp_j}\bigr),$$

$$\mathbb{E}K_{n,r} = \sum_{j=1}^{\infty}\binom{n}{r} p_j^r(1-p_j)^{n-r}, \qquad \mathbb{E}K_r(t) = \sum_{j=1}^{\infty}\mathrm{e}^{-tp_j}\frac{(tp_j)^r}{r!},$$

$$\mathbb{E}M_{n,r} = \sum_{j=1}^{\infty}\binom{n}{r} p_j^{r+1}(1-p_j)^{n-r}, \qquad \mathbb{E}M_r(t) = \sum_{j=1}^{\infty}\mathrm{e}^{-tp_j}p_j\frac{(tp_j)^r}{r!}.$$

Formulas for higher moments can also be computed explicitly but their expression, especially in the binomial setting where a lot of dependencies are involved, often has an impractical form.

This classical occupancy setting naturally models a host of different application areas, including computational linguistics, ecology, and biology. Urns may represent species, and we are interested in the number of distinct species observed in a sample (the ecological diversity) or in the probability of the unobserved species. In linguistics, urns may represent words. In both of these applications, the independence assumption of the random variables $\{U_i\}_{i=1,\dots,n}$ may seem unrealistic. For instance in a sentence, the probability of appearance of a word strongly depends on the previous words, both for grammatical and semantic reasons. Likewise, the nucleotides in a DNA sequence do not form an i.i.d. sample. In $n$-gram models, independence is only conditional and the observations are assumed to satisfy a Markovian hypothesis: the probability of occurrence of a word depends on the $n-1$ previous words. But the i.i.d. case, although very simple, yields results that are interesting in themselves, and upon which a more sophisticated framework may be built.

Many practical questions may now be formulated in this setting. If we double the duration of a first experiment, how many yet unobserved specimens would we find (how does $K_{2n,r}$ compare to $K_{n,r}$ [resp. $K_r(2t)$ to $K_r(t)$]) (Fisher *et al.* [21])? If certain cypher keys have been observed, what is the probability for the next to be different (how does one estimate $M_{n,0}$)? For instance, Good [24] and Turing observed that $(n+1)\mathbb{E}M_{n,0} = \mathbb{E}K_{n+1,1}$ for all $n \geq 1$, and proposed to estimate the missing mass using the Jackknife estimator $G_{n,0} = K_{n,1}/n$ (the proportion of symbols seen just once).

*Contributions*. To study the Good–Turing estimator or other quantities that depend significantly on the small-count portion of the observations, we need to understand the occupancy counts well. Our contribution here is to give sharp concentration inequalities with explicit and reasonable constants, for $K_n$, $K_{n,r}$, and $M_{n,0}$ [resp. $K(t)$, $K_r(t)$, $M_0(t)$]. We give distribution-free results, and then exhibit a vast domain where these results are tight, namely the domain of distributions with a heavier tail than the geometric. In this domain, the non-asymptotic exponential concentration properties that we establish are sharp in the sense that the exponents are order-optimal, precisely capturing the scale of the variance. For this reason, we dedicate a portion of the paper to establishing bounds on various variances.

*Organization*. The paper is organized as follows. In Section 2, we present our terminology and give a concise summary of the results. In Section 3 we present our variance bounds and concentration results for the occupancy counts and the missing mass in great generality. In Section 4, we specialize these results to regularly varying distributions, the aforementioned domain of distributions where concentration can be characterized tightly. We then elaborate on some applications in Section 5, and conclude with a discussion of the results and possible extensions in Section 6. We group all proofs in the end, in Section 7.

## 2. Summary of results

*Terminology*. Our concentration results mostly take the form of bounds on the log-Laplace transform. Our terminology follows closely (Boucheron *et al.* [13]). We say that the random variable $Z$ is *sub-Gaussian* on the right tail (resp. on the left tail) with variance factor $v$ if, for all $\lambda \geq 0$ (resp. $\lambda \leq 0$),

$$\log \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} \leq \frac{v\lambda^2}{2}. \tag{2.1}$$

We say that a random variable $Z$ is *sub-Poisson* with variance factor $v$ if, for all $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} \leq v\phi(\lambda), \tag{2.2}$$

with $\phi : \lambda \mapsto e^{\lambda} - \lambda - 1$.

We say that a random variable $Z$ is *sub-gamma on the right tail* with variance factor $v$ and scale parameter $c$ if

$$\log \mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq \frac{\lambda^2 v}{2(1-c\lambda)} \qquad \text{for every } \lambda \text{ such that } 0 \leq \lambda \leq 1/c. \tag{2.3}$$

The random variable $Z$ is *sub-gamma on the left tail* with variance factor $v$ and scale parameter $c$, if $-Z$ is sub-gamma on the right tail with variance factor $v$ and scale parameter $c$. If $Z$ is sub-Poisson with variance factor $v$, then it is sub-Gaussian on the left tail with variance factor $v$, and sub-gamma on the right tail with variance factor $v$ and scale parameter $1/3$.

These log-Laplace upper bounds then imply exponential tail bounds. For instance, inequality (2.3) results in a Bernstein-type inequality for the right tail, that is, for $s > 0$ our inequalities have the form

$$\mathbb{P}\left\{Z > \mathbb{E}[Z] + \sqrt{2vs} + cs\right\} \leq e^{-s},$$

while inequality (2.1) for all $\lambda \leq 0$ entails

$$\mathbb{P}\left\{Z < \mathbb{E}[Z] - \sqrt{2vs}\right\} \leq e^{-s}.$$

We present such results first without making distributional assumptions, beyond the structure of those quantities themselves. These concentrations then specialize in various settings, such as that of regular variation.

*Main results*. We proceed by giving a coarse description of our main results. In the Poisson setting, for each $r \geq 1$, $(\mathbb{I}\{X_j(t) = r\})_{j \geq 1}$ are independent, hence $K_r(t)$ is a sum of independent Bernoulli random variables, and it is not too surprising that it satisfies sub-Poisson, also known as Bennett, inequalities. For $\lambda \in \mathbb{R}$, we have:

$$\log \mathbb{E}\mathrm{e}^{\lambda(K_r(t)-\mathbb{E}K_r(t))} \leq \operatorname{var}\bigl(K_r(t)\bigr)\phi(\lambda) \leq \mathbb{E}\bigl[K_r(t)\bigr]\phi(\lambda).$$

The proofs are elementary and are based on the careful application of Efron–Stein–Steele inequalities and the entropy method (Boucheron *et al.* [13]).

As for the binomial setting, the summands are not independent but we can use negative association arguments (Dubhashi and Ranjan [17]) (see Section 7) to obtain Bennett inequalities for the cumulated occupancy counts $K_{n,\overline{r}}$. These hold either with the Jackknife variance proxy given by the Efron–Stein inequality, $r\mathbb{E}K_{n,r}$ or with the variance proxy stemming from the negative correlation of the summands, $\mathbb{E}K_{n,\overline{r}}$. Letting $v_{n,\overline{r}} = \min(r\mathbb{E}K_{n,r}, \mathbb{E}K_{n,\overline{r}})$, we have, for all $\lambda \in \mathbb{R}$:

$$\log \mathbb{E}\mathrm{e}^{\lambda(K_{n,\overline{r}}-\mathbb{E}K_{n,\overline{r}})} \leq v_{n,\overline{r}}\phi(\lambda).$$

This in turn implies a concentration inequality for $K_{n,r}$. Letting

$$v_{n,r} = 2\min\bigl(\max\bigl(r\mathbb{E}K_{n,r}, (r+1)\mathbb{E}K_{n,r+1}\bigr), \mathbb{E}K_{n,\overline{r}}\bigr),$$

we have, for all $s \geq 0$,

$$\mathbb{P}\bigl\{|K_{n,r} - \mathbb{E}K_{n,r}| \geq \sqrt{4v_{n,r}s} + 2s/3\bigr\} \leq 4\mathrm{e}^{-s}.$$

We obtain distribution-free bounds on the log-Laplace transform of $M_{n,0}$, which result in sub-Gaussian concentration on the left tail, sub-gamma concentration on the right tail with scale proxy $1/n$. More precisely, letting $v_n^- = 2\mathbb{E}K_2(n)/n^2$ and $v_n^+ = 2\mathbb{E}K_{\overline{2}}(n)/n^2$, we show that, for all $\lambda \leq 0$,

$$\log \mathbb{E}\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})} \leq v_n^- \frac{\lambda^2}{2},$$

and, for all $\lambda \geq 0$,

$$\log \mathbb{E}\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})} \leq v_n^+ \frac{\lambda^2}{2(1-\lambda/n)}.$$

Indeed, these results are distribution-free. But though the variance factor $v_n^-$ is a sharp bound for the variance of the missing mass, $v_n^+$ may be much larger. This leads us to look for distribution-specific conditions ensuring that $v_n^+$ captures the right order for the variance, such as by using a tail asymptotic stability condition as in extreme value theory.

Karlin [29] pioneered such a condition by assuming that the function $\vec{\nu} : (0, 1] \to \mathbb{N}$, defined by $\vec{\nu}(x) = |\{j \in \mathbb{N}_+, p_j \geq x\}|$ satisfies a regular variation assumption, namely $\vec{\nu}(1/n) \sim n^\alpha \ell(n)$ near $+\infty$, with $\alpha \in (0, 1]$ (see also Gnedin *et al.* [22], Ohannessian and Dahleh [35]). Here $\ell$ is a slowly varying function at $\infty$, i.e. for all $x$, $\ell(\tau x)/\ell(\tau) \to 1$ as $\tau \to \infty$. This condition allows us to compare the asymptotics of the various occupancy scores. In particular, in this framework $\mathbb{E}K_2(n)$ and $\mathbb{E}K_{\overline{2}}(n)$ have the same order of growth, and, divided by $n^2$ they both are of the same order as the variance of the missing mass. Hence, regular variation provides a framework in which our concentration inequalities are order-optimal.

To handle the case $\alpha = 0$, we move from *Karamata* to *de Haan* theory, and take $\vec{\nu}$ to have an extended regular variation property, with the additional assumption that $\mathbb{E}K_1(n)$ tends to $+\infty$. This domain corresponds to light-tailed distributions which are still heavier than the geometric. In this case, we manage to show the sub-gamma concentration of the missing mass only for $n$ large enough, that is, that there exists $n_0$ such that for all $n \geq n_0$, for $\lambda > 0$, we have $\log \mathbb{E}e^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})} \leq (v_n\lambda^2)/2(1-\lambda/n)$, with $v_n \asymp \operatorname{Var} M_{n,0}$.

Back to our examples of applications, considerable insight may be gained from these concentration results. For instance, heavy tails lead to multiplicative concentration for $M_{n,0}$ and the consistency of the Good–Turing estimator: $\frac{G_{n,0}}{M_{n,0}} \xrightarrow{p} 1$. Generally, new estimators can be derived and shown to be consistent in a unified framework, once one is able to estimate $\alpha$ consistently. For instance, when $\vec{\nu}(1/\cdot)$ is regularly varying with index $\alpha$, $\hat{\alpha} = K_{n,1}/K_n$ is a consistent estimator of $\alpha$. Then, to estimate the number of new species in a sample twice the size of the original, we immediately get that $\widehat{K}_{2n} = K_n + \frac{2^{\widehat{\alpha}}-1}{\widehat{\alpha}}K_{n,1}$ is a consistent estimator of $K_{2n}$. This methodology is very similar to extreme value theory (Beirlant *et al.* [7]): harnessing limiting expressions and tail parameter estimation. These results strengthen and extend the contribution of Ohannessian and Dahleh [35], which is restricted to power-laws and implicit constants in the inequalities. Beyond consistency results, we also obtain confidence intervals for the Good–Turing estimator in the Poisson setting, using empirical quantities.

*Historical notes and related work*. There exists a vast literature on the occupancy scheme, as formulated here and in many other variations. The most studied problems are the asymptotic behavior of $K_n$ and $K_{n,r}$. This is done often in a finite context, or a scaling model where probabilities remain mostly uniform. Of particular relevance to this paper, we mention the work of Karlin [29], who built on earlier work by Bahadur [3], credited as one of the first to study the infinite occupancy scheme. Karlin's main results were to establish central limit theorems in an infinite setting, under a condition of regular variation. He also derived strong laws of large numbers. Gnedin *et al.* [22] present a general review of these earlier results, as well as more contemporary work on this problem. The focus continues to be central limit theorems, or generally asymptotic results. For example, the work of Hwang and Janson [28] (effectively) provides a local limit theorem for $K_n$ provided that the variance tends to infinity. Somewhat less asymptotic results have also been proposed, in the form of normal approximations, such as in the work of Bogachev *et al.* [12] and Barbour and Gnedin [4].

Besides occupancy counts analysis, a distinct literature investigates the number of species and missing mass problems. These originated in the works of Fisher *et al.* [21], Good [24], and Good and Toulmin [25], and generated a long line of research to this day (Bunge and Fitzpatrick [14]). Here, instead of characterizing the asymptotic behavior of these quantities, one is interested in estimating $K_{\lambda n} - K_n$ for a $\lambda > 1$, that is the number of discoveries when the sample size is multiplied by $\lambda$, or estimating $M_{n,0}$: estimators are proposed, and then their statistical properties are studied. One recently studied property by McAllester and Schapire [33] and McAllester and Ortiz [32], is that of concentration, which sets itself apart from the CLT-type results in that it is non-asymptotic in nature. Based on this, Ohannessian and Dahleh [35] showed that in the regular variation setting of Karlin, one could show multiplicative concentration, and establish strong consistency results. Conversely, characterizing various aspects of concentration allows one to systematically design new estimators. For example, this was illustrated in Ohannessian and Dahleh [35] for the estimation of rare probabilities, to both justify and extend Good's (Good [24]) work that remains relevant in some of the aforementioned applications, especially computational linguistics. Such concentration results for rare probabilities have been also used in the general probability estimation problem, such as by Acharya *et al.* [1].

# 3. Distribution-free concentration

## 3.1. Occupancy counts

3.1.1. *Variance bounds*

In order to understand the fluctuations of occupancy counts $K_n$, $K(t)$, $K_{n,r}$, $K_r(t)$, we start by reviewing and stating variance bounds. We start with the Poisson setting where occupancy counts are sums of independent Bernoulli random variables with possibly different success probabilities, and thus variance computations are straightforward. Exact expressions may be derived (see, for example, Gnedin *et al.* [22], equation (4)), but ingenuity may be used to derive more tractable and tight bounds. We start by stating a well-known connection between the variance of the number of occupied urns and the expected number of singletons (Gnedin *et al.* [22], Karlin [29]). In the binomial setting, similar bounds can be derived using the Efron–Stein–Steele inequalities, outlined in Section 7.1.1 (see Boucheron *et al.* [13], Section 3.1).

**Proposition 3.1.** *In the Poissonized setting, we have*

$$\frac{\mathbb{E}K_1(2t)}{2} \le \operatorname{Var}\bigl(K(t)\bigr) \le \mathbb{E}K_1(t).$$

*In the binomial setting, we have*

$$\operatorname{Var}(K_n) \le \mathbb{E}\bigl[K_{n,1}(1 - M_{n,0})\bigr] \le \mathbb{E}K_{n,1}.$$

The upper bounds in these two propositions parallel each other, but in the binomial setting, we cannot hope to establish lower bounds like $\mathbb{E}K_{cn,1}/c \le \operatorname{Var}(K_n)$ for some constant $c > 0$ in full generality. To see this, consider the following example which shows that the variance of $K(t)$ and of $K_n$ may differ significantly, and that the variance of $K_n$ may be much smaller than the expected value of $K_{n,1}$.

***Example 1.*** In the so-called *birthday paradox* scenario, $n$ balls are thrown independently into $n^2$ urns with uniform probabilities $1/n^2$. In the Poisson setting for time $t = n$, since we have $\mathbb{E}K(n) = \sum_j (1 - \mathrm{e}^{-np_j}) = n^2(1 - \mathrm{e}^{-1/n})$, using an upper and lower Taylor expansion we can obtain the bounds:

$$n - \frac{1}{2} \le \mathbb{E}K(n) \le n - \frac{1}{2} + \frac{1}{6n}.$$

Since $\mathrm{Var}(K(n)) = \sum_j \mathrm{e}^{-np_j}(1-\mathrm{e}^{-np_j}) = \mathbb{E}K(2n) - \mathbb{E}K(n)$, it follows that:

$$n - \frac{1}{6n} \leq \mathrm{Var}\bigl(K(n)\bigr) \leq n + \frac{1}{12n}.$$

Meanwhile, we have $\mathbb{E}[K_1(n)] = \sum_j np_j \mathrm{e}^{-np_j} = n\mathrm{e}^{-1/n}$, which can be bounded similarly:

$$n - 1 \leq \mathbb{E}\bigl[K_1(n)\bigr] \leq n - 1 + \frac{1}{2n}.$$

We can thus see that the Poisson birthday paradox satisfies the spirit of Proposition 3.1, if not its letter (because of being outside of our fixed-$p$ setting). Namely, the Poisson quantities $\mathrm{Var}(K(n))$ and $\mathbb{E}K_1(n)$ are of the same order of magnitude, roughly $n$.

On the other hand, in the binomial setting, since $1 - M_{n,0} = K_n/n^2$,

$$\mathrm{Var}(K_n) \leq \mathbb{E}\bigl[K_{n,1}(1-M_{n,0})\bigr] = \mathbb{E}\biggl[K_{n,1}\frac{K_n}{n^2}\biggr] \leq \frac{1}{n}\mathbb{E}K_{n,1} \leq 1,$$

where we have used the same variance bound as in the proof of Proposition 3.1 (Section 7.2.1). While this implies that the upper bound $\mathrm{Var}(K_n) \leq \mathbb{E}K_{n,1}$ is satisfied, it also shows that a lower bound of the kind of $\mathbb{E}K_{cn,1}/c \leq \mathrm{Var}(K_n)$ is not possible, since similarly to $\mathbb{E}K_1(n)$, we have $\mathbb{E}K_{n,1} = \sum_j np_j(1-p_j)^n = n(1-\frac{1}{n^2})^n \geq n-1$.

Another straightforward bound on $\mathrm{Var}(K_n)$ comes from the fact that the Bernoulli variables $(\mathbb{I}_{\{X_{n,j}>0\}})_{j\geq 1}$ are negatively correlated. Thus, ignoring the covariance terms, we get

$$\mathrm{Var}(K_n) \leq \sum_{j=1}^{\infty} \mathrm{Var}(\mathbb{I}_{\{X_{n,j}>0\}}) = \sum_{j=1}^{\infty}(1-p_j)^n\bigl(1-(1-p_j)^n\bigr) = \mathbb{E}K_{2n} - \mathbb{E}K_n.$$

Let us denote this bound by $\mathrm{Var}^{\mathrm{ind}}(K_n) = \mathbb{E}K_{2n} - \mathbb{E}K_n$, as it is a variance proxy obtained by considering that the summands in $K_n$ are independent. One can observe that the expression of $\mathrm{Var}^{\mathrm{ind}}(K_n)$ is very similar to the variance in the Poissonized setting, $\mathrm{Var}(K(t)) = \mathbb{E}K(2t) - \mathbb{E}K(t)$. It is insightful to compare the true variance, the Poissonized proxy, and the negative correlation proxy, to quantify the price one pays by resorting to the latter two as an approximation for the first. We revisit this in more detail in Section 6.1.

We now investigate the fluctuations of the individual occupancy counts $K_{n,r}$ and $K_r(t)$, and that of the cumulative occupancy counts $K_{n,\overline{r}} = \sum_{j\geq r} K_{n,j}$ and $K_{\overline{r}}(t) = \sum_{j\geq r} K_j(t)$.

**Proposition 3.2.** *In the Poisson setting, for $r \geq 1$, $t \geq 0$,*

$$\mathrm{Var}\bigl(K_{\overline{r}}(t)\bigr) \leq \min\bigl(r\mathbb{E}K_r(t), \mathbb{E}K_{\overline{r}}(t)\bigr).$$

*In the binomial setting, for $r, n \geq 1$,*

$$\mathrm{Var}(K_{n,\overline{r}}) \leq \min(r\mathbb{E}K_{n,r}, \mathbb{E}K_{n,\overline{r}}).$$

For each setting, the first bound follows from Efron–Stein–Steele inequalities, the second from negative correlation. These techniques are presented briefly in Sections 7.1.1 and 7.1.2, respectively.

***Remark 3.1.*** Except for $r = 1$, there is no clear-cut answer as to which of these two bounds is the tightest. In the regular variation scenario with index $\alpha \in ]0, 1]$ as explored in Gnedin *et al.* [22], the two bounds are asymptotically of the same order, indeed,

$$\frac{r\mathbb{E}K_{n,r}}{\mathbb{E}K_{n,\overline{r}}} \underset{n\to+\infty}{\sim} \alpha,$$

see Section 4 for more on this.

Bounds on $\mathrm{Var}(K_r(t))$ can be easily derived as $K_r(t)$ is a sum of independent Bernoulli random variables, however, noticing that $K_{n,r} = K_{n,\overline{r}} - K_{n,\overline{r+1}}$ and that $K_{n,\overline{r}}$ and $K_{n,\overline{r+1}}$ are positively correlated, the following bound is immediate.

**Proposition 3.3.** *In the Poisson setting, for $r \geq 1$, $t \geq 0$,*

$$\mathrm{Var}\big(K_r(t)\big) \leq \mathbb{E}K_r(t).$$

*In the binomial setting, for $r, n \geq 1$,*

$$\begin{aligned}\mathrm{Var}(K_{n,r}) &\leq \min\big(r\mathbb{E}K_{n,r} + (r+1)\mathbb{E}K_{n,r+1}, \mathbb{E}K_{n,\overline{r}} + \mathbb{E}K_{n,\overline{r+1}}\big)\\ &\leq 2\min\big(\max\big(r\mathbb{E}K_{n,r}, (r+1)\mathbb{E}K_{n,r+1}\big), \mathbb{E}K_{n,\overline{r}}\big).\end{aligned}$$

3.1.2. *Concentration inequalities*

Concentration inequalities refine variance bounds. These bounds on the logarithmic moment generating functions are indeed Bennett (sub-Poisson) inequalities with the variance upper bounds stated in the preceding section. For $K_{n,\overline{r}}$, the next proposition gives a Bernstein inequality where the variance factor is, up to a constant factor, the Efron–Stein upper bound on the variance.

**Proposition 3.4.** *Let $r \geq 1$, and let $v_{n,\overline{r}} = \min(r\mathbb{E}K_{n,r}, \mathbb{E}K_{n,\overline{r}})$. Then, for all $\lambda \in \mathbb{R}$,*

$$\log \mathbb{E}\mathrm{e}^{\lambda(K_{n,\overline{r}} - \mathbb{E}K_{n,\overline{r}})} \leq v_{n,\overline{r}}\phi(\lambda),$$

*with $\phi : \lambda \mapsto \mathrm{e}^{\lambda} - \lambda - 1$.*

It is worth noting that the variance bound $\mathbb{E}K_{n,\overline{r}}$ in this concentration inequality can also be obtained using a variant of Stein's method known as size-biased coupling (Bartroff *et al.* [5], Chen *et al.* [15]).

A critical element of the proof of Proposition 3.4 is to use the fact that each $K_{n,\overline{r}}$ is a sum of negatively associated random variables (Section 7.1.2). This is not the case for $K_{n,r}$, and thus negative association cannot be invoked directly. To deal with this, we simply use the observation of Ohannessian and Dahleh [35] that since $K_{n,r} = K_{n,\overline{r}} - K_{n,\overline{r+1}}$, the concentration of $K_{n,r}$ follows from that of those two terms. We can show the following.

**Proposition 3.5.** *Let*

$$v_{n,r} = 2\min\bigl(\max\bigl(r\mathbb{E}K_{n,r}, (r+1)\mathbb{E}K_{n,r+1}\bigr), \mathbb{E}K_{n,\bar{r}}\bigr).$$

*Then, for* $s \geq 0$,

$$\mathbb{P}\bigl\{|K_{n,r} - \mathbb{E}K_{n,r}| \geq \sqrt{4v_{n,r}s} + 2s/3\bigr\} \leq 4\mathrm{e}^{-s}.$$

## 3.2. Missing mass

### 3.2.1. *Variance bound*

Recall that $M_{n,0} = \sum_{j=1}^{\infty} p_j \mathbb{I}_{\{X_{n,j}=0\}} = \sum_{j=1}^{\infty} p_j Y_j$, and we can readily show that the summands are negatively associated weighted Bernoulli random variables (Section 7.1.2). This results in a handy upper bound for the variance of the missing mass.

**Proposition 3.6.** *In the Poisson setting*,

$$\mathrm{Var}\bigl(M_0(t)\bigr) = 2\mathbb{E}K_2(t)/t^2 - \mathbb{E}K_2(2t)/2t^2 \leq 2\mathbb{E}K_2(t)/t^2,$$

*while in the binomial setting*,

$$\mathrm{Var}(M_{n,0}) \leq \sum_{j=1}^{\infty} p_j^2 \,\mathrm{Var}(Y_j) \leq \frac{2\mathbb{E}K_2(n)}{n^2}.$$

Note that whereas the expected value of the missing mass is connected to the number of singletons, its variance may be upper bounded using the number of *doubletons* (in the Poisson setting). This connection was already pointed out in Good [24] and Esty [20].

### 3.2.2. *Concentration of the left tail*

Moving on to the concentration properties of the missing mass, we first note that as a sum of weighted sub-Poisson random variables (following Boucheron *et al.* [13]), the missing mass is itself a sub-gamma random variable on both tails. It should not come as a surprise that the left tail of $M_{n,0}$ is sub-Gaussian with the variance factor derived from negative association. This had already been pointed out by McAllester and Schapire [33] and McAllester and Ortiz [32].

**Proposition 3.7 (McAllester and Ortiz [32]).** *In the Poisson setting, the missing mass* $M_0(t)$ *is sub-Gaussian on the left tail with variance factor the effective variance* $\mathrm{Var}(M_0(t)) = \sum_{j=1}^{\infty} p_j^2 \mathrm{e}^{-tp_j}(1-\mathrm{e}^{-tp_j})$.

*In the binomial setting, the missing mass* $M_{n,0}$ *is sub-Gaussian on the left tail with variance factor* $v = \sum_{j=1}^{\infty} p_j^2 \,\mathrm{Var}(Y_j)$ *or* $v_n^- = 2\mathbb{E}K_2(n)/n^2$.

*For* $\lambda \leq 0$,

$$\log \mathbb{E}\bigl[\mathrm{e}^{\lambda(M_{n,0} - \mathbb{E}M_{n,0})}\bigr] \leq \frac{v\lambda^2}{2} \leq \frac{2v_n^-\lambda^2}{2}.$$

### 3.2.3. *Concentration of the right tail*

The following concentration inequalities for the right tail of the missing mass mostly rely on the following proposition, which bounds the log-Laplace transform of the missing mass in terms of the sequence of expected occupancy counts $\mathbb{E}K_r(n)$, for $r \geq 2$.

**Proposition 3.8.** *For all* $\lambda > 0$,

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \leq \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \mathbb{E}K_r(n).$$

This suggests that if we have a uniform control on the expected occupancy scores $(\mathbb{E}K_r(t))_{r\geq 2}$, then the missing mass has a sub-gamma right tail, with some more or less accurate variance proxy, and scale factor $1/n$.

The next theorem shows that the missing mass is sub-gamma on the right tail with variance proxy $2\mathbb{E}K_{\overline{2}}(n)/n^2$ and scale proxy $1/n$. Despite its simplicity and its generality, this bound exhibits an intuitively correct scale factor: if there exist symbols with probability of order $1/n$, they offer the major contribution to the fluctuations of the missing mass.

**Theorem 3.9.** *In the binomial as well as in the Poisson setting, the missing mass is sub-gamma on the right tail with variance factor* $v_n^+ = 2\mathbb{E}K_{\overline{2}}(n)/n^2$ *and scale factor* $1/n$. *For* $\lambda \geq 0$,

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \leq \frac{v_n^+\lambda^2}{2(1-\lambda/n)}.$$

*If the sequence* $(\mathbb{E}K_r(n))_{r\geq 2}$ *is non-increasing, the missing mass is sub-gamma on the right tail with variance factor* $v_n^- = 2\mathbb{E}K_2(n)/n^2$ *and scale factor* $1/n$,

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \leq \frac{v_n^-\lambda^2}{2(1-\lambda/n)}.$$

***Remark 3.2.*** McAllester and Ortiz [32] and Berend and Kontorovich [8] point out that for each Bernoulli random variable $Y_j$, for all $\lambda \in \mathbb{R}$

$$\log \mathbb{E}\mathrm{e}^{\lambda(Y_j-\mathbb{E}Y_j)} \leq \frac{\lambda^2}{4\mathrm{c}_{\mathrm{LS}}(\mathbb{E}Y_j)},$$

where $\mathrm{c}_{\mathrm{LS}}(p) = \log((1-p)/p)/(1-2p)$ (or 2 if $p = 1/2$) is the optimal logarithmic Sobolev constant for Bernoulli random variables with success probability $p$ (this sharp and nontrivial result has been proven independently by a number of people: Kearns and Saul [30], Berend and Kontorovich [8], Raginsky and Sason [38], Berend and Kontorovich [9]; the constant also appears early on in the exponent of one of Hoeffding's inequalities (Hoeffding [27], Theorem 1, equation (2.2)). From this observation, thanks to the negative association of the $(Y_j)_{j\geq 1}$, it follows that the missing mass is sub-Gaussian with variance factor

$$w_n = \sum_{j=1}^{\infty} \frac{p_j^2}{2\mathrm{c}_{\mathrm{LS}}((1-p_j)^n)} \leq \sum_{j=1}^{\infty} \frac{p_j^2}{2\log((1-p_j)^{-n})} \leq \sum_{j=1}^{\infty} \frac{p_j^2}{2np_j} \leq \frac{1}{2n}. \tag{3.1}$$

An upper bound on $w_n$ does not mean that $w_n$ is necessarily larger than $\mathbb{E}K_{\overline{2}}(n)/n^2$. Nevertheless, it is possible to derive a simple lower bound on $w_n$ that proves to be of order $O(1/n)$.

Assume that the sequence $(p_j)_{j\geq 1}$ is such that $p_j \leq 1/4$ for all $j \geq 1$. Then

$$\begin{aligned}
w_n &\geq \sum_{j:p_j\geq 1/n} \frac{p_j^2}{2\mathrm{c}_{\mathrm{LS}}((1-p_j)^n)} \\
&\geq \sum_{j:p_j\geq 1/n} \frac{p_j^2(1-2(1-p_j)^n)}{2n\log(1/(1-p_j))} \\
&\geq \sum_{j:p_j\geq 1/n} \frac{p_j^2(1-2/\mathrm{e})}{2np_j/(1-p_j)} \\
&\geq \frac{3(1-2/\mathrm{e})}{8n}\left(1-\sum_{j:p_j<1/n} p_j\right) \\
&\geq \frac{3}{32n}\left(1-\sum_{j:p_j<1/n} p_j\right),
\end{aligned}$$

and the statement follows from the observation that $\lim_{n\to\infty}\sum_{j:p_j<1/n} p_j = 0$.

The variance factor $w_n$ from (3.1) is usually larger than $2\mathbb{E}K_{\overline{2}}(n)/n^2$. In the scenarios discussed in Section 4, $(2\mathbb{E}K_{\overline{2}}(n)/n^2)/w_n$ even tends to 0 as $n$ tends to infinity.

# 4. Regular variation

*Motivation*. Are the variance bounds in the results of Section 3 tight? In some pathological situations, this may not be the case. Consider the following example (which revisits Example 1).

***Example 2.*** We may challenge the tail bounds offered by Proposition 3.7 and Theorem 3.9 in the simplest setting where we have $k$ symbols all of which have equal probabilities $1/k$. Then the missing mass is $1-K_n/k$, its variance is $\mathrm{Var}(K_n)/k^2$. In the birthday paradox setting ($k=n^2$), $\mathrm{Var}(M_{n,0}) \leq 1/n^4$, and the variance bound $2\mathbb{E}K_{\overline{2}}(n)/n^2$ is not tight. Indeed, one can verify that $\mathbb{E}K_{\overline{2}}(n) \geq \frac{1}{2}(1-\frac{1}{n})$ so that $v_n^+ \geq \frac{1}{n^2}-\frac{1}{n^3}$. However, in what is called the *central domain* in Kolchin *et al.* [31], that is when $k\to\infty$ while $n/k\to t\in\mathbb{R}_+$, the tail bounds become relevant. The variance of $K_n$ is equivalent to $k\mathrm{e}^{-t}(1-\mathrm{e}^{-t})$ while its expectation is equivalent to $k(1-\mathrm{e}^{-t})$.

Note that in this setting all $\mathbb{E}K_r(n)$ and $\mathbb{E}K_{n,r}$ are of the same order of magnitude as $\mathbb{E}K_n$, indeed $\mathbb{E}K_r(n)/\mathbb{E}K(n) \to \mathrm{e}^{-t}t^r/(r!(1-\mathrm{e}^{-t}))$.

These examples are illustrative although they do not fall in the fixed $p$ regime we are considering in this paper. We use them because they have tractable expressions, and they provide informative diagnostics. To parallel the phenomenon of mismatched variance proxies in our setting, one can simply look at the geometric distribution for a concrete example. If $(p_k)_{k\geq 1}$ defines a geometric distribution $p_k = (1-q)^{k-1}q$, then $\mathbb{E}K_2(n)$ remains bounded, while $\mathbb{E}K_{\overline{2}}(n)$ scales like $\log n$ as $n$ tends to infinity.

In particular, we may conjecture that Theorem 3.9 is likely to be sharp when the first terms of the sequence $(\mathbb{E}K_r(n))_{r\geq 2}$ grow at the same rate as $\mathbb{E}K(n)$, or at least as $\mathbb{E}K_{\overline{2}}(n)$, which is not the case in the birthday paradox setting of Example 2. We see in what follows that the regular variation framework introduced by Karlin [29] leads to such asymptotic equivalents. The most useful aspect of these equivalent growth rates is a simple characterization of the variance of various quantities, particularly relative to their expectation. We focus on the right tail of the missing mass, which exhibits the highest sensitivity to this asymptotic behavior, by trying to specialize Theorem 3.9 under regular variation.

*Definition*. Regularly varying frequencies can be seen as generalizations of power-law frequencies. One possible definition is as follows: for $\alpha \in (0, 1)$, the sequence $(p_j)_{j\geq 1}$ is said to be regularly varying with index $-1/\alpha$ if, for all $\kappa \in \mathbb{N}_+$,

$$\frac{p_{\kappa j}}{p_j} \underset{j\to\infty}{\sim} \kappa^{-1/\alpha}.$$

It is easy to see that pure power laws do indeed satisfy this definition. However, in order to extend the regular variation hypothesis to $\alpha = 0$ and 1, we need a more flexible definition, which requires some new notation. This definition relies on the *counting function* $\vec{\nu}$, defined for all $x > 0$ by:

$$\vec{\nu}(x) = \big|\{j : j \geq 1, p_j \geq x\}\big|.$$

The overhead arrow is not a vector notation, and rather codifies that we are counting points with probability "to the right" of $x$. More precisely, letting $\nu$ be the *counting measure* defined by

$$\nu(\mathrm{d}x) = \sum_{j=1}^{\infty} \delta_{p_j}(\mathrm{d}x),$$

then, for all $x > 0$, $\vec{\nu}(x) = \nu[x, 1]$.

Henceforth, following Karlin [29], we say that the probability mass function $(p_j)_j$ is regularly varying with index $\alpha \in [0, 1]$, if $\vec{\nu}(1/\cdot)$ is $\alpha$-regularly varying in the neighbourhood of $\infty$, which reads as

$$\vec{\nu}(1/x) \underset{x\to\infty}{\sim} x^{\alpha}\ell(x),$$

where $\ell$ is a slowly varying function, that is, for all $x > 0$, $\lim_{\tau\to+\infty} \ell(\tau x)/\ell(\tau) = 1$. We use the notation $\vec{\nu}(1/\cdot) \in \mathrm{RV}_\alpha$ to indicate that $\vec{\nu}(1/\cdot)$ is $\alpha$-regularly varying.

We now note that when $\alpha \in (0,1)$, the regular variation assumption on $(p_j)_{j\geq 1}$ is indeed equivalent to the regular variation assumption on the counting function $\vec{\nu}$ (see Gnedin et al. [22], Proposition 23): if $(p_j)_{j\geq 1}$ is regularly varying with index $-1/\alpha$ as $j$ tends to infinity, then $\vec{\nu}(1/\cdot)$ is $\alpha$-regularly varying, that is $\lim_{x\to\infty} \vec{\nu}(1/(qx))/\vec{\nu}(1/x) = q^\alpha$ for $q > 0$ (see also Bingham *et al.* [11]). The second definition however lends itself more easily to generalization to $\alpha = 0$ and 1.

In what follows, we treat these three cases separately: the nominal regular variation case with $\alpha \in (0,1)$ strictly, the *fast variation* case with $\alpha = 1$, and the *slow variation* case with $\alpha = 0$.

In the latter case, that is if frequencies $p_j$ are regularly varying with index 0, we find that the mere regular variation hypothesis is not sufficient to obtain asymptotic formulas. For this reason, we introduce further control in the form of an *extended* regular variation hypothesis (given by Definition 1 of Section 4.3).

***Remark 4.1.*** Before we proceed, as further motivation, we note that the regular variation hypothesis is very close to being a necessary condition for exponential concentration. For example, considering Proposition 3.8, we see that if the sampling distribution is such that the ratio $\mathbb{E}K_{\overline{2}}(t)/\mathbb{E}K_2(t)$ remains bounded, then we are able to capture the right variance factor. Now, defining the shorthand $\Phi_{\overline{2}}(t) = \mathbb{E}K_{\overline{2}}(t)$ and $\Phi_2(t) = \mathbb{E}K_2(t)$ following the notation of Gnedin *et al.* [22], we have

$$\Phi'_{\overline{2}}(t) = \frac{2\Phi_2(t)}{t}.$$

Hence, $\Phi_{\overline{2}}(t)/\Phi_2(t) = 2\Phi_{\overline{2}}(t)/t\Phi'_{\overline{2}}(t)$, and if instead of boundedness, we further require that this ratio converges to some finite limit, then, by the converse part of Karamata's theorem (see de Haan and Ferreira [16], Theorem B.1.5), we find that $\Phi_{\overline{2}}$ (and then $\Phi_2$) is regularly varying, which in turn implies that $\vec{\nu}(1/t)$ is regularly varying. We elaborate on this further in our discussions, in Section 6.2.

## 4.1. Case $\alpha \in (0,1)$

We first consider the case $0 < \alpha < 1$. The next theorem states that when the sampling distribution is regularly varying with index $\alpha \in (0,1)$, the variance factors in the Bernstein inequalities of Proposition 3.7 and Theorem 3.9 are of the same order as the variance of the missing mass.

**Theorem 4.1.** *Assume that the counting function $\vec{\nu}$ satisfies the regular variation condition with index $\alpha \in (0,1)$, then the missing mass $M_{n,0}$ (or $M_0(n)$) is sub-Gaussian on the left tail with variance factor $v_n^- = 2\mathbb{E}K_2(n)/n^2$ and sub-gamma on the right tail with variance factor $v_n^+ = 2\mathbb{E}K_{\overline{2}}(n)/n^2$. The variance factors satisfy*

$$\lim_n \frac{v_n^-}{\operatorname{Var}(M_{n,0})} = \frac{1}{1-2^{\alpha-2}},$$

$$\lim_n \frac{v_n^+}{\operatorname{Var}(M_{n,0})} = \frac{2}{\alpha(1-2^{\alpha-2})},$$

*and thus*

$$\lim_n \frac{v_n^-}{v_n^+} = \frac{\alpha}{2}.$$

The second ratio deteriorates when $\alpha$ approaches 0, implying that the variance factor for the right tail gets worse for lighter tails. We do not detail the proof of Theorem 4.1, except to note that it follows from Proposition 3.7, Theorem 3.9, and the following asymptotics (see also Gnedin *et al.* [22], Ohannessian and Dahleh [35]).

**Theorem 4.2 (Karlin [29]).** *If the counting function $\vec{\nu}$ is regularly varying with index $\alpha \in (0, 1)$, for all $r \geq 1$,*

- $K_n \overset{a.s.}{\sim} \mathbb{E}K_n \sim_{+\infty} \Gamma(1-\alpha)n^{\alpha}\ell(n)$,
- $K_{n,r} \overset{a.s.}{\sim} \mathbb{E}K_{n,r} \sim_{+\infty} \frac{\alpha\Gamma(r-\alpha)}{r!}n^{\alpha}\ell(n)$,
- $\mathrm{Var}(M_{n,0}) \sim \alpha\Gamma(2-\alpha)(1-2^{\alpha-2})n^{\alpha-2}\ell(n)$

*and the same hold for the corresponding Poissonized quantities.*

Note that all expected occupancy counts are of the same order, and the asymptotics for $\mathbb{E}K_{\overline{2}}(n)$ follows directly from the difference between $\mathbb{E}K(n)$ and $\mathbb{E}K_1(n)$.

## 4.2. Fast variation, $\alpha = 1$

We refer to the regular variation regime with $\alpha = 1$ as *fast variation.*[1] From the perspective of concentration, this represents a relatively "easy" scenario. In a nutshell, this is because the variance of various quantities grows much slower than their expectation.

The result of this section is to simply state that Theorem 4.1 continues to hold as is for $\alpha = 1$. The justification for this, however, is different. In particular, the asymptotics of Theorem 4.2 do not apply: the number of distinct symbols $K_n$ and the singletons $K_{n,1}$ continue to have comparable growth order, but now their growth dominates that of $K_{n,r}$ for all $r \geq 2$. Intuitively, under fast variation almost all symbols appear only once in the observation, with only a vanishing fraction of symbols appearing more than once. We formalize this in the following theorem.

**Theorem 4.3 (Karlin [29]).** *Assume $\vec{\nu}(1/x) = x\ell(x)$ with $\ell \in \mathrm{RV}_0$ (note that $\ell$ tends to 0 at $\infty$). Define $\ell_1 : [1,\infty) \to \mathbb{R}_+$ by $\ell_1(y) = \int_y^{\infty} u^{-1}\ell(u)\,\mathrm{d}u$. Then $\ell_1 \in \mathrm{RV}_0$ and $\lim_{t\to\infty} \ell_1(t)/\ell(t) = \infty$ and the following asymptotics hold*:

- $K_n \overset{a.s.}{\sim} \mathbb{E}K_n \sim_{+\infty} n\ell_1(n)$,
- $K_{n,1} \overset{a.s.}{\sim} \mathbb{E}K_{n,1} \sim_{+\infty} \mathbb{E}K_n$,
- $K_{n,r} \overset{a.s.}{\sim} \mathbb{E}K_{n,r} \sim_{+\infty} \frac{1}{r(r-1)}n\ell(n)$, $r \geq 2$,

*and the same hold for the corresponding Poissonized quantities.*

[1] Sometimes *rapid variation* is used Gnedin *et al.* [22], but this conflicts with Bingham *et al.* [11].

As the expected missing mass scales like $\mathbb{E}K_1(n)/n$ while its variance scales like $\mathbb{E}K_2(n)/n^2$, Theorem 4.3 quantifies our claim that this is an “easy” concentration. To establish Theorem 4.1, it remains to show that $\mathbb{E}K_{\overline{2}}(n)$ is also of the same order as $\mathbb{E}K_2(n)$, with the correct limiting ratio for $\alpha = 1$. For this, we give the following proposition, which is in fact sufficient to prove Theorem 4.1 for both $0 < \alpha < 1$ and $\alpha = 1$.

**Proposition 4.4.** *Assume that the counting function $\vec{\nu}$ satisfies the regular variation condition with index $\alpha \in (0, 1]$, then for all $r \geq 2$,*

$$K_{\overline{r}}(n) \underset{+\infty}{\sim} \frac{\Gamma(r-\alpha)}{(r-1)!}\vec{\nu}(1/n) \qquad \textit{almost surely.}$$

Thus, when $\alpha = 1$, $\mathbb{E}K_r(n)$ and $\mathbb{E}K_{\overline{r}}(n)$ for $r \geq 2$ all grow like $n\ell(n)$, which is dominated by the $n\ell_1(n)$ growth of $\mathbb{E}K(n)$ and $\mathbb{E}K_1(n)$, as $\ell(n)/\ell_1(n) \to 0$. Specializing for $r = 2$, we do find that our proxies still capture the right order of the variance of the missing mass, and that we have the desired limit of Theorem 4.1, $\lim_n v_n^-/v_n^+ = \frac{1}{2}$.

***Remark 4.2.*** When $0 < \alpha < 1$, another good variance proxy would have been $2\mathbb{E}K(n)/n^2$. For $\alpha = 1$, however, singletons should be removed to get the correct order.

We also note that when $\alpha = 1$, the missing mass is even more stable. If we let $v_n$ denote either $2\mathbb{E}K_2(n)/n^2$ or $2\mathbb{E}K_{\overline{2}}(n)/n^2$, then we have the following comparison between the expectation and the fluctuations of the missing mass, with the appropriate constants:

$$\frac{\sqrt{v_n}}{\mathbb{E}M_{n,0}} \sim \begin{cases} c_\alpha \cdot n^{-\alpha/2}\dfrac{\sqrt{\ell(n)}}{\ell(n)}, & \text{for } 0 < \alpha < 1\,, \\ c_1 \cdot n^{-1/2}\dfrac{\sqrt{\ell(n)}}{\ell_1(n)}, & \text{for } \alpha = 1. \end{cases}$$

## 4.3. Slow variation, $\alpha = 0$

The setting where the counting function $\vec{\nu}$ satisfies the regular variation condition with index 0 represents a challenge. We refer to this regime simply as *slow variation*. Recall that this means that $\vec{\nu}(z/n)/\vec{\nu}(1/n)$ converges to 1 as $n$ goes to infinity, yet to deal with this case we need to control the speed of this convergence, exemplified by the notion of extended regular variation that was introduced by de Haan (see Bingham *et al.* [11], de Haan and Ferreira [16]). As we illustrate in the end of this section, one may face rather irregular behavior without such a hypothesis.

***Definition 1.*** *A measurable function $\ell : \mathbb{R}^+ \to \mathbb{R}^+$ has the extended slow variation property, if there exists a nonnegative measurable function $a : \mathbb{R}^+ \to \mathbb{R}^+$ such that for all $x > 0$*

$$\lim_{\tau\to\infty} \frac{\ell(\tau x) - \ell(\tau)}{a(\tau)} \underset{\tau\to\infty}{\to} \log(x).$$

*The function $a(\cdot)$ is called an auxiliary function. When a function $\ell$ has the extended slow variation property with auxiliary function $a$, we denote it by $\ell \in \Pi_a$.*

Note that the auxiliary function is always slowly varying and grows slower than the original function, namely it satisfies $\lim_{\tau\to\infty} \ell(\tau)/a(\tau) = \infty$. Furthermore, any two possible auxiliary functions are asymptotically equivalent, that is if $a_1$ and $a_2$ are both auxiliary functions for $\ell$, then $\lim_{t\to\infty} a_1(t)/a_2(t) = 1$.

The notion of extended slow variation and the auxiliary function give us the aforementioned control needed to treat the $\alpha = 0$ case on the same footing as the $0 < \alpha < 1$ case. In particular, in what follows in this section we assume that $\vec{\nu}(1/\cdot) \in \Pi_a$, with the additional requirement that the auxiliary function $a$ tends to $+\infty$.

***Remark 4.3.*** This domain corresponds to light-tailed distributions just above the geometric distribution (the upper-exponential part of Gumbel's domain). For the geometric distribution with frequencies $p_j = (1-q)q^{k-1}$, $j = 1, 2, \dots$, the counting function satisfies $\vec{\nu}(1/n) \sim_\infty \log_{1/q}(n) \in \mathrm{RV}_0$, but the auxiliary function $a(n) = \log(1/q)$ does not tend to infinity. Frequencies of the form $p_j = cq^{\sqrt{j}}$ on the other hand do fit this framework.

**Theorem 4.5 (Gnedin *et al.* [22]).** *Assume that $\ell(t) = \vec{\nu}(1/t)$ is in $\Pi_a$, where $a$ is slowly varying and tends to infinity. The following asymptotics hold for each $r \geq n$:*

- $K_n \overset{\mathbb{P}}{\sim} \mathbb{E}K_n \sim_{+\infty} \ell(n)$,
- $K_{n,r} \overset{\mathbb{P}}{\sim} \mathbb{E}K_{n,r} \sim_{+\infty} \frac{a(n)}{r}$,
- $K_{n,\bar{r}} \overset{\mathbb{P}}{\sim} \mathbb{E}K_{n,\bar{r}} \sim_{+\infty} \ell(n)$,
- $M_{n,r} \overset{\mathbb{P}}{\sim} \mathbb{E}M_{n,r} \sim_{+\infty} \frac{a(n)}{n}$.

*The same equivalents hold for the corresponding Poissonized quantities.*

***Remark 4.4.*** In this case, the expectations $(\mathbb{E}K_{n,r})_{r\geq 1}$ are of the same order but are much smaller than $\mathbb{E}K_n$, and the variables $K_n$ and $K_{n,\bar{s}}$ are all almost surely equivalent to $\ell(n)$. It is also remarkable that all the expected masses $(\mathbb{E}M_{n,r})_{r\geq 1}$ are equivalent.

The variance of the missing mass is of order $2\mathbb{E}K_2(n)/n^2 \sim a(n)/n^2$, whereas the proxy $2\mathbb{E}K_{\bar{2}}(n)/n^2$ is of much faster order $2\ell(n)/n^2$, and is thus inadequate. By exploiting more carefully the regular variation hypothesis, we obtain uniform control over $(\mathbb{E}K_r(n))_{r\geq 1}$ for large enough $n$, leading to a variance proxy of the correct order.

**Theorem 4.6.** *Assume that $\ell$ defined by $\ell(x) = \vec{\nu}(1/x)$ is in $\Pi_a$ where the slowly varying function $a$ tends to infinity, and let $v_n = 12a(n)/n^2$. We have*:

1. $\mathrm{Var}(M_{n,0}) \sim \frac{3a(n)}{4n^2}$, *thus* $v_n \asymp \mathrm{Var}(M_{n,0})$.
2. *There exists $n_0 \in \mathbb{N}$ that depends on $\vec{\nu}$ such that for all $n > n_0$, for all $\lambda > 0$,*

$$\log \mathbb{E}\left[e^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \leq \frac{v_n \lambda^2}{2(1-\lambda/n)}.$$

*The same results hold for $M_0(t)$.*

***Remark 4.5.*** By standard Chernoff bounding, Theorem 4.6 implies that there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $s \geq 0$,

$$\mathbb{P}\left\{M_{n,0} \geq \mathbb{E}M_{n,0} + \sqrt{2v_n s} + \frac{s}{n}\right\} \leq \mathrm{e}^{-s}.$$

4.3.1. *Too slow variation*

We conclude this section by motivating why it is crucial to have a heavy-enough tail in order to obtain meaningful concentration. For example, even under regular variation when $\alpha = 0$, but $\vec{\nu}$ is not in a de Haan class $\Pi_a$ with $a(n) \to \infty$, the behavior of the occupancy counts and their moments may be quite irregular. In this section, we collect some observations on those light-tailed distributions. We start with the geometric distribution which represents in many respects a borderline case.

The geometric case is an example of slow variation: $\vec{\nu}(1/\cdot) \in \mathrm{RV}_0$. Indeed, with $p_k = (1-q)^{k-1}q$, $0 < q < 1$, we have

$$\begin{aligned}
\vec{\nu}(x) &= \sum_{k=1}^{+\infty} \mathbb{I}_{\{p_k \geq x\}} \\
&= \left|k \in \mathbb{N}_+, (1-q)^{k-1}q \geq x\right| \\
&= 1 + \left\lfloor \frac{\log(x/q)}{\log(1-q)} \right\rfloor,
\end{aligned}$$

and thus $\vec{\nu}(x) \sim_{x \to 0} \ell(1/x)$, with $\ell$ slowly varying.

In this case, $\mathrm{Var}(K(n)) = \mathbb{E}K(2n) - \mathbb{E}K(n) \to \frac{\log(2)}{\log(1/1-q)}$.

**Proposition 4.7.** *When the sampling distribution is geometric with parameter $q \in (0,1)$, letting $M_n = \max(X_1, \dots, X_n)$,*

$$\mathbb{E}M_n \geq \mathbb{E}K_n \geq \mathbb{E}M_n - \frac{1-q}{q^2}.$$

In the case of geometric frequencies, the missing mass can fluctuate widely with respect to its expectation, and one cannot expect to obtain sub-gamma concentration with both the correct variance proxy and scale factor $1/n$. Indeed, intuitively, the symbol which primarily contributes to the missing mass' fluctuations, is the quantile of order $1 - 1/n$. With $F(k) = \sum_{j=1}^{k} p_j$, and $F^{\leftarrow}$ the generalized inverse of $F$,

$$\begin{aligned}
j^* = F^{\leftarrow}(1 - 1/n) &= \inf\{j \geq 1, F(j) \geq 1 - 1/n\} \\
&= \inf\left\{j \geq 1, \sum_{k>j} p_k \leq 1/n\right\}.
\end{aligned}$$

Omitting the slowly varying functions, when $\vec{\nu}(1/\cdot) \in \mathrm{RV}_\alpha$, $0 < \alpha < 1$, $j^*$ is of order $n^{\alpha/(1-\alpha)}$ and $p_{j^*}$ is of order $n^{-1/(1-\alpha)}$. The closer to 1 is $\alpha$, the smaller the probability of $j^*$. When $\alpha$

goes to 0, this probability becomes $1/n$. With geometric frequencies, $j^*$ is $\frac{\log(n)}{\log(1/1-q)}$ and $p_{j^*}$ is $\frac{q}{n(1-q)}$. Hence, around the quantile of order $1 - 1/n$, there are symbols which may contribute significantly to the missing mass' fluctuations.

Another interesting case consists of distributions which are very light-tailed, in the sense that $\frac{p_{k+1}}{p_k} \to 0$ when $k \to \infty$. An example of these is the Poisson distribution $\mathcal{P}(\lambda)$, for which $\frac{p_{k+1}}{p_k} = \frac{\lambda}{k} \to_{k\to+\infty} 0$. The next proposition shows that for such concentrated distributions, the missing mass essentially concentrates on two points.

**Proposition 4.8.** *In the infinite urns scheme with probability mass function* $(p_k)_{k\in\mathbb{N}}$, *if* $p_k > 0$ *for all* $k$ *and* $\lim_{k\to\infty} \frac{p_{k+1}}{p_k} = 0$, *then there exists a sequence of integers* $(u_n)_{n\in\mathbb{N}}$ *such that*

$$\lim_{n\to\infty} \mathbb{P}\big\{M_{n,0} \in \big\{\overline{F}(u_n), \overline{F}(u_n+1)\big\}\big\} = 1,$$

*where* $\overline{F}(k) = \sum_{j>k} p_j$.

# 5. Applications

## 5.1. Estimating the regular variation index

When working in the regular variation setting, the most basic estimation task is to estimate the regular variation index $\alpha$. We already mentioned in Section 2 the fact that, when $\vec{\nu} \in \mathrm{RV}_\alpha, \alpha \in (0,1)$, the ratio $K_{n,1}/K_n$ provides a consistent estimate of $\alpha$. This is actually only one among a family of estimators of $\alpha$ that one may construct. The next result shows this, and is a direct consequence of Proposition 4.4.

**Proposition 5.1.** *If* $\vec{\nu} \in \mathrm{RV}_\alpha, \alpha \in (0,1]$, *then for all* $r \geq 1$

$$\frac{rK_{n,r}}{K_{n,\overline{r}}}$$

*is a strongly consistent estimator of* $\alpha$.

Thus, writing $k_n = \max\{r, K_{n,r} > 0\}$, at time $n$, we can have up to $k_n$ non-trivial estimators of $\alpha$. One would expect these estimators to offer various bias-variance trade-offs, and one could ostensibly select an "optimal" $r$ via model selection.

## 5.2. Estimating the missing mass

The Good–Turing estimation problem (Good [24]) is that of estimating $M_{n,r}$ from the observation $(U_1, U_2, \ldots, U_n)$. For large scores $r$, designing estimators for $M_{n,r}$ is straightforward: we assume that the empirical distribution mimics the sampling distribution, and that the empirical probabilities $\frac{rK_{n,r}}{n}$ are likely to be good estimators. The question is more delicate for rare events.

In particular, for $r=0$, it may be a bad idea to assume that there is no missing mass $M_{n,0}=0$, that is to assign a zero probability to the set of symbols that do not appear in the sample. Various "smoothing" techniques have thus developed, in order to adjust the maximum likelihood estimator and obtain more accurate probabilities.

In particular, Good–Turing estimators attempt to estimate $(M_{n,r})_r$ from $(K_{n,r})_r$ for all $r$. They are defined as

$$G_{n,r} = \frac{(r+1)K_{n,r+1}}{n}.$$

The rationale for this choice comes from the following observations.

$$\mathbb{E}G_{n,0} = \frac{\mathbb{E}[K_{n,1}]}{n} = \mathbb{E}M_{n-1,0} = \mathbb{E}M_{n,0} + \frac{\mathbb{E}M_{n,1}}{n} \tag{5.1}$$

and

$$\mathbb{E}G_{n,r} = \frac{(r+1)\mathbb{E}K_{n,r+1}}{n} = \mathbb{E}M_{n-1,r}. \tag{5.2}$$

In the Poisson setting, there is no bias: $\mathbb{E}G_r(t) = (r+1)\frac{\mathbb{E}K_{r+1}(t)}{t} = \mathbb{E}M_r(t)$.

Here, we primarily focus on the estimation of the missing masses $M_{n,0}$ and $M_0(t)$, though most of the methodology extends also to $r>0$, with the appropriate concentration results. From (5.1) and (5.2), Good–Turing estimators look like slightly biased estimators of the relevant masses. In particular, the bias $\mathbb{E}G_{n,0} - \mathbb{E}M_{n,0}$ is always positive but smaller than $1/n$. It is however far from obvious to determine scenarios where these estimators are consistent and where meaningful confidence regions can be constructed.

When trying to estimate the missing mass $M_{n,0}$ or $\mathbb{E}M_{n,0}$, consistency needs to be redefined since the estimand is not a fixed parameter of interest but a random quantity whose expectation further depends on $n$. Additive consistency, that is bounds on $\widehat{M}_{n,0} - M_{n,0}$ is not a satisfactory notion, because, as $M_{n,0}$ tends to 0, the trivial constant estimator 0 would be universally asymptotically consistent. Relative consistency, that is control on $(\widehat{M}_{n,0} - M_{n,0})/M_{n,0}$ looks like a much more reasonable notion. It is however much harder to establish.

In order to establish relative consistency of a missing mass estimator, we have to check that $\mathbb{E}[\widehat{M}_{n,0} - M_{n,0}]$ is not too large with respect to $\mathbb{E}M_{n,0}$, and that both $\widehat{M}_{n,0}$ and $M_{n,0}$ are concentrated around their mean values.

As shown in Ohannessian and Dahleh [35], the Good–Turing estimator of the missing mass is not universally consistent in this sense. This occurs principally in very light tails, such as those described in Section 4.3.1.

**Proposition 5.2 (Ohannessian and Dahleh [35]).** *When the sampling distribution is geometric with small enough $q \in (0,1)$, there exists $\eta > 0$, and a subsequence $n_i$ such that for $i$ large enough, $G_{n_i,0}/M_{n_i,0} = 0$ with probability no less than $\eta$.*

On the other hand, the concentration result of Corollary 4.1 gives a law of large numbers for $M_{n,0}$ (by a direct application of the Borel–Cantelli lemma), which in turn implies the strong multiplicative consistency of the Good–Turing estimate.

**Corollary 5.3.** *We have the following two regimes of consistency for the Good–Turing estimator of the missing mass.*

(i) *If the counting function $\vec{\nu}$ is such that $\mathbb{E}K_{n,2}/\mathbb{E}K_{n,1}$ remains bounded and $\mathbb{E}K_{n,1} \to +\infty$ (in particular, when $\vec{\nu}$ is regularly varying with index $\alpha \in (0,1]$ or $\alpha = 0$ and $\vec{\nu} \in \Pi_a$ with $a \to \infty$),*

$$\frac{M_{n,0}}{\mathbb{E}M_{n,0}} \xrightarrow{\mathbb{P}} 1,$$

*and the Good–Turing estimator of $M_{n,0}$ defined by $G_{n,0} = K_{n,1}/n$, is multiplicatively consistent in probability:*

$$\frac{G_{n,0}}{M_{n,0}} \xrightarrow{\mathbb{P}} 1.$$

(ii) *If furthermore $\mathbb{E}K_{n,\overline{2}}/\mathbb{E}K_{n,1}$ remains bounded and if, for all $c > 0$, $\sum_{n=0}^{\infty} \exp(-c\mathbb{E} \times K_{n,1}) < \infty$ (in particular, when $\vec{\nu}$ is regularly varying with index $\alpha \in (0,1]$), then these two convergences occur almost surely.*

***Remark 5.1.*** One needs to make assumptions on the sampling distribution to guarantee the consistency of the Good–Turing estimator. In fact, there is no hope to find a universally consistent estimator of the missing mass without any such restrictions, as shown recently by Mossel and Ohannessian [34].

Consistency is a desirable property, but the concentration inequalities provide us with more power, in particular in terms of giving confidence intervals that are asymptotically tight. For brevity, we focus here on the Poisson setting to derive concentration inequalities which in turn yield confidence intervals. A similar, but somewhat more tedious, methodology yields confidence intervals in the binomial setting as well.

### 5.2.1. *Concentration inequalities for $G_0(t) - M_0(t)$*

In the Poisson setting, the analysis of the Good–Turing estimator is illuminating. As noted earlier, the first pleasant observation is that the Good–Turing estimator is an unbiased estimator of the missing mass. Second, the variance of $G_0(t) - M_0(t)$ is simply related to occupancy counts:

$$\operatorname{Var}\bigl(G_0(t) - M_0(t)\bigr) = \frac{1}{t^2}\bigl(\mathbb{E}K_1(t) + 2\mathbb{E}K_2(t)\bigr). \tag{5.3}$$

Third, simple yet often tight concentration inequalities can be obtained for $G_0(t) - M_0(t)$.

**Proposition 5.4.** *The random variable $G_0(t) - M_0(t)$ is sub-gamma on the right tail with variance factor $\operatorname{Var}(G_0(t) - M_0(t))$ and scale factor $1/t$, and sub-gamma on the left tail with variance factor $3\mathbb{E}K(t)/t^2$ and scale factor $1/t$.*

*For all $\lambda \geq 0$,*

(i) $\log \mathbb{E}e^{\lambda(G_0(t) - M_0(t))} \leq \operatorname{Var}(G_0(t) - M_0(t)) t^2 \phi(\frac{\lambda}{t})$, *and*

(ii) $\log \mathbb{E} e^{\lambda(M_0(t)-G_0(t))} \le \frac{3\mathbb{E}K(t)}{2t^2} \frac{\lambda^2}{1-\lambda/t}$.

We are now in a position to build confidence intervals for the missing mass.

**Proposition 5.5.** *With probability larger than* $1-4\delta$, *the following hold*

$$M_0(t) \le G_0(t) + \frac{1}{t}\left(\sqrt{6K(t)\log\frac{1}{\delta}} + 5\log\frac{1}{\delta}\right),$$

$$M_0(t) \ge G_0(t) - \frac{1}{t}\left(\sqrt{2\bigl(K_1(t)+2K_2(t)\bigr)\log\frac{1}{\delta}} + 4\log\frac{1}{\delta}\right).$$

To see that these confidence bounds are asymptotically tight, consider the following central limit theorem. A similar results can be paralleled in the binomial setting.

**Proposition 5.6.** *If the counting function* $\vec{\nu}$ *is regularly varying with index* $\alpha \in (0,1]$, *the following central limit theorem holds for the ratio* $G_0(t)/M_0(t)$:

$$\frac{\mathbb{E}K_1(t)}{\sqrt{\mathbb{E}K_1(t)+2\mathbb{E}K_2(t)}}\left(\frac{G_0(t)}{M_0(t)} - 1\right) \rightsquigarrow \mathcal{N}(0,1).$$

***Remark 5.2.*** Note that when $\alpha = 1$, this convergence occurs faster: the speed is of order $\sqrt{n\ell_1(n)}$ instead of $\sqrt{n^\alpha \ell(n)}$.

## 5.3. Estimating the number of species

Fisher's number of species problem (Fisher *et al.* [21]) consists of estimating $K_{(1+\tau)n} - K_n$ for $\tau > 0$, the number of distinct new species one would observe if the data collection runs for an additional fraction $\tau$ of time. This was posed primarily within the Poisson model in the original paper (Fisher *et al.* [21]) and later by Efron and Thisted [19], but the same question may also be asked in the binomial model. The following estimates come from straightforward computations on the asymptotics given in Theorems 4.2, 4.3 and 4.5.

**Proposition 5.7.** *If the counting function* $\vec{\nu}$ *is regularly varying with index* $\alpha \in (0,1]$, *letting* $\hat{\alpha}$ *be any of the estimates* $rK_{n,r}/K_{n,\overline{r}}$ *of* $\alpha$ *from Proposition* 5.1, *then any of the following quantities*

$$\bigl(\tau^{\hat{\alpha}} - 1\bigr)K_n, \qquad \frac{\tau^{\hat{\alpha}}-1}{\hat{\alpha}}K_{n,1} \quad \text{and} \quad \left(\prod_{k=2}^{r}\frac{k}{k-1-\hat{\alpha}}\right)\frac{\tau^{\hat{\alpha}}-1}{\hat{\alpha}}K_{n,r}, \qquad r \ge 2,$$

*is a strongly consistent estimate of* $K_{\tau n} - K_n$, *the number of newly discovered species when the sample size is multiplied by* $\tau$.

*If the counting function* $\vec{\nu}$ *is in* $\Pi_a$, *with* $a(n) \to +\infty$, *then, for each* $r \ge 1$,

$$\log(\tau) r K_{n,r}$$

*is an estimate of $K_{\tau n} - K_n$, consistent in probability.*

# 6. Discussion

To conclude the paper, we review our results in a larger context, and propose some connections, extensions, and open problems.

## 6.1. The cost of Poissonization and negative correlation

Resorting to Poissonization or negative correlation may have a price. It may lead to variance overestimates. Gnedin *et al.* ([22], Lemma 1), asserts that for some constant $c$

$$\left|\operatorname{Var}\bigl(K(n)\bigr) - \operatorname{Var}(K_n)\right| \le \frac{c}{n} \max\bigl(1, \mathbb{E}K_1(n)^2\bigr).$$

This bound conveys a mixed message. As $\mathbb{E}K_1(n)/n$ tends to 0, it asserts that

$$\left|\operatorname{Var}\bigl(K(n)\bigr) - \operatorname{Var}(K_n)\right| / \mathbb{E}K_1(n)$$

tends to 0. But there exist scenarios where $\mathbb{E}K_1(n)^2/n$ tends to infinity. It is shown in Gnedin *et al.* [22] that $\mathbb{E}K_1(n)^2/(n \operatorname{Var}(K(n)))$ tends to 0, so that, as soon as $n \operatorname{Var}(K(n))$ tends to infinity (which might not always be the case), the two variances $\operatorname{Var}(K_n)$ and $\operatorname{Var}(K(n))$ are asymptotically equivalent.

It would be interesting to find necessary and sufficient conditions under which there is equivalence. Though these aren't generally known, it is instructive to compare $\operatorname{Var}(K(n))$, $\operatorname{Var}(K_n)$ and $\operatorname{Var}^{\text{ind}}(K_n)$ the variance upper bound obtained from negative correlation by bounding their differences. For instance, one can show that for any sampling distribution we have:

$$\frac{\mathbb{E}K_2(2n)}{n} \le \operatorname{Var}\bigl(K(n)\bigr) - \operatorname{Var}^{\text{ind}}(K_n) \le \frac{2\mathbb{E}K_2(n)}{n}$$

and

$$0 \le \operatorname{Var}^{\text{ind}}(K_n) - \operatorname{Var}(K_n) \le \frac{(\mathbb{E}K_{n,1})^2}{n} - \frac{\mathbb{E}K_{2n,2}}{2n-1}.$$

These bounds are insightful but, without any further assumptions on the sampling distribution, they are not sufficient to prove asymptotic equivalence.

## 6.2. Extensions of regular variation

The regular variation hypothesis is an elegant framework, allowing one to derive, thanks to Karamata and Tauberian theorems, simple and intelligible equivalents for various moments. As we have seen in Remark 4.1, regular variation comes very close to being a necessary condition for exponential concentration. It may however seem too stringent. Without getting too specific, let us mention that other less demanding hypotheses also yield the asymptotic relative orders that work in favor of the concentration of the missing mass. For instance, referring back to Remark 4.1, one could instead ask for:

$$0 < \liminf_{t\to\infty} \frac{\Phi_{\overline{2}}(t)}{\Phi_2(t)} \le \limsup_{t\to\infty} \frac{\Phi_{\overline{2}}(t)}{\Phi_2(t)} < \infty.$$

Recalling that $\Phi'_{\overline{2}}(t) = \frac{2\Phi_2(t)}{t}$, and applying Corollary 2.6.2. of Bingham *et al.* [11], one obtains that $\Phi_2$ is in the class $OR$ of $O$-regularly varying functions and $\Phi_{\overline{2}}$ is in the class $ER$ of extended regularly varying functions, that is, for all $\lambda \ge 1$

$$0 < \liminf_{t\to\infty} \frac{\Phi_2(\lambda t)}{\Phi_2(t)} \le \limsup_{t\to\infty} \frac{\Phi_2(\lambda t)}{\Phi_2(t)} < \infty$$

and

$$\lambda^d \le \liminf_{t\to\infty} \frac{\Phi_{\overline{2}}(\lambda t)}{\Phi_{\overline{2}}(t)} \le \limsup_{t\to\infty} \frac{\Phi_{\overline{2}}(\lambda t)}{\Phi_{\overline{2}}(t)} \le \lambda^c,$$

for some constants $c$ and $d$. Observe that this result, which is the equivalent of Karamata's theorem, differs from the regular variation setting, in the sense that the control on the derivative $\Phi_2$ is looser than the one on $\Phi_{\overline{2}}$, whereas, in the Karamata theorem, both the function and its derivative inherit the regular variation property.

We can in turn show that $\Phi(t) = \mathbb{E}K(t)$ is in the class $OR$ and, by Theorem 2.10.2 of Bingham *et al.* [11], this is equivalent to $\vec{\nu}(1/\cdot) \in OR$, as $\Phi$ is the Laplace–Stieltjes transform of $\vec{\nu}$.

## 6.3. Random measures

As noted by Gnedin *et al.* [22], the asymptotics for the moments of the occupancy counts in the regular variation setting is still valid when the frequencies $(p_j)_{j\ge1}$ are random, in which case the measure $\nu$ is defined by

$$\mathbb{E}\left[\sum_{j=1}^{\infty} f(p_j)\right] = \int_0^1 f(x)\nu(\mathrm{d}x),$$

for all functions $f \ge 0$. We can also define the measure $\nu_1$ by

$$\mathbb{E}\left[\sum_{j=1}^{\infty} p_j f(p_j)\right] = \int_0^1 f(x)\nu_1(\mathrm{d}x),$$

for all functions $f \ge 0$. This measure corresponds to the distribution of the frequency of the first discovered symbol.

For instance, when $(p_j)_{j\ge1}$ are Poisson–Dirichlet($\alpha$,0) with $0 < \alpha < 1$, the measure $\nu_1(\mathrm{d}x)$ is the size-biased distribution of PD($\alpha$, 0), that is Beta $\mathcal{B}(1-\alpha,\alpha)$ (see Pitman and Yor [37]).

Thus, we have:

$$\begin{aligned}\nu_1[0,x] &= \frac{1}{\mathcal{B}(1-\alpha,\alpha)}\int_0^x t^{-\alpha}(1-t)^{\alpha-1}\,\mathrm{d}t \\ &\underset{x\to 0}{\sim} \frac{x^{1-\alpha}}{(1-\alpha)\mathcal{B}(1-\alpha,\alpha)}\end{aligned}$$

and, by Gnedin *et al.* [22], Proposition 13, this is equivalent to

$$\vec{\nu}(x) \underset{x\to 0}{\sim} \frac{1}{\alpha\mathcal{B}(1-\alpha,\alpha)} x^{-\alpha}.$$

Thus, denoting by $N(x)$ the random number of frequencies $p_j$ which are larger than $x$, the expectation $\vec{\nu}(x) = \mathbb{E}N(x)$ is regularly varying. One can also show that the mass-partition mechanism of the distribution $\mathrm{PD}(\alpha,0)$ almost surely generates $N(x)$ to be regularly varying. To see this, refer to Pitman and Yor [37], Proposition 10 or to Bertoin [10], Proposition 2.6, which assert that the limit

$$L := \lim_{n\to\infty} np_n^{\alpha}$$

exists almost surely. This is equivalent to

$$N(x) \underset{x\to 0}{\sim} x^{-\alpha}L \qquad \text{almost surely.}$$

The $\mathrm{PD}(\alpha,0)$ distribution can be generated through a Poisson process with intensity measure $\nu([x,\infty]) = cx^{-\alpha}$. Without entering into further details, let us mention that similar almost sure results hold even when the intensity measure $\nu$ is not a strict power, but satisfies the property

$$\nu\big([x,\infty]\big) \underset{x\to 0}{\sim} x^{-\alpha}\ell(x),$$

with $\ell$ slowly varying, Gnedin [23], Section 6. Working with a regular variation hypothesis thus gives us more flexibility than assuming specific Bayesian priors.

# 7. Proofs

## 7.1. Fundamental techniques

### 7.1.1. *Efron–Stein–Steele inequalities*

Our variance bounds mostly follow from the Efron–Stein–Steele inequality (Efron and Stein [18]), which states that when a random variable is expressed as a function of many independent random variables, its variance can be controlled by the sum of the local fluctuations.

**Theorem 7.1.** *Let $\mathcal{X}$ be some set, $(X_1, X_2, \ldots, X_n)$ be independent random variables taking values in $\mathcal{X}$, $f : \mathcal{X}^n \to \mathbb{R}$ be a measurable function of $n$ variables, and $Z = f(X_1, X_2, \ldots, X_n)$.*

*For all $i \in \{1, \dots, n\}$, let $X^{(i)} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ and $\mathbb{E}^{(i)} Z = \mathbb{E}[Z|X^{(i)}]$. Then, letting $v = \sum_{i=1}^n \mathbb{E}[(Z - \mathbb{E}^{(i)} Z)^2]$,*

$$\operatorname{Var}[Z] \le v.$$

*If $X'_1, \dots, X'_n$ are independent copies of $X_1, \dots, X_n$, then letting $Z'_i = f(X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n)$,*

$$v = \sum_{i=1}^n \mathbb{E}\bigl[\bigl(Z - Z'_i\bigr)_+^2\bigr] \le \sum_{i=1}^n \mathbb{E}\bigl[(Z - Z_i)^2\bigr],$$

*where the random variables $Z_i$ are arbitrary $X^{(i)}$-measurable and square-integrable random variables.*

### 7.1.2. *Negative association*

The random variables $K_n$, $K_{n,r}$, and $M_{n,r}$ are sums or weighted sums of Bernoulli random variables. These summands depend on the scores $(X_{n,j})_{j \ge 1}$ and therefore are not independent. Transforming the fixed-$n$ binomial setting into a continuous time Poisson setting is one way to circumvent this problem. This is the Poissonization method. In this setting, the score variables $(X_j(n))_{j \ge 1}$ are independent Poisson variables with respective means $np_j$. Results valid for the Poisson setting can then be transferred to the fixed-$n$ setting, up to approximation costs. For instance, Gnedin *et al.* [22] (Lemma 1) provide bounds on the discrepancy between expectations and variances in the two settings. (See also our discussion in Section 6.1.)

Another approach to deal with the dependence is to invoke the notion of negative association, which provides a systematic comparison between moments of certain monotonic functions of the occupancy scores. In our present setting, this will primarily be useful for bounding the logarithmic moment generating function, which is an expectation of products, by products of expectations, thus recovering the structure of independence. This has already been used to derive exponential concentration for occupancy counts (see Dubhashi and Ranjan [17], Shao [39], McAllester and Schapire [33], Ohannessian and Dahleh [35]). It is also useful for bounding variances. We use this notion throughout the proofs, and therefore present it here formally.

***Definition 2 (Negative association).*** *Real-valued random variables $Z_1, \dots, Z_K$ are said to be negatively associated if, for any two disjoint subsets $A$ and $B$ of $\{1, \dots, K\}$, and any two real-valued functions $f : \mathbb{R}^{|A|} \mapsto \mathbb{R}$ and $g : \mathbb{R}^{|B|} \mapsto \mathbb{R}$ that are both either coordinate-wise non-increasing or coordinate-wise non-decreasing, we have*:

$$\mathbb{E}\bigl[f(Z_A) \cdot g(Z_B)\bigr] \le \mathbb{E}\bigl[f(Z_A)\bigr] \cdot \mathbb{E}\bigl[g(Z_B)\bigr].$$

In particular, as far as concentration properties are concerned, sums of negatively associated variables can only do better than sums of independent variables.

**Theorem 7.2 (Dubhashi and Ranjan [17]).** *For each $n \in \mathbb{N}$, the occupancy scores $(X_{n,j})_{j \ge 1}$ are negatively associated.*

As monotonic functions of negatively associated variables are also negatively associated, the variables $(\mathbb{I}_{\{X_{n,j}>0\}})_{j\geq 1}$ (respectively, $(\mathbb{I}_{\{X_{n,j}=0\}})_{j\geq 1}$) are negatively associated as increasing (respectively, decreasing) functions of $(X_{n,j})_{j\geq 1}$. This is of pivotal importance for our proofs of concentration results for $K_n$ and $M_{n,0}$. For $r \geq 1$, the variables $(\mathbb{I}_{\{X_{n,j}=r\}})_{j\geq 1}$ appearing in $K_{n,r}$ are not negatively associated. However, following Ohannessian and Dahleh [35], one way to deal with this problem is to observe that

$$K_{n,r} = K_{n,\overline{r}} - K_{n,\overline{r+1}},$$

recalling that $K_{n,\overline{r}} = \sum_{j=1}^{\infty} \mathbb{I}_{\{X_{n,j}\geq r\}}$ is the number of urns that contain at least $r$ balls and that the Bernoulli variables appearing in $K_{n,\overline{r}}$ are negatively associated.

7.1.3. *Potter's inequalities and its variants*

One useful result from regular variation theory is provided by Potter's inequality (see Bingham *et al.* [11], de Haan and Ferreira [16], for proofs and refinements).

**Theorem 7.3 (Potter–Drees inequalities).**

(i) *If* $f \in \mathrm{RV}_\gamma$, *then for all* $\delta > 0$, *there exists* $t_0 = t_0(\delta)$, *such that for all* $t, x\colon \min(t, tx) > t_0$,

$$(1-\delta)x^\gamma \min\bigl(x^\delta, x^{-\delta}\bigr) \leq \frac{f(tx)}{f(t)} \leq (1+\delta)x^\gamma \max\bigl(x^\delta, x^{-\delta}\bigr).$$

(ii) *If* $\ell \in \Pi_a$, *then for all* $\delta_1, \delta_2$, *there exists* $t_0$ *such that for all* $t \geq t_0$, *for all* $x \geq 1$,

$$(1-\delta_2)\frac{1-x^{\delta_1}}{\delta_1} - \delta_2 < \frac{\ell(tx)-\ell(t)}{a(t)} < (1+\delta_2)\frac{x^{\delta_1}-1}{\delta_1} + \delta_2.$$

## 7.2. Occupancy counts

7.2.1. *Variance bounds for occupancy counts*

**Proof of Proposition 3.1.** Recall that in the Poisson setting,

$$\frac{\mathrm{d}\mathbb{E}K(t)}{\mathrm{d}t} = \frac{\mathbb{E}K_1(t)}{t} = \mathbb{E}M_0(t).$$

This entails

$$\operatorname{Var}\bigl(K(t)\bigr) = \sum_{j=1}^{\infty} \mathrm{e}^{-tp_j}\bigl(1-\mathrm{e}^{-tp_j}\bigr) = \mathbb{E}K(2t) - \mathbb{E}K(t) = \int_t^{2t} \mathbb{E}M_0(s)\,\mathrm{d}s.$$

Now, as $\mathbb{E}M_0(s)$ is non-increasing,

$$\frac{\mathbb{E}K_1(2t)}{2} = t\mathbb{E}M_0(2t) \leq \operatorname{Var}\bigl(K(t)\bigr) \leq t\mathbb{E}M_0(t) = \mathbb{E}K_1(t).$$

Moving on to the binomial setting, let $K_n^i$ denote the number of occupied urns when the $i$th ball is replaced by an independent copy. Then

$$\operatorname{Var}(K_n) \le \mathbb{E}\left[\sum_{i=1}^{n}\left(K_n - K_n^i\right)_+^2\right],$$

where $(K_n - K_n^i)_+$ denotes the positive part. Now, $K_n - K_n^i$ is positive if and only if ball $i$ is moved from a singleton into in a nonempty urn. Thus $\operatorname{Var}(K_n) \le \mathbb{E}[K_{n,1}(1 - M_{n,0})]$. □

**Proof of Proposition 3.2.** The bound $r\mathbb{E}K_{n,r}$ follows from the Efron–Stein inequality: denoting by $K_{n,\overline{r}}^{(i)}$ the number of cells with occupancy score larger than $r$ when ball $i$ is removed, then

$$K_{n,\overline{r}} - K_{n,\overline{r}}^{(i)} = \begin{cases} 1, & \text{if ball } i \text{ is in a } r\text{-ton}, \\ 0, & \text{otherwise.} \end{cases}$$

And thus, we get $\sum_{i=1}^{n}(K_{n,\overline{r}} - K_{n,\overline{r}}^{(i)})^2 = rK_{n,r}$.

The second bound follows from the negative association of the variables $(\mathbb{I}_{\{X_{n,j}\ge r\}})_j$ (negative correlation is actually sufficient):

$$\operatorname{Var}\left(\sum_{j=1}^{\infty}\mathbb{I}_{\{X_{n,j}\ge r\}}\right) \le \sum_{j=1}^{\infty}\operatorname{Var}(\mathbb{I}_{\{X_{n,j}\ge r\}}) \le \mathbb{E}K_{n,\overline{r}}.$$

□

7.2.2. *Concentration inequalities for occupancy counts*

**Proof of Proposition 3.4.** Let $X_{n,j}$ denote the occupancy score of cell $j$, $j \in \mathbb{N}$, then

$$K_{n,\overline{r}} = \sum_{j=1}^{\infty}\mathbb{I}_{\{X_{n,j}\ge r\}}.$$

As noted in Section 7.1.2, $K_{n,\overline{r}}$ is a sum of negatively associated Bernoulli random variables. Moreover, the Efron–Stein inequality implies that for each $j \in \mathbb{N}$,

$$\operatorname{Var}(\mathbb{I}_{\{X_{n,j}\ge r\}}) \le r\mathbb{E}\mathbb{I}_{\{X_{n,j}=r\}}.$$

Thus we have

$$\begin{aligned} \log \mathbb{E}e^{\lambda(K_{n,\overline{r}} - \mathbb{E}K_{n,\overline{r}})} &\le \sum_{j=1}^{\infty}\log \mathbb{E}e^{\lambda(\mathbb{I}_{\{X_{n,j}\ge r\}} - \mathbb{E}\mathbb{I}_{\{X_{n,j}\ge r\}})} \\ &\le \sum_{j=1}^{\infty}\operatorname{Var}(\mathbb{I}_{\{X_{n,j}\ge r\}})\phi(\lambda) \end{aligned}$$

$$
\begin{aligned}
&\le \phi(\lambda) \sum_{j=1}^{\infty} r \mathbb{E}\mathbb{I}_{\{X_{n,j}=r\}} \\
&= \phi(\lambda) r \mathbb{E} K_{n,r},
\end{aligned}
$$

where the first inequality comes from negative association, the second inequality is Bennett's inequality for Bernoulli random variables, and the last inequality comes from the Efron–Stein inequality. The other bound comes from the fact that $\mathrm{Var}(\mathbb{I}_{\{X_{n,j}\ge r\}}) \le \mathbb{E}\mathbb{I}_{\{X_{n,j}\ge r\}}$. □

**Proof of Proposition 3.5.** As $K_{n,r} = K_{n,\overline{r}} - K_{n,\overline{r+1}}$,

$$
\begin{aligned}
&\{K_{n,r} \ge \mathbb{E}K_{n,r} + x\} \\
&\quad \subseteq \left\{K_{n,\overline{r}} \ge \mathbb{E}K_{n,\overline{r}} + \frac{x}{2}\right\} \cup \left\{K_{n,\overline{r+1}} \le \mathbb{E}K_{n,\overline{r+1}} - \frac{x}{2}\right\}.
\end{aligned}
$$

By Proposition 3.4, Bernstein inequalities hold for both $K_{n,\overline{r}}$ and $K_{n,\overline{r+1}}$, with variance proxies $\mathbb{E}rK_{n,r}$ (or $\mathbb{E}K_{n,\overline{r}}$) and $(r+1)K_{n,r+1}$ (or $\mathbb{E}K_{n,\overline{r+1}} \le \mathbb{E}K_{n,\overline{r}}$) respectively. Hence,

$$
\begin{aligned}
&\mathbb{P}\{K_{n,r} \ge \mathbb{E}K_{n,r} + x\} \\
&\quad \le \exp\left(-\frac{x^2/4}{2(r\mathbb{E}K_{n,r} + x/6)}\right) + \exp\left(-\frac{x^2/4}{2((r+1)\mathbb{E}K_{n,r+1})}\right) \\
&\quad \le 2\exp\left(-\frac{x^2/4}{2(\max(r\mathbb{E}K_{n,r}, (r+1)K_{n,r+1}) + x/6)}\right).
\end{aligned}
$$

The same reasoning works for the alternative variance proxies and for the left tails. □

## 7.3. Missing mass

7.3.1. *Variance bounds for the missing mass*

**Proof of Proposition 3.6.** In the Poisson setting,

$$
\mathrm{Var}\big(M_0(t)\big) = \sum_{j=1}^{\infty} p_j^2 \mathrm{e}^{-tp_j}\big(1 - \mathrm{e}^{-tp_j}\big) \le \sum_{j=1}^{\infty} p_j^2 \mathrm{e}^{-tp_j} = \frac{2}{t^2}\mathbb{E}K_2(t).
$$

In the binomial setting, by negative correlation,

$$
\mathrm{Var}(M_{n,0}) \le \sum_{j=1}^{\infty} p_j^2\big(1 - (1-p_j)^n\big)(1-p_j)^n \le \sum_{j=1}^{\infty} p_j^2 \mathrm{e}^{-np_j} = \frac{2}{n^2}\mathbb{E}K_2(n).
$$

□

7.3.2. *Concentration inequalities for the missing mass*

**Proof of Proposition 3.7.** For all $\lambda \in \mathbb{R}$,

$$\begin{aligned}\log \mathbb{E}\big[e^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\big] &= \log \mathbb{E}\big[e^{\lambda \sum_{j=1}^{\infty} p_j(Y_j-\mathbb{E}Y_j)}\big] \\ &\le \sum_{j=1}^{\infty} \log \mathbb{E}\big[e^{\lambda p_j(Y_j-\mathbb{E}[Y_j])}\big] \\ &\le \sum_{j=1}^{\infty} (1-p_j)^n\big(1-(1-p_j)^n\big)\phi(\lambda p_j),\end{aligned}$$

where the first inequality comes from negative association, and the second is Bennett's inequality for Bernoulli random variables.

Noting that $\lim_{\lambda\to 0_-} \phi(\lambda)/\lambda^2 = \lim_{\lambda\to 0_+} \phi(\lambda)/\lambda^2 = 1/2$, the function $\lambda \mapsto \phi(\lambda)/\lambda^2$ has a continuous increasing extension on $\mathbb{R}$. Hence, for $\lambda \le 0$, we have $\phi(\lambda) \le \lambda^2/2$.

Thus, for $\lambda < 0$,

$$\begin{aligned}\log \mathbb{E}\big[e^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\big] &\le \sum_{j=1}^{\infty} p_j^2(1-p_j)^n\big(1-(1-p_j)^n\big)\frac{\lambda^2}{2} \\ &= \sum_{j=1}^{\infty} p_j^2 \operatorname{Var}[Y_j]\frac{\lambda^2}{2}.\end{aligned}$$

Recall that $\sum_{j=1}^{\infty} p_j^2 \operatorname{Var}[Y_j] \le 2\mathbb{E}K_2(n)/n^2$ (Proposition 3.6). □

**Proof of Proposition 3.8.** From the beginning of the proof of Proposition 3.7, that is, thanks to negative association and to the fact that each Bernoulli random variable satisfies a Bennett inequality,

$$\log \mathbb{E}\big[e^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\big] \le \sum_{j=1}^{\infty} e^{-np_j}\phi(\lambda p_j).$$

Now, using the power series expansion of $\phi$,

$$\begin{aligned}\sum_{j=1}^{\infty} e^{-np_j}\phi(\lambda p_j) &= \sum_{j=1}^{\infty} e^{-np_j} \sum_{r=2}^{\infty} \frac{(\lambda p_j)^r}{r!} \\ &= \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \sum_{j=1}^{\infty} e^{-np_j}\frac{(np_j)^r}{r!}.\end{aligned}$$

We recognize that for each $r$, $\sum_{j=1}^{\infty} \mathrm{e}^{-np_j} \frac{(np_j)^r}{r!} = \mathbb{E}K_r(n)$, so that

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \le \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \mathbb{E}K_r(n).$$

□

**Proof of Theorem 3.9.** Using Proposition 3.8 and noticing that for each $r \ge 2$, $\mathbb{E}K_r(n) \le \mathbb{E}K_{\overline{2}}(n)$, we immediately obtain that

$$\begin{aligned}\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] &\le \mathbb{E}K_{\overline{2}}(n) \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \\ &= \lambda^2 \frac{\mathbb{E}K_{\overline{2}}(n)/n^2}{1-\lambda/n},\end{aligned}$$

which concludes the proof. □

## 7.4. Regular variation

**Proof of Proposition 4.4.** By monotonicity of $K_{n,\overline{r}}$, we have the following strong law for any sampling distribution

$$K_{n,\overline{r}} = \sum_{s=r}^{\infty} K_{n,s} \underset{+\infty}{\sim} \sum_{s=r}^{\infty} \mathbb{E}K_s(n) \qquad \text{a.s.,}$$

(see Gnedin *et al.* [22], the discussion after Proposition 2). Recall that $X_j(n) \sim \mathcal{P}(np_j)$ and that, if $Y \sim \mathcal{P}(\lambda)$, then $\mathbb{P}[Y \le k] = \frac{\Gamma(k+1,\lambda)}{k!}$, where $\Gamma(z,x) = \int_x^{+\infty} \mathrm{e}^{-t} t^{z-1}\,\mathrm{d}t$ is the incomplete Gamma function. Hence

$$\begin{aligned}\sum_{s=r}^{\infty} \mathbb{E}K_s(n) &= \sum_{j=1}^{\infty} \mathbb{P}\left[X_j(n) \ge r\right] \\ &= \sum_{j=1}^{\infty} \frac{1}{(r-1)!} \int_0^{np_j} \mathrm{e}^{-t} t^{r-1}\,\mathrm{d}t \\ &= \frac{1}{(r-1)!} \int_0^1 \int_0^{nx} \mathrm{e}^{-t} t^{r-1} \mathrm{d}t \cdot \nu(\mathrm{d}x) \\ &= \frac{1}{(r-1)!} \int_0^1 n\mathrm{e}^{-nx} (nx)^{r-1} \vec{\nu}(x)\,\mathrm{d}x \\ &= \frac{1}{(r-1)!} \int_0^{+\infty} \mathrm{e}^{-z} z^{r-1} \vec{\nu}(z/n)\,\mathrm{d}z \\ &\underset{+\infty}{\sim} \frac{\vec{\nu}(1/n)}{(r-1)!} \Gamma(r-\alpha).\end{aligned}$$

In particular,

$$K_{n,\overline{r}} \underset{+\infty}{\sim} \frac{r K_{n,r}}{\alpha} \qquad \text{a.s.} \qquad \square$$

**Proof of Theorem 4.6.** Let us recall Proposition 3.8:

$$\log \mathbb{E}\big[e^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\big] \le \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \mathbb{E}K_r(n).$$

Now, bounding each $\mathbb{E}K_r(n)$ by $\mathbb{E}K_{\overline{2}}(n)$ is not sufficient to get the right order for the variance: $\mathbb{E}K_{\overline{2}}(n)$ is of order $\ell(n)$ whereas $\operatorname{Var} M_{n,0}$ is of order $a(n)/n^2$.

We explore more carefully the structure of $\mathbb{E}K_r(n)$ and show that these quantities are uniformly (in $r$) bounded by a function of order $a(n)$ for large enough $n$, that is, that there exists $n^* \in \mathbb{N}$ and $C \in \mathbb{R}_+$ such that for all $n \ge n^*$, for all $r \ge 1$, $\mathbb{E}K_r(n) \le Ca(n)$.

Before going into the proof, we observe that for $r \ge n/a(n)$, the result is true. Indeed, from the identity $\sum_{r=1}^{\infty} r\mathbb{E}K_r(n) = n$, we deduce that $r\mathbb{E}K_{\overline{r}}(n) \le n$, so that for $r \ge n/a(n)$, $\mathbb{E}K_{\overline{r}}(n) \le a(n)$. Thus, we assume that $r \le n/a(n)$.

First, we easily deal with the contribution to $\mathbb{E}K_r(n)$ of the symbols with probability less than $1/n$. Indeed

$$I_1^r := \int_0^{1/n} e^{-nx} \frac{(nx)^r}{r!} \nu(dx) \le \int_0^{1/n} e^{-nx} \frac{(nx)^2}{2!} \nu(dx) \le \mathbb{E}K_2(n).$$

As $\mathbb{E}K_2(n) \sim a(n)/2$, for all $\delta_0$, there exists $n_0$ such that for all $n \ge n_0$, for all $r \ge 1$, $I_1^r \le (1+\delta_0)/2$.

For the contribution of the symbols with probability larger than $1/n$, integration by part and change of variable yield:

$$\begin{aligned}
I_2^r &:= \int_{1/n}^1 e^{-nx} \frac{(nx)^r}{r!} \nu(dx) \\
&= \left[e^{-nx} \frac{(nx)^r}{r!} \big(-\vec{\nu}(x)\big)\right]_{1/n}^1 + \int_{1/n}^1 e^{-nx} \frac{n^r}{r!} \big(rx^{r-1} - nx^r\big) \vec{\nu}(x)\, dx \\
&= \frac{\vec{\nu}(1/n)e^{-1}}{r!} + \int_1^{\infty} e^{-z} \left(\frac{z^{r-1}}{(r-1)!} - \frac{z^r}{r!}\right) \vec{\nu}(z/n)\, dz.
\end{aligned}$$

As $\int_1^{\infty} e^{-z}(\frac{z^{r-1}}{(r-1)!} - \frac{z^r}{r!})\, dz = -\mathbb{P}[\mathcal{P}(1) = r] = -e^{-1}/r!$, we can rearrange the previous expression:

$$I_2^r = a(n) \int_1^{\infty} e^{-z} \left(\frac{z^{r-1}}{(r-1)!} - \frac{z^r}{r!}\right) \frac{\vec{\nu}(z/n) - \vec{\nu}(1/n)}{a(n)}\, dz.$$

Notice that when $z \in [1, r]$, the integrand is negative, so we simply ignore this part of the integral and restrict ourselves to

$$I_3^r := \int_r^\infty \mathrm{e}^{-z}\left(\frac{z^r}{r!} - \frac{z^{r-1}}{(r-1)!}\right)\frac{\vec{\nu}(1/n) - \vec{\nu}(z/n)}{a(n)}\,\mathrm{d}z,$$

which we try to bound by a constant term for $n$ greater than some integer that does not depend on $r$.

The main ingredient of our proof is the next version of the Potter–Drees inequality (see Theorem 7.3 in Section 7.1.3 and de Haan and Ferreira [16], point 4 of Corollary B.2.15): for $\ell \in \Pi_a$, for arbitrary $\delta_1, \delta_2$, there exists $t_0$ such that for all $t \geq t_0$, and for all $x \leq 1$ with $tx \geq t_0$,

$$(1-\delta_2)\frac{1 - x^{-\delta_1}}{\delta_1} - \delta_2 < \frac{\ell(t) - \ell(tx)}{a(t)} < (1+\delta_2)\frac{x^{-\delta_1} - 1}{\delta_1} + \delta_2.$$

Thus, for arbitrary $\delta_1, \delta_2$, there exists $n_1$ such that, for all $n \geq n_1$, for all $z \in [1, n/n_1]$,

$$\frac{\vec{\nu}(1/n) - \vec{\nu}(z/n)}{a(n)} \leq (1+\delta_2)\frac{z^{\delta_1} - 1}{\delta_1} + \delta_2.$$

As $r \leq n/a(n)$, taking, if necessary, $n$ large enough so that $a(n) \geq n_1$, we have $r \leq n/n_1$ and

$$\begin{aligned} I_3^r &\leq \int_r^{n/n_1} \mathrm{e}^{-z}\left(\frac{z^r}{r!} - \frac{z^{r-1}}{(r-1)!}\right)\left((1+\delta_2)\frac{z^{\delta_1} - 1}{\delta_1} + \delta_2\right)\mathrm{d}z \\ &\quad + \int_{n/n_1}^\infty \mathrm{e}^{-z}\left(\frac{z^r}{r!} - \frac{z^{r-1}}{(r-1)!}\right)\frac{\vec{\nu}(1/n) - \vec{\nu}(z/n)}{a(n)}\,\mathrm{d}z \\ &=: I_4^r + I_5^r, \end{aligned}$$

with

$$\begin{aligned} I_4^r &\leq \delta_2 + \frac{1+\delta_2}{\delta_1}\int_r^\infty \mathrm{e}^{-z}\left(\frac{z^{r+\delta_1}}{r!} - \frac{z^r}{r!} + \frac{z^{r-1}}{(r-1)!} - \frac{z^{r-1+\delta_1}}{(r-1)!}\right)\mathrm{d}z \\ &\leq \delta_2 + \frac{1+\delta_2}{\delta_1}\int_r^\infty \mathrm{e}^{-z}\left(\frac{z^{r+\delta_1}}{r!} - \frac{z^{r-1+\delta_1}}{(r-1)!}\right)\mathrm{d}z \\ &= \delta_2 + \frac{1+\delta_2}{\delta_1}\left(\frac{\Gamma(r+1+\delta_1, r)}{\Gamma(r+1)} - \frac{\Gamma(r+\delta_1, r)}{\Gamma(r)}\right), \end{aligned}$$

where $\Gamma(a, x) = \int_x^\infty \mathrm{e}^{-t} t^{a-1}\,\mathrm{d}t$ is the incomplete Gamma function. Using the fact that $\Gamma(a, x) = (a-1)\Gamma(a-1, x) + x^{a-1}\mathrm{e}^{-x}$, we have

$$\begin{aligned} I_4^r &\leq \delta_2 + \frac{1+\delta_2}{\delta_1\Gamma(r+1)}\big(\Gamma(r+1+\delta_1, r) - (r+\delta_1)\Gamma(r+\delta_1, r) + \delta_1\Gamma(r+\delta_1, r)\big) \\ &= \delta_2 + \frac{1+\delta_2}{\delta_1}\left(\frac{r^{r+\delta_1}\mathrm{e}^{-r}}{r!} + \delta_1\frac{\Gamma(r+\delta_1, r)}{\Gamma(r+1)}\right). \end{aligned}$$

By Stirling's inequality, for all $r$,

$$\frac{r^{r+\delta_1}\mathrm{e}^{-r}}{r!} \le r^{\delta_1}(2\pi r)^{-1/2}.$$

Thus, taking $\delta_1 = 1/4$, the right-hand term is uniformly bounded by 1. And $\frac{\Gamma(r+\delta_1, r)}{\Gamma(r+1)}$ is also bounded by 1. Thus

$$I_4^r \le \delta_2 + \frac{1+\delta_2}{\delta_1}(1+\delta_1)$$

and

$$\begin{aligned} I_5^r &\le \frac{\vec{\nu}(1/n)}{a(n)} \int_{n/n_1}^{\infty} \mathrm{e}^{-z}\left(\frac{z^r}{r!} - \frac{z^{r-1}}{(r-1)!}\right)\mathrm{d}z \\ &= \frac{\vec{\nu}(1/n)}{a(n)}\mathrm{e}^{-n/n_1}\frac{(n/n_1)^r}{r!} \\ &\le \frac{\vec{\nu}(1/n)}{a(n)}\mathrm{e}^{-n/n_1}\frac{(n/n_1)^{\lfloor n/n_1\rfloor}}{\lfloor n/n_1\rfloor!}. \end{aligned}$$

By Stirling's inequality, this bound is smaller than $\frac{\vec{\nu}(1/n)}{a(n)}(2\pi(n/n_1))^{-1/2}$, which tends to 0 as $n \to \infty$. Thus, there exists $n_2$ such that for all $n \ge n_2$, and all $r \le n/n_1$, $I_5^r \le \delta_2$.

In the end, we get that for all $\delta_0 \ge 0$, $\delta_1$ with $0 \le \delta_1 \le 1/4$, and $\delta_2 \ge 0$, there exists $n^* = \max(n_0, n_1, n_2)$ such that for all $n \ge n^*$, for all $r \ge 1$,

$$\mathbb{E}K_r(n) \le a(n)\left(\frac{1+\delta_0}{2} + \delta_2 + \frac{1+\delta_2}{\delta_1}(1+\delta_1) + \delta_2\right).$$

Taking for instance $\delta_1 = 1/4$ and $\delta_0 = \delta_2 = 1/15$, we have that for large enough $n$ and for all $r \ge 1$,

$$\mathbb{E}K_r(n) \le 6a(n)$$

and

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0} - \mathbb{E}M_{n,0})}\right] \le \frac{12a(n)}{n^2} \cdot \frac{\lambda^2}{2(1-\lambda/n)}. \qquad \square$$

**Proof of Proposition 4.8.** Under the condition of the Proposition 4.8, from Grübel and Hitczenko [26], with probability tending to 1, the sample is gap-free, hence the missing mass is $\overline{F}(\max(X_1, \ldots, X_n))$.

The condition of the proposition implies the condition described in Anderson [2], i.e. $\lim_{n\to+\infty} \frac{\overline{F}(n+1)}{\overline{F}(n)} = 0$, to ensure the existence of a sequence of integers $(u_n)_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} \mathbb{P}\left\{\max(X_1, \ldots, X_n) \in \{u_n, u_n+1\}\right\} = 1. \qquad \square$$

## 7.5. Applications

**Proof of Corollary 5.3.** Let us assume that $\mathbb{E}K_{n,1} \to \infty$. Using the fact that $0 \le \mathbb{E}G_{n,0} - \mathbb{E}M_{n,0} \le 1/n$, we notice that as soon as $\mathbb{E}K_{n,1} \to \infty$, $\mathbb{E}G_{n,0} \sim_{+\infty} \mathbb{E}M_{n,0}$. Now by Chebyshev's inequality,

$$\mathbb{P}\left[\left|\frac{M_{n,0}}{\mathbb{E}M_{n,0}} - 1\right| > \varepsilon\right] \le \frac{\mathrm{Var}(M_{n,0})}{\varepsilon^2(\mathbb{E}M_{n,0})^2} \le \frac{2\mathbb{E}K_2(n)}{\varepsilon^2 n^2(\mathbb{E}M_{n,0})^2}$$
$$\sim \frac{2(\mathbb{E}K_{n,2} + o(1))}{\varepsilon^2(\mathbb{E}K_{n,1})^2},$$

where we used that $|\mathbb{E}K_2(n) - \mathbb{E}K_{n,2}| \to 0$ (see Lemma 1, Gnedin *et al.* [22]). On the other hand,

$$\mathbb{P}\left[\left|\frac{K_{n,1}}{\mathbb{E}K_{n,1}} - 1\right| > \varepsilon\right] \le \frac{\mathrm{Var}(K_{n,1})}{\varepsilon^2(\mathbb{E}K_{n,1})^2} \le \frac{\mathbb{E}K_{n,1} + 2\mathbb{E}K_{n,2}}{\varepsilon^2(\mathbb{E}K_{n,1})^2},$$

showing that if, furthermore, $\mathbb{E}K_{n,2}/\mathbb{E}K_{n,1}$ remains bounded, the ratios $M_{n,0}/\mathbb{E}M_{n,0}$, $G_{n,0}/\mathbb{E}G_{n,0}$ and thus $M_{n,0}/G_{n,0}$ converge to 1 in probability. To get almost sure convergence, we use Theorem 3.9 to get that when $\mathbb{E}K_{n,1} \to \infty$,

$$\mathbb{P}\left[\left|\frac{M_{n,0}}{\mathbb{E}M_{n,0}} - 1\right| > \varepsilon\right] \le 2\exp\left(-\frac{\varepsilon^2(\mathbb{E}M_{n,0})^2}{2(2\mathbb{E}K_{\overline{2}}(n)/n^2 + \mathbb{E}M_{n,0}/n)}\right)$$
$$= 2\exp\left(-\frac{\varepsilon^2(\mathbb{E}K_{n,1} + o(\mathbb{E}K_{n,1}))^2}{2(2\mathbb{E}K_{n,\overline{2}} + \mathbb{E}K_{n,1} + o(\mathbb{E}K_{n,1}))}\right).$$

If $\mathbb{E}K_{n,\overline{2}}/\mathbb{E}K_{n,1}$ remains bounded, this becomes smaller than $c_1\exp(-c_2\varepsilon^2\mathbb{E}K_{n,1})$. Hence, if $\exp(-c\mathbb{E}K_{n,1})$ is summable for all $c > 0$, we can apply the Borel–Cantelli lemma and obtain the almost sure convergence of $M_{n,0}/\mathbb{E}M_{n,0}$ to 1. Moreover, by Proposition 3.5,

$$\mathbb{P}\left[\left|\frac{K_{n,1}}{\mathbb{E}K_{n,1}} - 1\right| > \varepsilon\right] \le 4\exp\left(-\frac{\varepsilon^2(\mathbb{E}K_{n,1})^2}{2(4\max(\mathbb{E}K_{n,1}, 2\mathbb{E}K_{n,2}) + 2/3)}\right),$$

which shows that under these assumptions $K_{n,1}/\mathbb{E}K_{n,1}$ also tends to 1 almost surely. □

**Proof of Proposition 5.4.** The random variable $G_0(t) - M_0(t)$ is a sum of independent, centered and bounded random variables, namely

$$G_0(t) - M_0(t) = \frac{1}{t}\sum_{j=1}^{\infty} \mathbb{I}_{X_j(t)=1} - tp_j\mathbb{I}_{X_j(t)=0}.$$

Bound (i) follows immediately from the observation that each $\mathbb{I}_{X_j(t)=1} - tp_j\mathbb{I}_{X_j(t)=0}$ satisfies a Bennett inequality,

$$\begin{aligned}\log \mathbb{E}\mathrm{e}^{\lambda(G_0(t)-M_0(t))} &\leq \sum_{j=1}^{\infty} \mathrm{Var}(\mathbb{I}_{X_j(t)=1} - tp_j\mathbb{I}_{X_j(t)=0})\phi\left(\frac{\lambda}{t}\right) \\ &= \mathrm{Var}\bigl(G_0(t) - M_0(t)\bigr)t^2\phi\left(\frac{\lambda}{t}\right).\end{aligned}$$

Bound (ii) follows from the observation that each $\mathbb{I}_{X_j(t)=0} - \frac{1}{tp_j}\mathbb{I}_{X_j(t)=1}$ satisfies a Bennett inequality,

$$\begin{aligned}\log \mathbb{E}\mathrm{e}^{\lambda(M_0(t)-G_0(t))} &\leq \sum_{j=1}^{\infty} \mathrm{Var}\left(\mathbb{I}_{X_j(t)=0} - \frac{1}{tp_j}\mathbb{I}_{X_j(t)=1}\right)\phi(\lambda p_j) \\ &= \sum_{j=1}^{\infty}\left(1 + \frac{1}{tp_j}\right)\mathrm{e}^{-tp_j}\phi(\lambda p_j) \\ &= \sum_{r\geq 2}\left(\frac{\lambda}{t}\right)^r \sum_{j=1}^{\infty}\left(1 + \frac{1}{tp_j}\right)\mathrm{e}^{-tp_j}\frac{(tp_j)^r}{r!} \\ &= \sum_{r\geq 2}\left(\frac{\lambda}{t}\right)^r\left(\mathbb{E}K_r(t) + \frac{1}{r}\mathbb{E}K_{r-1}(t)\right) \\ &\leq \sum_{r\geq 2}\left(\frac{\lambda}{t}\right)^r \frac{3\mathbb{E}K(t)}{2},\end{aligned}$$

which concludes the proof. □

**Proof of Proposition 5.5.** With probability greater than $1 - 2\delta$, by Proposition 5.4,

$$G_0(t) - M_0(t) \leq \frac{1}{t}\sqrt{2\bigl(\mathbb{E}K_1(t) + 2\mathbb{E}K_2(t)\bigr)\log\frac{1}{\delta}} + \frac{\log(1/\delta)}{3t}$$

and

$$G_0(t) - M_0(t) \geq -\frac{1}{t}\sqrt{6\mathbb{E}K(t)\log\frac{1}{\delta}} - \frac{\log(1/\delta)}{t}.$$

We may now invoke concentration inequalities for $K_1(t) + 2K_2(t)$ and $K(t)$. Indeed, with probability greater than $1 - \delta$, $K(t) \geq \mathbb{E}K(t) - \sqrt{2\mathbb{E}K(t)\log\frac{1}{\delta}}$ which entails $\sqrt{\mathbb{E}K(t)} \leq \sqrt{K(t) + \frac{\log(1/\delta)}{2}} + \sqrt{\frac{\log(1/\delta)}{2}}$.

We have $2K_2(t) + K_1(t) \geq 2\mathbb{E}K_2(t) + \mathbb{E}K_1(t) - \sqrt{4(2\mathbb{E}K_2(t) + \mathbb{E}K_1(t))\log\frac{1}{\delta}}$ with probability greater than $1-\delta$, which entails

$$\sqrt{2\mathbb{E}K_2(t) + \mathbb{E}K_1(t)} \leq \sqrt{\big(2K_2(t) + K_1(t)\big) + \log\frac{1}{\delta}} + \sqrt{\log\frac{1}{\delta}},$$

which concludes the proof. □

**Proof of Proposition 5.6.** The covariance matrix $\mathrm{Cov}(t)$ of $(G_0(t), M_0(t))$ can be written in terms of the expected occupancy counts as

$$\mathrm{Cov}(t) = \frac{1}{t^2}\begin{pmatrix} \mathbb{E}K_1(t) & 0 \\ 0 & 2\mathbb{E}K_2(t) \end{pmatrix} - \frac{\mathbb{E}K_2(2t)}{2t^2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

From Karlin [29], we have

$$\mathrm{Cov}(t)^{-1/2}\begin{pmatrix} G_0(t) - \mathbb{E}G_0(t) \\ M_0(t) - \mathbb{E}M_0(t) \end{pmatrix} \rightsquigarrow \mathcal{N}(0, I_2),$$

where $I_2$ is the identity matrix, which can be rewritten as

$$\Sigma(t)^{-1/2}\begin{pmatrix} \dfrac{G_0(t)}{\mathbb{E}G_0(t)} - 1 \\ \dfrac{M_O(t)}{\mathbb{E}M_0(t)} - 1 \end{pmatrix} \rightsquigarrow \mathcal{N}(0, I_2),$$

with $\Sigma(t) = (\mathbb{E}G_0(t))^{-2}\,\mathrm{Cov}(t)$.

The delta method applied to the function $(x_1, x_2) \mapsto x_1/x_2$ yields

$$\left((1 \; -1)\Sigma(t)\begin{pmatrix} 1 \\ -1 \end{pmatrix}\right)^{-1/2}\left(\frac{G_0(t)}{M_0(t)} - 1\right) \rightsquigarrow \mathcal{N}(0, 1)$$

and

$$\left((1 \; -1)\Sigma(t)\begin{pmatrix} 1 \\ -1 \end{pmatrix}\right)^{-1/2} = \frac{\mathbb{E}K_1(t)}{\sqrt{\mathbb{E}K_1(t) + 2\mathbb{E}K_2(t)}},$$

which concludes the proof. □

***Remark 7.1.*** The proof for the binomial setting is very similar, the only difficulty being that $\mathbb{E}G_{n,0}$ and $\mathbb{E}M_{n,0}$ are no longer equal. However, the bias becomes negligible with respect to the fluctuations, that is, for $v_n$ either $n^\alpha \ell(n)$ or $n\ell_1(n)$

$$\sqrt{v_n}\left(\frac{\mathbb{E}G_{n,0}}{\mathbb{E}M_{n,0}} - 1\right) \underset{n\to\infty}{\to} 0.$$

# References


[1] Acharya, J., Jafarpour, A., Orlitsky, A. and Suresh, A.T. (2013). Optimal probability estimation with applications to prediction and classification. In *Colt* 2013. *J. of Mach. Learn. Research—Proc. Track* **30** 764–796.

[2] Anderson, C.W. (1970). Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *J. Appl. Probab.* **7** 99–113. MR0256441

[3] Bahadur, R.R. (1960). On the number of distinct values in a large sample from an infinite discrete distribution. *Proc. Nat. Inst. Sci. India Part A* **26** 67–75. MR0137256

[4] Barbour, A.D. and Gnedin, A.V. (2009). Small counts in the infinite occupancy scheme. *Electron. J. Probab.* **14** 365–384. MR2480545

[5] Bartroff, J., Goldstein, L. and Işlak, Ü. (2014). Bounded size biased couplings, log concave distributions and concentration of measure for occupancy models. Preprint. Available at arXiv:1402.6769.

[6] Batu, T., Fischer, E., Fortnow, L., Kumar, R., Rubinfeld, R. and White, P. (2001). Testing random variables for independence and identity. In 42*nd IEEE Symposium on Foundations of Computer Science* (*Las Vegas*, *NV*, 2001) 442–451. Los Alamitos, CA: IEEE Computer Soc. MR1948733

[7] Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004). *Statistics of Extremes*: *Theory and Applications*. *Wiley Series in Probability and Statistics*. Chichester: Wiley. MR2108013

[8] Berend, D. and Kontorovich, A. (2013). On the concentration of the missing mass. *Electron. Commun. Probab.* **18** 1–7. MR3011530

[9] Berend, D. and Kontorovich, A. (2015). A finite sample analysis of the Naive Bayes classifier. *J. Mach. Learn. Res.*. **16** 1519–1545.

[10] Bertoin, J. (2006). *Random Fragmentation and Coagulation Processes*. *Cambridge Studies in Advanced Mathematics* **102**. Cambridge: Cambridge Univ. Press. MR2253162

[11] Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1989). *Regular Variation*. *Encyclopedia of Mathematics and Its Applications* **27**. Cambridge: Cambridge Univ. Press. MR1015093

[12] Bogachev, L.V., Gnedin, A.V. and Yakubovich, Y.V. (2008). On the variance of the number of occupied boxes. *Adv. in Appl. Math.* **40** 401–432. MR2412154

[13] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities*. Oxford: Oxford Univ. Press. MR3185193

[14] Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: A review. *J. Amer. Statist. Assoc.* **88** 364–373.

[15] Chen, L. H. Y., Goldstein, L. and Shao, Q.-M. (2010). *Normal Approximation by Stein's Method*. Berlin: Springer.

[16] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory*: *An Introduction*. *Springer Series in Operations Research and Financial Engineering*. New York: Springer. MR2234156

[17] Dubhashi, D. and Ranjan, D. (1998). Balls and bins: A study in negative dependence. *Random Structures Algorithms* **13** 99–124. MR1642566

[18] Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596. MR0615434

[19] Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.

[20] Esty, W.W. (1982). Confidence intervals for the coverage of low coverage samples. *Ann. Statist.* **10** 190–196. MR0642730

[21] Fisher, R.A., Corbet, A.S. and Williams, C.B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* **12** 42–58.

[22] Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probab. Surv.* **4** 146–171. MR2318403
[23] Gnedin, A.V. (2010). Regeneration in random combinatorial structures. *Probab. Surv.* **7** 105–156. MR2684164
[24] Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264. MR0061330
[25] Good, I.J. and Toulmin, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63. MR0077039
[26] Grübel, R. and Hitczenko, P. (2009). Gaps in discrete random samples. *J. Appl. Probab.* **46** 1038–1051. MR2582705
[27] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363
[28] Hwang, H.-K. and Janson, S. (2008). Local limit theorems for finite and infinite urn models. *Ann. Probab.* **36** 992–1022. MR2408581
[29] Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17** 373–401. MR0216548
[30] Kearns, M. and Saul, L. (1998). Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* 311–319. San Mateo, CA: Morgan Kaufmann.
[31] Kolchin, V.F., Sevast'yanov, B.A. and Chistyakov, V.P. (1978). *Random Allocations*. Washington, DC: V.H. Winston & Sons. MR0471016
[32] McAllester, D. and Ortiz, L. (2004). Concentration inequalities for the missing mass and for histogram rule error. *J. Mach. Learn. Res.* **4** 895–911. MR2076001
[33] McAllester, D.A. and Schapire, R.E. (2000). On the convergence rate of Good–Turing estimators. In *Colt* 2000 1–6.
[34] Mossel, E. and Ohannessian, M.I. (2015). On the impossibility of learning the missing mass. Preprint. Available at arXiv:1503.03613.
[35] Ohannessian, M.I. and Dahleh, M.A. (2012). Rare probability estimation under regularly varying heavy tails. In *Colt* 2012. *J. of Mach. Learn. Research—Proc. Track* **23** 21.1–21.24.
[36] Orlitsky, A., Santhanam, N.P. and Zhang, J. (2004). Universal compression of memoryless sources over unknown alphabets. *IEEE Trans. Inform. Theory* **50** 1469–1481. MR2095850
[37] Pitman, J. and Yor, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25** 855–900. MR1434129
[38] Raginsky, M. and Sason, I. (2013). *Concentration of Measure Inequalities in Information Theory, Communications and Coding* **10**. Boston: Now Publishers.
[39] Shao, Q.-M. (2000). A comparison theorem on moment inequalities between negatively associated and independent random variables. *J. Theoret. Probab.* **13** 343–356. MR1777538
[40] Valiant, G. and Valiant, P. (2011). Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC'11—Proceedings of the* 43*rd ACM Symposium on Theory of Computing* 685–694. New York: ACM. MR2932019

*Received December 2014 and revised May 2015*