Optimal prediction of the number of unseen species

Alon Orlitsky[a], Ananda Theertha Suresh[b,1], and Yihong Wu[c]

[a]Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093; [b]Google Research, New York, NY 10011; and [c]Department of Statistics, Yale University, New Haven, CT 06511

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved September 16, 2016 (received for review May 14, 2016)

Estimating the number of unseen species is an important problem in many scientific endeavors. Its most popular formulation, introduced by Fisher et al. [Fisher RA, Corbet AS, Williams CB (1943) *J Animal Ecol* 12(1):42–58], uses *n* samples to predict the number *U* of hitherto unseen species that would be observed if $t \cdot n$ new samples were collected. Of considerable interest is the largest ratio *t* between the number of new and existing samples for which *U* can be accurately predicted. In seminal works, Good and Toulmin [Good I, Toulmin G (1956) *Biometrika* 43(102):45–63] constructed an intriguing estimator that predicts *U* for all $t \le 1$. Subsequently, Efron and Thisted [Efron B, Thisted R (1976) *Biometrika* 63(3):435–447] proposed a modification that empirically predicts *U* even for some $t > 1$, but without provable guarantees. We derive a class of estimators that provably predict *U* all of the way up to $t \propto \log n$. We also show that this range is the best possible and that the estimator's mean-square error is near optimal for any *t*. Our approach yields a provable guarantee for the Efron–Thisted estimator and, in addition, a variant with stronger theoretical and experimental performance than existing methodologies on a variety of synthetic and real datasets. The estimators are simple, linear, computationally efficient, and scalable to massive datasets. Their performance guarantees hold uniformly for all distributions, and apply to all four standard sampling models commonly used across various scientific disciplines: multinomial, Poisson, hypergeometric, and Bernoulli product.

species estimation | extrapolation model | nonparametric statistics

Species estimation is an important problem in numerous scientific disciplines. Initially used to estimate ecological diversity (1–4), it was subsequently applied to assess vocabulary size (5, 6), database attribute variation (7), and password innovation (8). Recently, it has found a number of bioscience applications, including estimation of bacterial and microbial diversity (9–12), immune receptor diversity (13), complexity of genomic sequencing (14), and unseen genetic variations (15).

All approaches to the problem incorporate a statistical model, with the most popular being the "extrapolation model" introduced by Fisher, Corbet, and Williams (16) in 1943. It assumes that n independent samples $X^n \triangleq X_1, \ldots, X_n$ were collected from an unknown distribution p, and calls for estimating

$$U \triangleq U\left(X^n, X_{n+1}^{n+m}\right) \triangleq \left|\left\{X_{n+1}^{n+m}\right\} \setminus \left\{X^n\right\}\right|,$$

the number of hitherto unseen symbols that would be observed if m additional samples $X_{n+1}^{n+m} \triangleq X_{n+1}, \ldots, X_{n+m}$ were collected from the same distribution.

In 1956, Good and Toulmin (17) predicted U by a fascinating estimator that has since intrigued statisticians and a broad range of scientists alike (18). For example, in the Stanford University Statistics Department brochure (19), published in the early 1990s and slightly abbreviated here, Bradley Efron credited the problem and its elegant solution with kindling his interest in statistics. As we shall soon see, Efron, along with Ronald Thisted, went on to make significant contributions to this problem.

In the early 1940s, naturalist Corbet had spent 2 y trapping butterflies in Malaya. At the end of that time, he constructed a table (see below) to show how many times he had trapped various butterfly species. For example, 118 species were so rare that Corbet had trapped only one specimen of each, 74 species had been trapped twice each, etc.

Frequency	1	2	3	4	5	...	14	15
Species	118	74	44	24	29	...	12	6

Corbet returned to England with his table, and asked R. A. Fisher, the greatest of all statisticians, how many new species he would see if he returned to Malaya for another 2 y of trapping. This question seems impossible to answer, because it refers to a column of Corbet's table that doesn't exist, the "0" column. Fisher provided an interesting answer that was later improved on [by Good and Toulmin (17)]. The number of new species you can expect to see in 2 y of additional trapping is

$$118 - 74 + 44 - 24 + \cdots - 12 + 6 = 75.$$

This example evaluates the Good–Toulmin estimator for the special case where the original and future samples are of equal size, namely $m = n$. To describe the estimator's general form, we need only a modicum of nomenclature.

Preliminaries

The *prevalence* $\Phi_i \triangleq \Phi_i(X^n)$ of an integer $i \ge 0$ in X^n is the number of symbols appearing i times in X^n. For example, for X^7 = *bananas*, $\Phi_1 = 2$ and $\Phi_2 = \Phi_3 = 1$, and, in Corbet's table, $\Phi_1 = 118$

Significance

Many scientific applications ranging from ecology to genetics use a small sample to estimate the number of distinct elements, known as "species," in a population. Classical results have shown that *n* samples can be used to estimate the number of species that would be observed if the sample size were doubled to 2*n*. We obtain a class of simple algorithms that extend the estimate all the way to *n* log *n* samples, and we show that this is also the largest possible estimation range. Therefore, statistically speaking, the proverbial bird in the hand is worth log *n* in the bush. The proposed estimators outperform existing ones on several synthetic and real datasets collected in various disciplines.

Author contributions: A.O., A.T.S., and Y.W. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence should be addressed. Email: s.theertha@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1607774113/-/DCSupplemental.

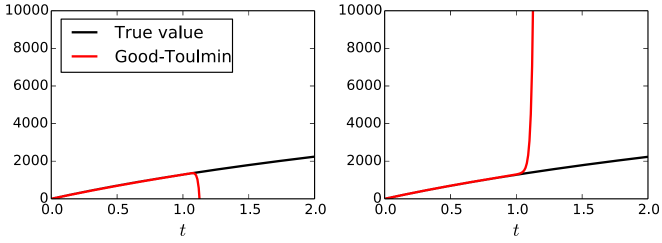

Fig. 1. U^{GT} as a function of t for two random samples of size $n = 5{,}000$ generated by a Zipf distribution $p_i \propto 1/(i+10)$ for $1 \le i \le 10{,}000$.

and $\Phi_2 = 74$. Let $t \triangleq m/n$ be the ratio of the number of future and past samples so that $m = tn$. Good and Toulmin estimated U by the surprisingly simple formula

$$U^{\mathrm{GT}} \triangleq U^{\mathrm{GT}}(X^n, t) \triangleq -\sum_{i=1}^{\infty} (-t)^i \Phi_i. \qquad \mathbf{[1]}$$

They showed that, for all $t \le 1$, U^{GT} is nearly unbiased, and that, although U can be as high as nt,*

$$\mathbb{E}\left(U^{\mathrm{GT}} - U\right)^2 \lesssim nt^2;$$

hence, in expectation, U^{GT} approximates U to within just $\sqrt{n}t$. Fig. 1 shows that, for the ubiquitous Zipf distribution, U^{GT} indeed approximates U well for all $t \le 1$. Naturally, it is desirable to predict U for as large a t as possible. However, as $t > 1$ increases, U^{GT} grows as $(-t)^i \Phi_i$ for the largest i such that $\Phi_i > 0$. Hence, whenever any symbol appears more than once, U^{GT} grows superlinearly in t, eventually far exceeding U that grows at most, linearly in t. Fig. 1 also shows that, for the same Zipf distribution, for $t > 1$, indeed, U^{GT} does not approximate U at all.

To predict U for $t > 1$, Good and Toulmin (17) suggested using the Euler transform (20) that converts an alternating series into another series with the same sum, and heuristically often converges faster. Interestingly, Efron and Thisted (5) showed that, when the Euler transform of U^{GT} is truncated after k terms, it can be expressed as another simple linear estimator,

$$U^{\mathrm{ET}} \triangleq \sum_{i=1}^{n} h_i^{\mathrm{ET}} \cdot \Phi_i, \qquad \mathbf{[2]}$$

where

$$h_i^{\mathrm{ET}} \triangleq -(-t)^i \cdot \mathbb{P}\left(\mathrm{Bin}\left(k, \frac{1}{1+t}\right) \ge i\right),$$

and

$$\mathbb{P}\left(\mathrm{Bin}\left(k, \frac{1}{1+t}\right) \ge i\right) = \begin{cases} \sum_{j=i}^{k} \binom{k}{j} \dfrac{t^{k-j}}{(1+t)^k} & i \le k, \\ 0 & i > k, \end{cases}$$

is the binomial tail probability that decays with i, thereby moderating the rapid growth of $(-t)^i$.

Over the years, U^{ET} has been used by numerous researchers in a variety of scenarios and a multitude of applications. However, despite its widespread use and robust empirical results, no provable guarantees have been established for its performance or that of any related estimator when $t > 1$. The lack of theoretical understanding has also precluded clear guidelines for choosing the parameter k in U^{ET}.

Methodology and Results

We construct a family of estimators that provably predict U optimally not just for constant $t > 1$ but all of the way up to $t \propto \log n$; this shows that, per each observed sample, we can infer properties of $\log n$ yet unseen samples. The proof technique is general and provides a disciplined guideline for choosing the parameter k for U^{ET} as well as a better-performing modification of U^{ET}.

Smoothed Good–Toulmin Estimator. To obtain a new class of estimators, we too start with U^{GT}, but, unlike U^{ET} that was derived from U^{GT} via analytical considerations aimed at improving the convergence rate, we take a probabilistic view that controls the bias and variance of U^{GT} and balances the two to obtain a more efficient estimator.

Note that what renders U^{GT} inaccurate when $t > 1$ is not its bias but its high variance due to the exponential growth of the coefficients $(-t)^i$ in **[1]**; in fact U^{GT} is the unique unbiased estimator for all t and n in the closely related Poisson sampling model (*SI Appendix*, section 1.2). Therefore, it is tempting to truncate the series **[1]** at the ℓ^{th} term and use the partial sum estimator

$$U^{\ell} \triangleq -\sum_{i=1}^{\ell} (-t)^i \Phi_i. \qquad \mathbf{[3]}$$

However, as *Lemma 2* shows, as long as $t > 1$, regardless of the choice of ℓ, there exist certain distributions so that most of the symbols typically appear ℓ times and, hence, the last term in **[3]** dominates, resulting in a large bias and inaccurate estimates.

To resolve this problem, we propose the Smoothed Good–Toulmin (SGT) estimator that truncates **[1]** at an independent random location L and averages over the distribution of L,

$$U^{\mathrm{L}} \triangleq \mathbb{E}_L\left[-\sum_{i=1}^{L} (-t)^i \Phi_i\right].$$

The key insight is that, because the bias of U^{ℓ} typically alternates signs with ℓ, averaging over different cutoff locations can significantly reduce the bias by taking advantage of the cancellation. Furthermore, the SGT estimator can also be expressed simply as a linear combination of prevalences

$$U^{\mathrm{L}} = \mathbb{E}_L\left[-\sum_{i \ge 1} (-t)^i \Phi_i 1_{L \ge i}\right] = -\sum_{i \ge 1} (-t)^i \mathbb{P}(L \ge i) \Phi_i.$$

Choosing different smoothing distributions for L yields different linear estimators, where the tail probability $\mathbb{P}(L \ge i)$ compensates for the exponential growth of $(-t)^i$, thereby stabilizing the variance. Surprisingly, although the motivation and approach are quite different, SGT estimators include U^{ET} in **[2]** as a special case corresponding to binomial smoothing $L \sim \mathrm{Bin}(k, \frac{1}{1+t})$; this provides an intuitive probabilistic interpretation of U^{ET}, originally derived via Euler's transform and analytic considerations. In *Main Results*, we show that this interpretation leads to a theoretical guarantee for U^{ET} as well as improved estimators.

Main Results. Because $0 \le U \le nt$, we evaluate an estimator U^{E} by its worst-case normalized mean-square error (NMSE),

*For positive sequences a_n, b_n, denote $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if for some constant c, $a_n/b_n \le c$ for all $n \ge n_0$. Denote $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

$$\mathcal{E}_{n,t}(U^{\mathrm{E}}) \triangleq \sup_{p} \mathbb{E}_p \left(\frac{U^{\mathrm{E}} - U}{nt}\right)^2.$$

This criterion conservatively evaluates the estimator on the worst possible distribution. The trivial estimator that always predicts $nt/2$ new symbols achieves NMSE 1/4, and we would like to construct estimators with vanishing NMSE that estimate U up to an error that diminishes with n, regardless of the data-generating distribution; in particular, we are interested in the largest t for which this is possible.

Relating the bias and variance of U^{L} to the moment generating function and the exponential generating function of L (see *Theorem 3* and *SI Appendix*, section 2.3), we obtain the following performance guarantee for SGT estimators with appropriately chosen smoothing distributions.

Theorem 1. *For Poisson or binomially distributed L with the parameters given in Table 1, for all* $t \geq 1$ *and* $n \in \mathbb{N}$,

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \lesssim \frac{1}{n^{1/t}}.$$

Theorem 1 provides a principled way for tuning the parameter k for the Efron–Thisted estimator U^{ET} and a provable guarantee for its performance, shown in Table 1. It also shows that a modification of U^{ET} with $q = \frac{2}{t+2}$ enjoys even faster convergence rate and, as empercially demonstrated in *Experiments*, outperforms the original version of U^{ET} as well as other state-of-the-art estimators.

Furthermore, SGT estimators are essentially optimal as witnessed by the following matching minimax lower bound.

Theorem 2. *There exist universal constants* c, c', such that, for all $t \geq c$, $n \in \mathbb{N}$, *and any estimator* U^{E},

$$\mathcal{E}_{n,t}(U^{\mathrm{E}}) \gtrsim \frac{1}{n^{c'/t}}.$$

Theorems 1 and *2* determine the limit of predictability up to a constant factor.

Corollary 1. *For all* $\delta > 0$,

$$\max\{t : \mathcal{E}_{n,t}(U^{\mathrm{E}}) < \delta \textit{ for some } \mathrm{U}^{\mathrm{E}}\} \asymp \frac{\log n}{\log \frac{1}{\delta}}.$$

Concurrent to this work, ref. 21 proposed a linear programming algorithm to estimate U; however, their NMSE is $O(\frac{t}{\log n})$, which is exponentially weaker than the optimal result $O(n^{-1/t})$ in *Theorem 1*. Furthermore, the computational cost far exceeds those of our linear estimators.

The rest of the paper is organized as follows. We first describe the four statistical models commonly used across various scientific disciplines, namely, the multinomial, Poisson, hypergeometric, and Bernoulli product models. Among the four models, Poisson is the simplest to analyze, for which we outline the proof of *Theorem 1*. In *SI Appendix*, we prove similar results for the other three statistical models and also prove the lower bound for the multinomial and Poisson models. Finally, we demonstrate the efficiency and practicality of our estimators on a variety of synthetic and data sets.

Table 1. NMSE of SGT estimators for three smoothing distributions

Smoothing distribution	Parameters	$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \lesssim$
Poisson (r)	$r = \frac{1}{2t}\log_e\frac{n(t+1)^2}{t-1}$	$n^{-1/t}$
Binomial (k, q)	$k = \left\lfloor\frac{1}{2}\log_2\frac{nt^2}{t-1}\right\rfloor$, $q = \frac{1}{t+1}$	$n^{-\log_2(1+1/t)}$
Binomial (k, q)	$k = \left\lfloor\frac{1}{2}\log_3\frac{nt^2}{t-1}\right\rfloor$, $q = \frac{2}{t+2}$	$n^{-\log_3(1+2/t)}$

Because, for any $t \geq 1$, $\log_3(1+2/t) \geq \log_2(1+1/t) \geq 1/t$, binomial smoothing with $q = 2/(2+t)$ yields the best convergence rate.

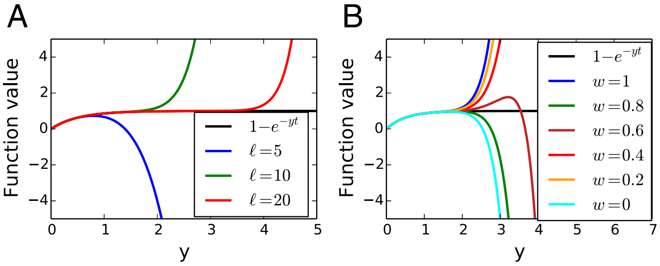

Fig. 2. Approximation of $1 - e^{-2y}$ by (*A*) ℓ-term Taylor approximation and (*B*) averages of 10- and 11-term Taylor approximation.

Statistical Models

The extrapolation paradigm has been applied to several statistical models. In all of them, an initial sample of size related to n is collected, resulting in a set S_{old} of observed symbols. We consider collecting a new sample of size related to m that would result in a yet unknown set S_{new} of observed symbols, and we would like to estimate

$$|S_{\mathrm{new}} \backslash S_{\mathrm{old}}|,$$

the number of unseen symbols that will appear in the new sample. For example, for the observed sample *bananas* and future sample **sonatas**, $S_{\mathrm{old}} = \{\mathtt{a}, \mathtt{b}, \mathtt{n}, \mathtt{s}\}$, $S_{\mathrm{new}} = \{\mathtt{a}, \mathtt{n}, \mathtt{o}, \mathtt{s}, \mathtt{t}\}$, and $|S_{\mathrm{new}} \backslash S_{\mathrm{old}}| = |\{\mathtt{o}, \mathtt{t}\}| = 2$.

Four statistical models have been commonly used in the literature (cf. survey in refs. 3 and 4), and our results apply to all of them. The first three statistical models are also referred to as the abundance models, and the last one is often referred to as the incidence model in ecology (4).

Multinomial. This is Good and Toulmin's original model where the samples are independently and identically distributed (i.i.d.), and the initial and new samples consist of exactly n and m symbols, respectively. Formally, $X^{n+m} = X_1, \ldots, X_{n+m}$ are generated independently according to an unknown discrete distribution of finite or even infinite support, $S_{\mathrm{old}} = \{X^n\}$, and $S_{\mathrm{new}} = \{X_{n+1}^{n+m}\}$.

Hypergeometric. This model corresponds to a sampling-without-replacement variant of the multinomial model. Specifically, X^{n+m} are drawn uniformly without replacement from an unknown collection of symbols that may contain repetitions, for example, an urn with some white and black balls. Again, $S_{\mathrm{old}} = \{X^n\}$ and $S_{\mathrm{new}} = \{X_{n+1}^{n+m}\}$.

Poisson. As in the multinomial model, the samples are also i.i.d., but the sample sizes, instead of being fixed, are Poisson distributed. Formally, $N \sim \mathrm{poi}(n)$, $M \sim \mathrm{poi}(m)$, and X^{N+M} are generated independently according to an unknown discrete distribution, $S_{\mathrm{old}} = \{X^N\}$, and $S_{\mathrm{new}} = \{X_{N+1}^{N+M}\}$.

Bernoulli Product. In this model, we observe signals from a collection of independent processes over subset of an unknown set $\mathcal{X}$. Every $x \in \mathcal{X}$ is associated with an unknown probability $0 \leq p_x \leq 1$, where the probabilities do not necessarily sum to 1. Each sample X_i is a subset of $\mathcal{X}$ where symbol $x \in \mathcal{X}$ appears with probability p_x and is absent with probability $1 - p_x$, independently of all other symbols. $S_{\mathrm{old}} = \cup_{i=1}^{n} X_i$ and $S_{\mathrm{new}} = \cup_{i=n+1}^{n+m} X_i$.

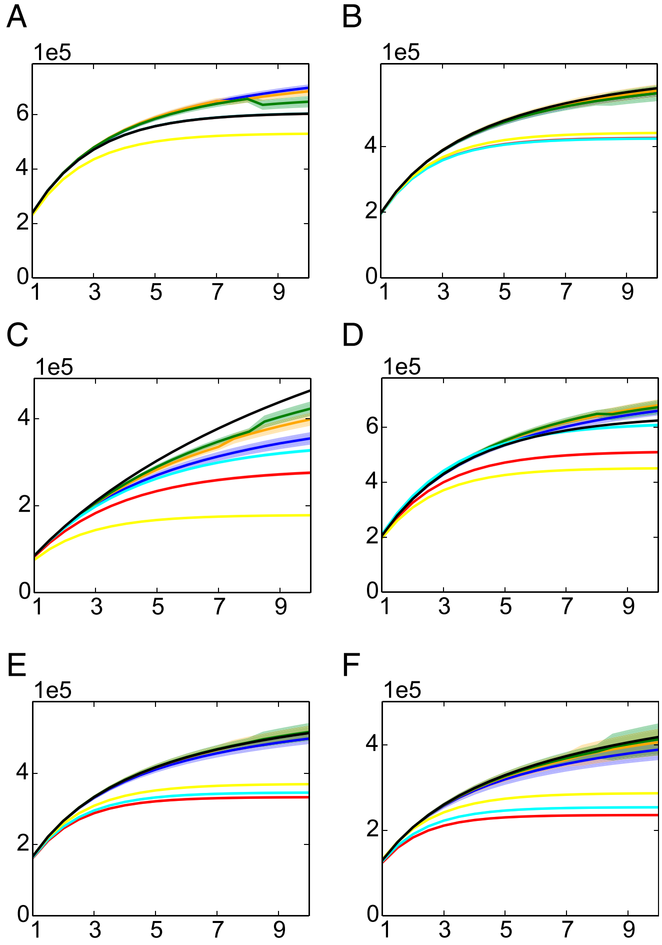

Fig. 3. Comparisons of unseen species estimates as a function of t for six distributions (*A*) uniform, (*B*) distributions with 2 steps $\frac{1}{2k}\times\frac{k}{2}\cup\frac{3}{2k}\times\frac{k}{2}$, (*C*) Zipf distribution with parameter 1 $\left(p_i\propto\frac{1}{i}\right)$, (*D*) Zipf distribution with parameter 1/2 $\left(p_i\propto\frac{1}{i^{1/2}}\right)$, (*E*) Dirichlet-1 prior, (*F*) Dirichlet-1/2 prior. All experiments have distribution support size 10^6, $n=5\cdot 10^5$, and are averaged over 100 iterations. The true value is shown in black, and estimated values are colored, with the solid line representing their means and the shaded band corresponding to one SD. The parameters of the SGT estimators are chosen based on Table 1.

We close this section by discussing two problems that are closely related to the extrapolation model, support size estimation and missing mass estimation that correspond to $m=\infty$ and $m=1$, respectively. The probability that the next sample is new is precisely the expected value of U for $m=1$, which is the goal in the basic Good–Turing problem (22–25). On the other hand, any estimator U^{E} for U can be converted to a (not necessarily good) support size estimator by adding the number of observed symbols. Estimating the support size of an underlying distribution has been studied by both ecologists (1–3) and theoreticians (26–28); however, to make the problem nontrivial, all statistical models impose a lower bound on the minimum nonzero probability of each symbol, which is assumed to be known to the statistician. We discuss the connections and differences to our results in *SI Appendix*, section 5.

Theory

We present the construction of estimators and the analysis for the Poisson model. Extensions to other models are given in *SI Appendix*.

General Linear Estimators. Following ref. 5, we consider linear estimators of the form

$$U^{\mathrm{h}}=\sum_{i=1}^{\infty}\Phi_i\cdot h_i, \qquad [4]$$

where $h_1,h_2,\ldots$ can be identified with a formal power series $h(y)=\sum_{i=1}^{\infty}\frac{h_i y^i}{i!}$. For example, U^{GT} in **[1]** corresponds to the function $h(y)=1-e^{-yt}$. *Lemma 1* (proved in *SI Appendix*, section 2.1) bounds the bias and variance of an arbitrary linear estimator U^{h} using properties of the function h. This result will later be particularized to the SGT estimator. Let $\Phi_+\triangleq\sum_{i=1}^{\infty}\Phi_i$ denote the number of observed symbols.

Lemma 1. *The bias of* U^{h} is

$$\mathbb{E}\left[U^{\mathrm{h}}-U\right]=\sum_x e^{-\lambda_x}\left(h(\lambda_x)-\left(1-e^{-t\lambda_x}\right)\right),$$

where $\lambda_x\triangleq np_x$, and its variance satisfies

$$\mathrm{Var}\left(U^{\mathrm{h}}-U\right)\le\mathbb{E}[\Phi_+]\cdot\sup_{i\ge 1}h_i^2+\mathbb{E}[U].$$

Lemma 1 enables us to reduce the estimation problem to a task on approximating functions. Specifically, the goal is to approximate $1-e^{-yt}$ by a function $h(y)$ all of whose derivatives at zero have small magnitude.

Why Truncated Good–Toulmin Does Not Work. Before we discuss the SGT estimator, we show that the naive approach of truncating the GT estimator described earlier in **[3]** leads to poor prediction when $t>1$. The GT estimator corresponds to the perfect approximation

$$h^{\mathrm{GT}}(y)=1-e^{-yt};$$

however, $\sup_{i\ge 1}\left|h_i^{\mathrm{GT}}\right|=\max\{t,\lim_{m\to\infty}t^m\}$, which is infinity if $t>1$ and leads to large variance. To avoid this situation, a natural approach is to use the ℓ-term Taylor expansion of $1-e^{-yt}$ at 0, namely,

$$h^{\ell}(y)=-\sum_{i=1}^{\ell}\frac{(-yt)^i}{i!}, \qquad [5]$$

which corresponds to the estimator U^{ℓ} defined in **[3]**. Then $\sup_{i\ge 1}\left|h_i^{\ell}\right|=t^{\ell}$, and, by *Lemma 1*, the variance is, at most, $n(t^{\ell}+t)$. Hence if $\ell\le\log_t m$, the variance is, at most, $n(m+t)$. However, note that the ℓ-term Taylor approximation is a degree-ℓ polynomial that eventually diverges and deviates from $1-e^{-yt}$ as y increases, thereby incurring a large bias. Fig. 2*A* illustrates this phenomenon by plotting the function $1-e^{-yt}$ and its Taylor expansion with 5, 10, and 20 terms. Indeed, the next result, proved in *SI Appendix*, establishes the inconsistency of truncated GT estimators.

Lemma 2. *For some constant* $c>0$, for all $\ell\ge 0$, $t>1$, *and* $n\in\mathbb{N}$,

$$\mathcal{E}_{n,t}\left(U^{\ell}\right)\ge\frac{c(t-1)^5}{t^4}.$$

Smoothing by Random Truncation. As we saw in *Why Truncated Good–Toulmin Does Not Work*, the ℓ-term Taylor approximation, where all of the coefficients beyond the ℓ^{th} term are set to zero results in a high bias. Instead, one can choose a weighted average of several Taylor series approximations, whose biases may cancel each other leading to significant bias reduction; this is the main idea of smoothing. As an illustration, in Fig. 2*B*, we plot

$$wh^{10}+(1-w)h^{11}$$

for various value of weight $w\in[0,1]$. Notice that, for instance, $w=0.6$ leads to better approximation of $1-e^{-yt}$ than both h^{10} and h^{11}.

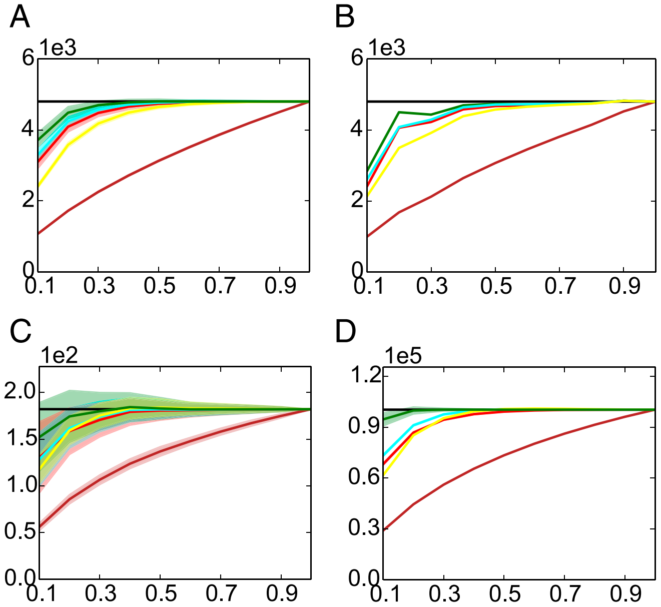

Fig. 4. Estimates for number of (*A*) distinct words in Hamlet with random sampling, (*B*) distinct words in Hamlet with consecutive sampling, (*C*) SLOTUs on human skin, and (*D*) last names, as a function of fraction of seen data.

A natural generalization of the above argument entails taking the weighted average of various Taylor approximations with respect to a given probability distribution over the set of nonnegative integers $\mathbb{Z}_+ \triangleq \{0,1,2,\ldots\}$. For a $\mathbb{Z}_+$-valued random variable L, consider the power series

$$h^{\mathrm{L}}(y) = \sum_{\ell=0}^{\infty} \mathbb{P}(L=\ell)\cdot h^{\ell}(y),$$

where h^{ℓ} is defined in **[5]**. Rearranging terms,

$$h^{\mathrm{L}}(y) = \sum_{\ell=0}^{\infty} \mathbb{P}(L=\ell) \sum_{i=1}^{\ell} \frac{-(-yt)^i}{i!} = -\sum_{i=1}^{\infty} \frac{(-yt)^i}{i!}\mathbb{P}(L\ge i).$$

Thus, the linear estimator with coefficients

$$h_i^{\mathrm{L}} = -(-t)^i \mathbb{P}(L \ge i) \qquad \textbf{[6]}$$

is precisely the SGT estimator U^{L}. Specific choices of smoothing distributions include the following:

$L=\infty$: the original Good–Toulmin estimator **[1]** without smoothing;

$L=\ell$ deterministically: this leads to the estimator U^{ℓ} in **[3]** corresponding to the ℓ-term Taylor approximation; and

$L \sim \mathrm{Bin}(k, 1/(1+t))$: the Efron–Thisted estimator **[2]**, where k is a tuning parameter to be chosen.

Our main results use Poisson and binomial smoothing. To study the performance of the corresponding estimators, we first upper bound the bias and variance for any smoothing distribution. The following key result is proved in *SI Appendix*.

Theorem 3. *For any random variable L over $\mathbb{Z}_+$ and $t \ge 1$,*

$$\mathbb{E}\left[\left(U^{\mathrm{L}}-U\right)^2\right] \le \mathbb{E}[\Phi_+]\cdot\mathbb{E}^2\left[t^L\right] + \mathbb{E}[U] + (\mathbb{E}[\Phi_+]+\mathbb{E}[U])^2 \xi_L(t)^2,$$

where Φ_+ is the number of distinct observed symbols and

$$\xi_L(t) \triangleq \max_{0\le s<\infty} \left|\mathbb{E}\left[\frac{(-s)^L}{L!}\right]\right| e^{-s/t}.$$

We have therefore reduced the problem of controlling the mean-squared loss to that of bounding the moment generating function and the exponential generating function of the smoothing distribution. Applying *Theorem 3* to Poisson smoothing, $L \sim \mathrm{poi}(r)$,

$$\mathbb{E}\left[t^L\right] = e^{-r}\sum_{\ell=0}^{\infty} \frac{(rt)^{\ell}}{\ell!} = e^{r(t-1)}.$$

Furthermore,

$$\mathbb{E}\left[\frac{(-s)^L}{L!}\right] = e^{-r}\sum_{\ell=0}^{\infty} \frac{(-sr)^{\ell}}{(\ell!)^2} = e^{-r} J_0\left(2\sqrt{sr}\right),$$

where J_0 is the Bessel function of the first kind. It is well-known that J_0 takes values in $[-1,1]$ (cf. ref. 20, equation 9.1.60), hence

$$\xi_L(t) \le e^{-r}.$$

Substituting these bounds and optimizing over r yields the upper bound for Poisson smoothing previously announced in Table 1. Results for the binomial smoothing (including the ET estimator) can be obtained using similar but more delicate analysis (*SI Appendix*).

Experiments

We demonstrate the efficacy of our methods by comparing their performance with that of several state-of-the-art support size estimators: Chao–Lee estimator (1, 2), Abundance Coverage Estimator (ACE) (29), and the jackknife estimator (30), all three combined with the Shen–Chao–Lin method (31) of converting support size estimation to unseen species estimation.

We consider both synthetic data generated from various natural distributions and real data. Starting with the former, Fig. 3 shows the species discovery curve, the estimation of U as a function of t for various distributions. Note that the Chao–Lee and ACE estimators are designed specifically for uniform distributions, and, hence, in Fig. 3*A*, they coincide with the true value; however, for all other distributions, SGT estimators have the best overall performance.

Among the proposed estimators, the binomial-smoothing estimator with parameter $q=\frac{2}{2+t}$ has the strongest theoretical guarantee and empirical performance. Hence, for real data experiments, we only plot it to compare with the state of the art. We test the estimators on three real datasets where the samples size n ranges from a few hundreds to a million. For all these datasets, our estimator outperforms the existing procedures.

Corpus Linguistics. Fig. 4*A* shows the first real-data experiment of predicting the vocabulary size based on partial text. Shakespeare's play *Hamlet* consists of $n_{\mathrm{total}} = 31{,}999$ words, of which 4,804 are distinct. We randomly sample n of the n_{total} words without replacement, predict the number of unseen words in the remaining $n_{\mathrm{total}} - n$ ones, and add it to those observed. The results shown are averaged over 100 trials. Observe that the new estimator outperforms existing ones and that merely 20% of the data already yields an accurate estimation of the total number of distinct words. Fig. 4*B* repeats the experiment simply using the first n consecutive words in lieu of random sampling, in which case the SGT estimator also outperforms other schemes in a similar fashion.

Biota Analysis. Fig. 4*C* estimates the number of bacterial species on the human skin. Gao et al. (12) considered the forearm skin biota of six subjects. They identified $n_{\mathrm{total}} = 1{,}221$ clones consisting of 182 different species-level operational taxonomic units (SLOTUs). As before, we select n out of the n_{total} clones without replacement and predict the number of distinct SLOTUs found. Again the SGT estimate is more accurate than those of existing estimators and is reasonably accurate already with 20% of the data.

Census Data. Finally, Fig. 4*D* considers the 2000 United States Census (32), which lists all US last names corresponding to at least 100 individuals. With these many repetitions, even a small fraction of the data will contain almost all names. To make the estimation task nontrivial, we first subsample the data $n_{\text{total}} = 10^6$ and obtain a list of 100,328 distinct last names. As before, we estimate this number using n randomly sampled names, and the SGT estimator yields significantly more accurate estimations than the state of the art.

Observations. As argued in ref. 33, it is often useful for species estimators to be monotone and concave in the extrapolation ratio t, which, however, need not be satisfied by linear estimators such as Good−Toulmin or SGT estimators. In *SI Appendix*, section 6, we propose a simple modification of the SGT estimator that is both monotone and concave, which retains the good empirical performance of the original estimator.

ACKNOWLEDGMENTS. We thank Dimitris Achlioptas, David Tse, Chi-Hong Tseng, and Jinye Zhang for helpful discussions and comments. This work was partially completed while the authors were visiting the Simons Institute for the Theory of Computing at University of California, Berkeley, whose support is gratefully acknowledged. The research of Y.W. has been supported in part by National Science Foundation (NSF) Grants IIS-14-47879 and CCF-15-27105, and the research of A.O. and A.T.S. has been supported in part by NSF Grants CCF15-64355 and CCF-16-19448.

1. Chao A (1984) Nonparametric estimation of the number of classes in a population. *Scand J Stat* 11:256–270.
2. Chao A, Lee SM (1992) Estimating the number of classes via sample coverage. *J Am Stat Assoc* 87(417):210–217.
3. Bunge J, Fitzpatrick M (1993) Estimating the number of species: A review. *J Am Stat Assoc* 88(421):364–373.
4. Colwell RK, et al. (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J Plant Ecol* 5(1):3–21.
5. Efron B, Thisted R (1976) Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* 63(3):435–447.
6. Thisted R, Efron B (1987) Did Shakespeare write a newly-discovered poem? *Biometrika* 74(3):445–455.
7. Haas PJ, Naughton JF, Seshadri S, Stokes L (1995) Sampling-based estimation of the number of distinct values of an attribute. *Proceedings of the 21st VLDB Conference* (Morgan Kaufmann, Burlington, MA), pp 311–322.
8. Florencio D, Herley C (2007) A large-scale study of web password habits. *Proceedings of the 16th International Conference on World Wide Web* (Assoc Comput Machinery, New York), pp 657–666.
9. Kroes I, Lepp PW, Relman DA (1999) Bacterial diversity within the human subgingival crevice. *Proc Natl Acad Sci USA* 96(25):14547–14552.
10. Paster BJ, et al. (2001) Bacterial diversity in human subgingival plaque. *J Bacteriol* 183(12):3770–3783.
11. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ (2001) Counting the uncountable: Statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* 67(10):4399–4406.
12. Gao Z, Tseng CH, Pei Z, Blaser MJ (2007) Molecular analysis of human forearm superficial skin bacterial biota. *Proc Natl Acad Sci USA* 104(8):2927–2932.
13. Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor β−chain diversity in $\alpha\beta$ T cells. *Blood* 114(19):4099–4107.
14. Daley T, Smith AD (2013) Predicting the molecular complexity of sequencing libraries. *Nat Methods* 10(4):325–327.
15. Ionita-Laza I, Lange C, M Laird N (2009) Estimating the number of unseen variants in the human genome. *Proc Natl Acad Sci USA* 106(13):5008–5013.
16. Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *J Anim Ecol* 12(1):42–58.
17. Good I, Toulmin G (1956) The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* 43(1-2):45–63.
18. Kolata G (1986) Shakespeare's New Poem: An Ode to Statistics: Two statisticians are using a powerful method to determine whether Shakespeare could have written the newly discovered poem that has been attributed to him. *Science* 231(4736):335–336.
19. Efron B (1992) Excerpt from Stanford statistics department brochure. Available at https://statistics.stanford.edu/sites/default/files/1992_StanfordStatisticsBrochure.pdf. Accessed October 24, 2016.
20. Abramowitz M, Stegun IA (1964) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Wiley, New York).
21. Valiant G, Valiant P (2015) Instance optimal learning. arXiv:1504.05321.
22. Good IJ (1953) The population frequencies of species and the estimation of population parameters. *Biometrika* 40(3-4):237–264.
23. McAllester DA, Schapire RE (2000) On the convergence rate of Good-Turing estimators. *COLT '00 Proceedings of the 13th Conference on Learning Theory* (Morgan Kaufmann, Burlington, MA), pp 1–6.
24. Bickel PJ, Yahav JA (1986) On estimating the total probability of the unobserved outcomes of an experiment. *Lecture Notes–Monograph Series*, ed Van Ryzin J (Inst Math Stat, Beachwood, OH), Vol 8, pp 332–337.
25. Orlitsky A, Suresh AT (2015) Competitive distribution estimation: Why is good-turing good? *Advances in Neural Information Processing Systems 28*, eds Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (NIPS, La Jolla, CA), pp 2143–2151.
26. Raskhodnikova S, Ron D, Shpilka A, Smith A (2009) Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM J Comput* 39(3):813–842.
27. Valiant G, Valiant P (2011) Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing* (Assoc Comput Machinery, New York), pp 685–694.
28. Wu Y, Yang P (2015) Minimax rates of entropy estimation on large alphabets via best polynomial approximation. arxiv:1407.0381.
29. Chao A (2005) Species estimation and applications. Encyclopedia of Statistical Sciences, eds Balakrishnan N, Read CB, Vidakovic B (Wiley, New York), Vol 12, pp 7907–7916.
30. Smith EP, van Belle G (1984) Nonparametric estimation of species richness. *Biometrics* 40(1):119–129.
31. Shen TJ, Chao A, Lin CF (2003) Predicting the number of new species in further taxonomic sampling. *Ecology* 84(3):798–804.
32. US Census Bureau (2000) *Frequently Occuring Surnames from the Census 2000* (US Census Bureau, Washington, DC).
33. Boneh S, Boneh A, Caron RJ (1998) Estimating the prediction function and the number of unseen species in sampling with replacement. *J Am Stat Assoc* 93(441):372–379.