Estimating the unseen from multiple populations

Aditi Raghunathan [1] **Gregory Valiant** [1] **James Zou** [1 2]

Abstract

Given samples from a distribution, how many new elements should we expect to find if we continue sampling this distribution? This is an important and actively studied problem, with many applications ranging from unseen species estimation to genomics. We generalize this extrapolation and related unseen estimation problems to the multiple population setting, where population j has an unknown distribution D_j from which we observe n_j samples. We derive an optimal estimator for the total number of elements we expect to find among new samples across the populations. Surprisingly, we prove that our estimator's accuracy is independent of the number of populations. We also develop an efficient optimization algorithm to solve the more general problem of estimating multi-population frequency distributions. We validate our methods and theory through extensive experiments. Finally, on a real dataset of human genomes across multiple ancestries, we demonstrate how our approach for unseen estimation can enable cohort designs that can discover interesting mutations with greater efficiency.

1. Introduction

Given samples from a distribution, many settings in machine learning and statistics involves estimating properties of the *unseen* portion of the distribution, i.e. elements in the support of the distribution that are not observed in the samples collected so far. One important example of estimating the unseen is the problem of predicting the number of distinct new elements in additional samples collected. This question is famously illustrated by the case of Corbet's butterflies. Alexander Corbet was a British naturalist who spent two years in Malaya trapping butterflies. He found 118 rare species of butterflies for which he found only one specimen, another 74 species with two specimens, 44 with three specimens, etc. Corbet was naturally interested in the butterflies that are heretofore unseen. In particular, he wanted to estimate how many distinct *new* species of butterflies he can expect to discover if he were to conduct a new expedition to Malaya—such an estimate could help determine whether a new experiment is warranted. Good-Toulmin, extending earlier work of Ronald Fisher, came up with the remarkable estimate that the number of new species Corbet can expect to find is simply the alternating sum 118 - 74 + 44 - ... The Good-Toulmin estimator sparked the investigation into how to estimate the discovery rate of new elements and this remains an active area of research. Estimating the discovery rate has many important applications beyond the original species collection setting. In genomics, for example, an important question is: given the genetic variation already identified in the genomes of individuals from some population (say, East Asia), how many additional mutations do we expect to find by sequencing the genomes of additional individuals from East Asia. An accurate answer to this question can improve the cohort design of new population sequencing experiments.

Predicting the number of new elements is a particular instance of estimating the unseen. In other applications, one may want to estimate different statistics that also depend on the currently unobserved elements. For example, one may want to predict how many new elements will be observed at least twice (for reproducibility) or at most three times (if the focus is on rare elements). More generally, one may want to estimate the histogram of the underlying distribution, which summarizes the frequency distribution of all the elements (see Sec. 2 for precise definition) and from which these other statistics can be derived.

The unseen estimation literature has focused on the setting where there is a *single* distribution which generate current samples as well as any future samples. In practice, we often have *multiple* distinct distributions and we observe varying number of samples from each distribution. In the genomics example above, in addition to sequencing data from East Asians, we also have genome sequences of individuals from Europe, Africa, etc. The relevant question is: given we currently have the genomes of n_i individuals

[1]Stanford University, Stanford, CA [2]Chan Zuckerberg Biohub, San Francisco, CA. Correspondence to: Aditi Raghunathan <aditir@stanford.edu>, Gregory Valiant <valiant@stanford.edu>, James Zou <jamesz@stanford.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

from population i, $i \in \{1, ..., m\}$, and we have identified all the genetic variants in this group, how many total new mutations do we expect to find if we sequence additional b_i individuals from population i. Moreover, given a finite budget N_{new} of new genomes that we can sequence, how should we allocate this budget across the different populations to maximize the expected number of new mutations oberved? Similarly, suppose Corbet had also collected butterflies in Brunei and Indonesia, in addition to Malaya. Then he might want to know how many totally new species he can expect to find if he was to spend, say, another six month in Malaya and one year in Brunei. He might also be interested in estimating the joint frequency distribution of butterflies across all three regions.

Our contributions. In this paper, we address the general problem of estimating the unseen when we have samples from multiple populations, each corresponding to a potentially distinct distribution. Despite being very natural, this multi-population problem has not been systematically studied to the best of our knowledge. We derive a multi-population generalization of the Good-Toulmin estimator for the expected number of new elements. Surprisingly, we prove that the accuracy of our extrapolation estimator is independent of the number of populations. Moreover, it achieves the optimal super-linear extrapolation rate. Next, we develop an efficient optimization method to estimate the more general multi-population joint frequency distribution. This complements our extrapolation estimator, and outperforms the generalized Good-Toulmin estimator in most settings. This more general approach also enables predictions for other statistics of interest. We systematically validate these two algorithms on synthetic data as well as real datasets from population genetics and from English books. Moreover, we illustrate that by estimating the joint frequency distribution, we can significantly improve the discovery power under a budget constraint.

2. Related works

The problem of estimating the properties of the unobserved portion of a distribution, given n samples, and the related problem of estimating the number of new domain elements that are likely to be observed if an additional cn samples are collected, dates back to works of I.J. Good and A. Turing (Good, 1953), and R.A. Fisher (Fisher et al., 1943). This was quickly followed by (Good & Toulmin, 1956), which introduced the Good-Toulmin estimator. While the Good-Toulmin estimator is always unbiased, the variance increases rapidly for $c \geq 1$. Subsequent works, including (Efron & Thisted, 1976) have suggested "smoothing" approaches that tradeoff the bias and variance for this type of approach. The recent work of Orlitsky et al. (2016) describes a clever variant that achieves good performance for $c = O(\log n)$. This ability to accurately estimate the number of domain elements seen in a second sample of size up to $O(n \log n)$, where n denotes the size of the original sample, was concurrently shown via a different approach in (Valiant & Valiant, 2016). This logarithmic factor extrapolation matches the lower bounds of (Valiant & Valiant, 2011), to constant factors. The linear estimators that we propose in Section 4 for the multiple population setting, and their analysis, are extensions of the smoothed Good-Toulmin estimators of (Orlitsky et al., 2016).

A different approach to this problem was proposed by Efron & Thisted (1976), who considered a linear-programming approach to estimating this property by implicitly finding a label-less representation of the underlying distribution that was consistent with the observed frequency counts, then returning the support size of this distribution. This approach was adapted and rigorously analyzed in (Valiant & Valiant, 2011; 2013), who showed that it provably yields an accurate representation of the frequency distribution of the underlying distribution, which can subsequently be leveraged to yield estimates of distributional properties, including entropy, distance metrics between distributions, and approximations for the number of new elements that would be observed in larger samples. Recent works (Valiant & Valiant, 2016; Zou et al., 2016) also established that this approach can accurately estimate the number of new elements that will be observed in samples of size up to $O(n \log n)$. Our optimization-based algorithm, described in Section 5, generalizes this approach.

3. Definitions and examples

Let Ω denote the domain, and $D_1, ..., D_m$ denote m probability distributions over Ω. D_i represents the frequency of elements in population i. Note that it is not restrictive to assume that the populations share the same domain Ω since different D_i's may have distinct supports. We model the multi-population unseen estimation as a two stage process. In the first period, we observe n_j independent samples from the j-th population, $\{X_i^j\}_{i=1,...,n_j}^{j=1,...,m}$. This is the *seen* data. In period two, which is in the future, we will sample additional $t_j n_j$ samples from the j-th population, $\{Y_i^j\}_{i=1,...,t_j n_j}^{j=1,...,m}$. The period two samples are *unseen* and we would like to estimate some statistic $U(\{Y_i^j\}, \{X_i^j\})$. We can think of $t_j \geq 0$ as the extrapolation factors. If t_j is large, then we will obtain many more samples from population j in the second period compared to what we have, and the problem of estimating U could be more challenging. We can take t_j as given for the purpose of estimating U. We later discuss how we to leverage our estimator of U to optimize the t_j's in order to maximize the number of new discoveries. Note that in general, the n_j's and t_j's can differ arbitrarily across the populations.

A particularly important statistic is U = *the total number of new elements in* $\{Y_i^j\}$ *that are not observed in the period one samples* $\{X_i^j\}$. A good estimator for this U quantifies the expected information gain of the second period. In the one population setting, this statistic is the focus of Good-Toulmin and a large number of papers. Other useful choices of U could be the number of distinct new elements that are observed at least twice in $\{Y_i^j\}$, which could be relevant if we want some reproducibility.

Beyond estimating these single parameters, we could also hope to use the samples $\{X_i^j\}$ to estimate the *histogram* of $D_1, ..., D_m$. The multi-population *histogram*, defined below, captures all of the information about the populations, other than the labels of the domain.

Definition 3.1 (Multi-population histogram)**.** *Given a collection of m distributions $D_1, \ldots, D_m$ over a common domain Ω, the corresponding multi-population histogram H is a mapping from $[0,1]^m \setminus 0^m \mapsto \mathbb{N} \cup \{0\}$. For each $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots \alpha_m) \in [0,1]^m \setminus 0^m, H(\boldsymbol{\alpha}) = |\{y \in \Omega \mid D_j(y) = \alpha_j, 1 \le j \le m\}|$, where $D_i(y)$ is the probability mass of domain element y. in the ith distribution D_i.*

Any *symmetric* multi-population statistic—one that is invariant to permuting the labels of the domain—is a function of only the histogram. Such statistics include distance metrics between the distributions/populations, measures of the entropy of the populations, and the number of new elements that one is likely to observe in a second batch of samples. The multi-population histogram is also of intrinsic interest; in population genetics, H is exactly the joint frequency distribution of mutations, and reveals information about demographic history (e.g. historical variations in population size) and selective pressures. One benefit of focusing on the histogram is that, while it does not contain as much information as the actual labeled distributions, it can often be accurately recovered even when given too few samples to learn the (labeled) distributions to any significant accuracy (Valiant & Valiant, 2011).

Both for directly predicting U and estimating H, we rely on a label-less representation of the samples, termed the *fingerprint* of $\{X_i^j\}$. The fingerprint of the samples is the analog of the histogram of the distributions, and captures all the information of $\{X_i^j\}$ that is relevant for estimating symmetric statistics.

Definition 3.2 (Multi-population fingerprint)**.** *Given the samples $\{X_i^j\}$, its fingerprint is an m-dimensional tensor Φ whose $i_1 ... i_m$-th entry, $\phi_{i_1 \ldots i_m}$, is the number of distinct elements observed exactly i_j times in the samples from population j. Here each i_j can range from 0 to n_j.*

Example 3.1. *Suppose we have five samples from Population 1, (A, B, C, E, F), and seven from Population 2, (A, B, D, E, E, F, F). The corresponding 2-dimensional fingerprint of this data is given by the following matrix:*

	0	1	2
0	·	1	0
1	1	2	2

The $(1,1)$ entry is 2 because A, B are observed once in each set of samples; the $(1,0)$ entry is 1 because exactly one element, C, is observed once in the samples from Population 1 and zero times in the samples from Population 2. By convention, we omit the $(0,0)$ element.

4. A linear estimator

Unbiased estimator. Given the empirical fingerprints Φ and the extrapolation factors $t_j, j = 1, ..., m$, we define the following estimator

$$\hat{U} = - \sum_{i_1, \ldots, i_m : \sum i_j > 0} \left(\prod_{j=1}^{m} (-t_j)^{i_j} \right) \phi_{i_1 \ldots i_m}. \qquad (1)$$

$\hat{U}$ is a weighted alternating sum of the empirical fingerprints where the weights are determined by the extrapolation factors t_j.

Proposition 4.1. *For any number of populations m, and any extrapolation factors $t_j \ge 0, j = 1, ...m$, $\hat{U}$ is an unbiased estimator of U.*

Proof of the proposition appears in Appendix 7.

$\hat{U}$ is linear in the fingerprint entries. Its computational cost is linear in the total number of period one samples, $n = \sum_j n_j$, since there can be at most n non-zero fingerprint entries. To build more intuition for $\hat{U}$, we illustrate its application in two simple settings.

Example 4.2. *Consider the setting where all m distribution are identical, i.e. all the samples are drawn from the same discrete distribution D. Let $t_j = 1, \forall j$ for simplicity. After rearranging terms, $\hat{U}$ can be written as*

$$\hat{U} = \sum_{k=1} (-1)^{k+1} \left(\sum_{(i_1, \ldots, i_m) : \sum i_j = k} \phi_{i_1 \ldots i_m} \right).$$

Because the populations are identical, the sum in the parenthesis is just the number of elements that are observed k times from all the samples so far. Hence the general estimator $\hat{U}$ reduces to the one dimensional Good-Toulmin estimator when all m populations are identical.

Example 4.3. *Suppose the supports of the distributions D_i are disjoint. Then the only possible non-zero fingerprint entries are $\phi_{i_1 \ldots i_m}$ where exactly one of the i_j is great than 0 and all the other i_j's are zero. For simplicity, assume $t_j = 1$ for all j. Then $\hat{U} = \sum_{j=1}^{k} \sum_i (-1)^{k+1} \phi_i^k$, where ϕ_i^k is the marginal fingerprint entry of the number of elements that are observed i times in population k. Hence*

when the populations are disjoint, the expected number of new elements is the sum of the expected number of new elements in each population. When the populations have overlapping support, we have the nontrivial interaction terms due to the cross-population fingerprint entries.

General weighted linear estimator. While $\hat{U}$ is unbiased, its variance could be large if some of the extrapolation factors t_j's are greater than 1. This is because the powers of t_j appear in Eqn. 1. To address this issue, we introduce a general class of multi-population weighted linear estimators.

$$\hat{U}^W = -\sum_{i_j:\sum i_j>0} \left(\prod_{j=1}^{m}(-t_j)^{i_j}\right) \phi_{i_1,\ldots,i_m} W(i_1,\ldots,i_m).$$

We focus on a particular weighting scheme, which is an extension of that introduced in (Orlitsky et al., 2016): $W(i_1, i_2, \ldots i_m) = \mathbb{P}\left(L \geq \sum_{j\in A} i_j\right)$ where $L \sim \text{Poi}(r)$ and $A = \{j : t_j > 1\}$ are the populations that we would like to extrapolate beyond the original sample size. If $t_j \leq 1\ \forall j$, then W = 1 and $\hat{U}^W$ is just the unbiased estimator $\hat{U}$. The Poisson rate r is a tuning parameter that determines the bias/variance tradeoff of $\hat{U}^W$. As r increases, all the weights approaches 1 and $\hat{U}^W$ approaches the unbiased estimator $\hat{U}$. As r decreases, the fingerprint entries $\phi_{i_1\ldots i_m}$ with some large i_j's—which are also the terms with high variance—are weighted by a factor that is close to 0. This reduces the total variance of $\hat{U}^W$ at the cost of introducing bias. We will see how to set r as a function of the n_j's and t_j's in order to minimize the overall estimation error. In the rest of the paper, unless otherwise specified, we will use $\hat{U}^W$ to denote the multi-population linear estimator with Poisson weights.

Performance guarantee of the weighted estimator. We use relative mean squared error, $\mathbb{E}\left[\left(\frac{\hat{U}^W-U}{\sum n_j t_j}\right)^2\right]$, to quantify the performance of $\hat{U}^W$. This is a natural error metric, because $\sum n_j t_j$ is the number of samples in period two and we care about how the error in the predicted number of new elements scales with the number of samples. Without loss of generality, we can relabel the populations so that $t_1 = max_j t_j$. We are especially interested in the setting when $t_1 \geq 1$ (i.e. large extrapolation).

Proposition 4.4. *Suppose $t_1 = \max_j t_j \geq 1$ and the Poisson rate is $r = \frac{\log(\sum_j n_j(t_j+1))}{2t_1}$, then*

$$\mathbb{E}\left[\left(\frac{\hat{U}^W - U}{\sum n_j t_j}\right)^2\right] \leq \left(\frac{n_1 t_1 + \sum_j n_j}{n_1 t_1}\right) n_1^{-1/t_1}. \quad (2)$$

Remark 4.5 (log extrapolation factor)**.** *Suppose the ratio $\frac{n_1}{\sum_j n_j}$ is bounded, then Prop. 4.4 guarantees that for any $\epsilon > 0$, we can achieve $\mathbb{E}\left[\left(\frac{\hat{U}^W-U}{\sum n_j t_j}\right)^2\right] \leq \epsilon$ with $t_1 = O(\log n_1/\log(1/\epsilon))$. This means that $\hat{U}^W$ has low relative error even when the largest extrapolation factor t_1 is logarithmic in its initial sample size n_1.*

Remark 4.6 (no dependence on m)**.** *Note that the relative error in Eqn. 2* ***does not depend on the number of populations*** *m. This is somewhat surprising since the number of terms in $\hat{U}^W$ potentially grows exponentially with m and the variance of each fingerprint entry $\phi_{i_1\ldots i_m}$ also increases as the number of population increases. This population agnostic property of $\hat{U}^W$ guarantees its accuracy even when m is arbitrarily large.*

Remark 4.7 (lower bound)**.** *Here we have focused on a specific form of the estimator $\hat{U}^W$ where the weights W of the fingerprint entries correspond to the tail probability of Poisson distributions. A natural question is whether there exists a different form of the weights or a different estimator altogether that can consistently be more accurate than our current $\hat{U}^W$. The answer is essentially no due to the following lower bound for one population extrapolation (Orlitsky et al., 2016; Valiant & Valiant, 2011): There exists universal constants c, c' such that for all estimators $\hat{U}$, if the extrapolation factor $t > c$, then $\exists$ distribution such that $\mathbb{E}\left[\left(\frac{\hat{U}-U}{nt}\right)^2\right] \gtrsim n^{-c'/t}$. Here n is the number of samples drawn from this distribution in period one. This lower bound implies that in order to guarantee that the relative error is less than ϵ in general, the extrapolation factor can be at most $O(\log n/\log(1/\epsilon))$, matching Prop. 4.4.*

Outline of the proof of Prop. 4.4 (detailed analysis is in the Appendix). To analyze the relative error, we separately quantify the bias and variance of $\hat{U}^W$ in terms of n_j, t_j, r.

Lemma 4.8 (Bias)**.** *Let r denote the rate of the Poisson weights, then*

$$\left|\mathbb{E}[\hat{U}^W - U]\right| \leq \left(\sum_{j\in A} n_j(t_j+1)\right) e^{-r}$$

Lemma 4.9 (Variance)**.** *Without loss of generality, let $t_1 = \max_j t_j$ and suppose $t_1 \geq 1$ then*

$$Var(\hat{U}^W - U) \leq \sum n_j e^{2r(t_1-1)} + \sum_j n_j t_j.$$

To obtain the optimal r given in the statement of Prop. 4.4, we set r to balance the squared bias and variance.

5. Estimating the multi-population frequency distribution

While we have a linear estimator for the number of unseen elements in a new sample, it is challenging to con-

struct good estimators of other statistics (e.g. number of new elements observed ≥ 2) directly from the fingerprints. As discussed in Sec. 3, we can also take the less direct approach of first trying to estimating the true underlying multi-population histogram. Given an accurate reconstruction of this underlying histogram, we can then estimate any symmetric statistic of the future samples. We discuss some of the uses of such a representation in Section 5.

Recovering the frequency distribution The core of our algorithm to recover the multi-population histogram is a natural extension of the single population algorithm presented in Valiant & Valiant (2011; 2013).

Estimating the multi-population histogram: Core Approach.
Input: Multi-population fingerprint Φ of samples,
Output: Two estimates, $\hat{H}_{counts}$ and $\hat{H}_{ll}$ of histogram corresponding to the distributions underlying fingerprint Φ.

- Compute $\hat{H}_{counts}$ and $\hat{H}_{ll}$ minimizing the following expressions:

$$\hat{H}_{counts} = \arg\min_{H} \sum_{\boldsymbol{i}} \frac{1}{\sqrt{1+\Phi_{\boldsymbol{i}}}} |\Phi_{\boldsymbol{i}} - \hat{\Phi}(H)_{\boldsymbol{i}}|.$$

$$\hat{H}_{ll} = \arg\max_{H} \sum_{\boldsymbol{i}} \log poi(\Phi_{\boldsymbol{i}}, \hat{\Phi}(H)_{\boldsymbol{i}}),$$

$$\text{Where } [\hat{\Phi}(H)]_{\boldsymbol{i}} = \sum_{\boldsymbol{\alpha}} H(\alpha) \prod_{j=1}^{m} bino(\alpha_j, n_j, i_j).$$

The intuition behind these two optimization problems is the following. The histogram corresponding to a set of distributions is an *unlabeled* representation of the underlying distributions, hence it makes intuitive sense to try to recover the histogram that maximizes the likelihood of the *unlabeled* representation of the samples, namely the fingerprint Φ. Recent work (Acharya et al., 2016) provided rigorous support for this intuition. In general, however, this likelihood might be difficulty to compute. Nevertheless, an efficiently computable proxy for this likelihood can be obtained by treating the distribution of the fingerprint, corresponding to a histogram H, as a product distribution, with $\Phi_{i_1,\ldots,i_m}$ distributed according to the Poisson distribution with appropriate expectation $E_H[\Phi_{i_1,\ldots,i_m}]$. The recent central limit theorem for "Poisson Multinomials" from (Valiant & Valiant, 2011) provides at least some corroboration for the reasonableness of having a proxy for the log-likelihood that decomposes linearly across the different elements of Φ. The motivation for the $\frac{1}{\sqrt{1+\Phi_{\boldsymbol{i}}}}$ scaling on the first proxy likelihood function is that this expression penalizes discrepancies between the observed and expected fingerprint entries according to a rough approximation of the standard deviation of that fingerprint entry, as the variance of a Poisson random variable is equal to its expectation, and the observed fingerprint entry is an approximation for the expected fingerprint entry given the true underlying histogram.

The work (Valiant & Valiant, 2013) focused on recovering $\hat{H}_{counts}$, as this optimization problem can be formulated as a linear program, whose variables correspond to a fine discretization of the potential support of the histogram. Unfortunately, in the present multi-distribution setting, the number of variables required by this linear programming approach would scale exponentially with the number of distributions in question. Even for fingerprints derived from modest-sized samples from two distributions, the resulting linear program becomes impractical.

Instead of pursuing the linear programming based approach, we instead propose a black-box optimization approach to finding a histogram that optimizes either of the two proxy likelihood functions. In this optimization approach, the dimensionality of the optimization problem is specified by the user, and corresponds to the number of $(i_1, \ldots, i_m)$ tuples for which the returned histogram $\hat{H}$ is nonzero. Denoting this quantity by s, the resulting optimization problem can be regarded as the problem of specifying s vectors $(h_1, \alpha_{1,1}, \ldots, \alpha_{1,m}), \ldots, (h_s, \alpha_{s,1}, \ldots, \alpha_{s,m})$. These s vectors are then interpreted as a histogram H with $H(\boldsymbol{\alpha}_j) = h_j$ for all $j \in \{1, \ldots, s\}$, and $H(\boldsymbol{\alpha}) = 0$ for all other vectors $\boldsymbol{\alpha}$.

The one additional modification that leads to a substantial improvement in runtime is to only evaluate the proxy likelihood expressions for fingerprint entries $\Phi_{i_1,\ldots,i_m} \geq 2$. The intuition for this is two-fold. First, the number of vectors $(i_1, \ldots, i_m)$ for which $\Phi_{i_1,\ldots,i_m} = 0$ will scale exponentially with m, as opposed to scaling as some parameter of the sample sizes; this is clearly undesirable. Second, given that we wish to avoid evaluating the contribution to the proxy likelihood from fingerprint entries that are zero, we must now be careful in dealing with fingerprint entries that are equal to 1. Suppose we have 1 element with true probability $\frac{i}{n}$ and suppose we observe that fingerprint entry $\Phi_i = 1$, and the other fingerprints near i are 0. Since we are maximizing the likelihood that $\Phi_i = 1$ (without taking into account the nearby 0 entries), we would assign roughly $\sqrt{i}$ elements to probability $\frac{i}{n}$ which is undesirable. Removing the ones largely resolves this issue. Note that the $\Phi_j = 2$ entries do not cause as much of an issue, as such collisions are unlikely to occur in regions of the fingerprint in which there is not a significant number of domain elements.

In this one-distribution example, a constraint on the total probability mass being 1 would resolve this issue, though

analogs of this issue in the multiple distribution setting cannot be resolved in this way. Hence, we adopt the crude, but effective approach of viewing all the empirical fingerprint entries that are equal to 1 as being reflective of an element in the underlying set of distributions whose probability is close to the empirical probability of the corresponding element. We summarize the complete algorithm below:

Estimating the multi-population histogram: Full Algorithm.
Input: Multi-population fingerprint Φ derived from samples from m distributions of respective sizes $n_1, \ldots, n_m$.
Output: Two estimates, $\hat{H}_{counts}$ and $\hat{H}_{ll}$ of histogram corresponding to the distributions underlying fingerprint Φ.

- Remove fingerprint entries that are 1, and add to empirical portion of histogram:
  1. Initialize m-distribution histogram $\hat{H}_{emp}$ to be identically zero.
  2. For each vector $\boldsymbol{i} = (i_1, \ldots, i_m)$ such that $\Phi_{\boldsymbol{i}} = 1$, set $\hat{H}_{emp}(\frac{i_1}{n_1}, \ldots, \frac{i_m}{n_m}) = 1$.
- Compute $\hat{H}_{counts}$ and $\hat{H}_{ll}$ minimizing the following expressions:

$$\hat{H}_{counts} = \arg\min_{H} \sum_{\boldsymbol{i}:\Phi(\boldsymbol{i})\geq 2} \frac{1}{\sqrt{1+\Phi_{\boldsymbol{i}}}} |\Phi_{\boldsymbol{i}} - \hat{\Phi}(H)_{\boldsymbol{i}}|.$$

$$\hat{H}_{ll} = \arg\max_{H} \sum_{\boldsymbol{i}:\Phi(\boldsymbol{i})\geq 2} \log poi(\Phi_{\boldsymbol{i}}, \hat{\Phi}(H)_{\boldsymbol{i}}.$$

Where $\hat{\Phi}(H)_{\boldsymbol{i}} = \sum_{\boldsymbol{\alpha}} H(\alpha) \prod_{j=1}^{m} bino(\alpha_j, n_j, i_j)$.

Subject to the constraint that, together with $\hat{H}_{emp}$, the total mass in all the distributions is 1. Namely for all $i \in \{1, \ldots, m\}$,

$$\sum_{\boldsymbol{\alpha}} \alpha_i \hat{H}_{ll}(\boldsymbol{\alpha}) + \sum_{\boldsymbol{\alpha}} \alpha_i \hat{H}_{*}(\boldsymbol{\alpha}) = 1.$$

- Return the concatenation of the empirical portion of the histogram and the portion returned by the optimization: $\hat{H}_{count} := \hat{H}_{count} + \hat{H}_{emp}$, and $\hat{H}_{ll} := \hat{H}_{ll} + \hat{H}_{emp}$

Leveraging $\hat{H}$ for approximating the value of additional data. An accurate representation of the histogram corresponding to the multi-population distribution underlying a given set of observations can be leveraged to estimate a number of useful properties. These properties include estimating the number of new domain elements that would likely be seen given additional samples from the populations. Specifically, given a histogram $\hat{H}$, corresponding to m populations, we can estimate the expected number of distinct elements that will be observed in samples from the m populations of respective sizes $n_1, \ldots, n_m$ via the simple formula:

$$E[\text{num observed}] = \sum_{\boldsymbol{\alpha}} \hat{H}(\boldsymbol{\alpha}) \left(1 - \prod_{i=1}^{m} (1-\alpha_i)^{n_i}\right). \quad (3)$$

An accurate approximation to the histogram can also be leveraged to answer many other questions about the populations that can not be readily addressed via the linear estimators of Section 4. These include tasks such as estimating the amount of data that must be collected to capture, say, 99% of the mass of the distributions in question.

6. Experiments

Evaluating the weighted linear estimator for large m. We empirically evaluated the performance of the weighted linear estimator $\hat{U}^W$. The experiments were conducted for three types of distributions—Uniform, Dirichlet and Geometric—that are commonly used to evaluate extrapolation algorithms. Each experiment contains $m = 100$ populations. We have a total of 3000 distinct elements. In the Uniform setting, each population has support on 100 elements that are randomly sampled from the 3000. For Dirichlet, each population also has support on 100 random elements (from the 3000), and the weights on these 100 elements are sampled from a Dirichlet prior. For the Geometric experiments, each population corresponds to a random ordering of the 3000 elements and the k-th element is assigned probability $\propto (1-p)^k p$. In period one, ten samples are observed in each of the 100 populations. In period two, 95 randomly chosen populations have extrapolation factor $t \in [0, 1]$ and five populations have extrapolation factor $10t$. This simulates the setting where we can obtain substantially more samples from a subset of the populations.

Figure 1(a, b, c) shows the results of the experiments for Uniform, Dirichlet(1) and Geometric with $p = 0.05$ respectively. The results for other parameter settings are qualitatively similar. The black curves indicate the true number of distinct new elements we expect to observe in period two by sampling from the true underlying distributions. The red curves are the predictions of the weighted linear estimator (shaded regions indicate one standard deviation across 100 experiments). In all three settings, $\hat{U}^W$ provides accurate estimate with low variance when the maximum extrapolation factor is relatively small (≤ 3). For Uniform and Geometric distributions, the accuracy is high up to 10 fold extrapolation. For Zipf, the bias is low but variance becomes large for the maximum extrapolation factor around 10. The downward bias in the predictions

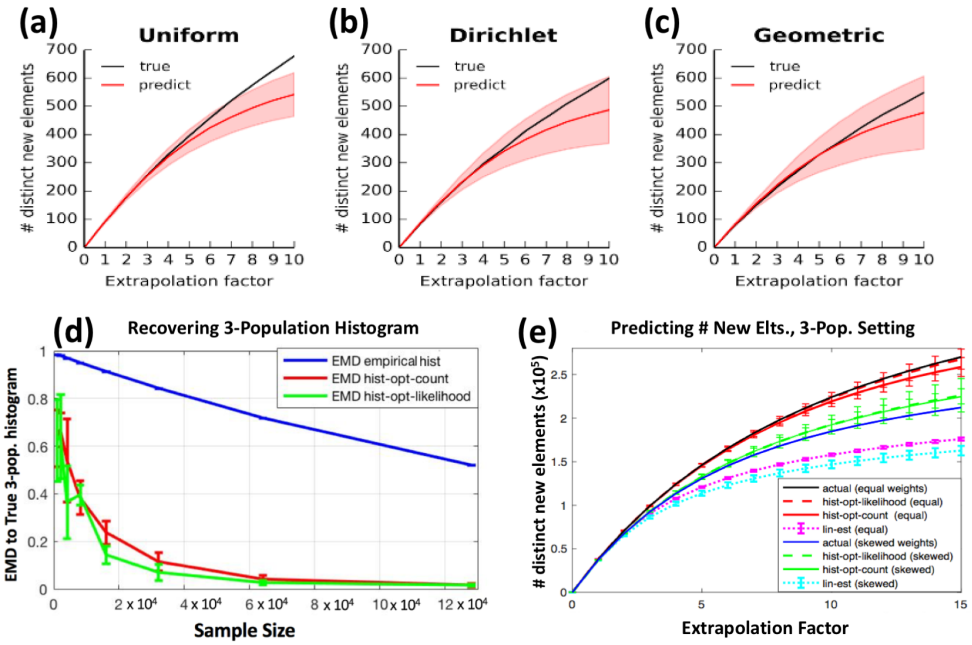

Figure 1. Performance of the weighted linear estimator of Prop. 4.4 for **(a)** Uniform, **(b)** Dirichlet and **(c)** Geometric distributions. Each experiment contains 100 populations. The x-axis corresponds to the maximum extrapolation factor among the 100 populations (t_1 in Prop. 4.4). The black curve indicates the true number of distinct new elements that we expect to observe in the new samples, and the red curve shows the predicted number of new elements. The red shaded region corresponds to one standard deviation over 100 independent experiments. **(d)** The 3-population earthmover distance (EMD) between the recovered histograms and the true histogram corresponding to the populations from which the samples were drawn. The blue line corresponds to the histogram of the empirical distribution of the samples, and the red and green lines correspond to the histograms returned by our multi-population histogram estimation algorithm, using the count-objective and likelihood objectives, respectively. Plots depict the mean and standard deviation over 5 independent runs. The true underlying distribution is supported on $4 \cdot 10^5$ domain elements. **(e)** Estimating the number of new domain elements that will be observed given additional samples in the same 3-population setting. Estimates are made for when the new samples are evenly distributed among the populations (equal) and when the majority come from one population (skewed). Error bars depict one standard deviation about the mean, calculated based on 10 independent trials.

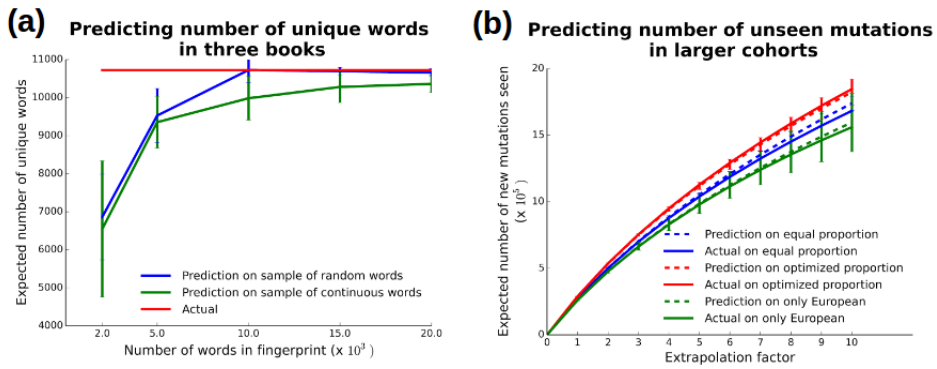

Figure 2. **(a)** Estimating the total number of unique words combining three different books using hist-opt-counts. Predictions are based on fingerprints of samples of words—sampled without replacement either randomly (blue) or from a contiguous block of text (green) from each book. Error bars depict one standard deviation over 10 independent runs. **(b)** Estimating the number of new mutations that would be observed given additional samples from four different populations using hist-opt-counts. We consider different ratios of sampling within these populations and observe the change in number of new mutations that would be observed.

is due to the weighting scheme. The relative error of the weighted estimator, $\left(\frac{\hat{U}^W - U}{\sum n_j t_j}\right)^2$, is 0.09, 0.08 and 0.08 for the Uniform, Dirichlet and Geometric distributions when the maximum extrapolation factor is 10. This confirms the theoretical results of Prop. 4.4 on the accuracy of the weighted linear estimator.

Evaluating the histogram estimators. We first validated the performance of $\hat{H}_{count}$ and $\hat{H}_{ll}$ on a three population setting with synthetic data. The true population consists of three uniform distributions over 200k elements, whose supports have 100k elements in common, and 100k elements unique to each distribution. In Figure 1(d), the x-axis corresponds to the number of samples we observe from each population, and the y-axis indicates the earthmover distance (EMD) between $\hat{H}_{count}$, $\hat{H}_{ll}$ and the true histogram. As a baseline, we also compute the EMD between the empirical histogram of the observed samples and the true histogram. $\hat{H}_{count}$ and $\hat{H}_{ll}$ performed roughly equally well and both are substantially better than the empirical estimator especially when the number of observed samples is small. Figure 1(e) illustrates the extrapolation accuracy of our histogram estimators. We estimated $\hat{H}_{count}$ and $\hat{H}_{ll}$ using 16K from each population, and then used Eqn. 3 to estimate the number of unseen elements in additional samples. We tested two different settings: 1) when the additional samples are equally drawn from the three populations, and 2) a skewed mixture where 5/6 of the new samples are from population 1 and 1/12 each are drawn from population 2 and 3. $\hat{H}_{count}$ and $\hat{H}_{ll}$ gave extremely accurate predictions. In comparison, the weighted linear estimator $\hat{U}^W$ was accurate for the initial extrapolations but has downward bias when the extrapolation increases, consistent with Fig. 1(a-c).

Additionally, we evaluate the performance of $\hat{H}_{count}$ on a real dataset, in which we sampled words from three books– *Hamlet* (32K total words), *Treasure Island* (40K) and *The Sun Also Rises* (72K). We used the true word frequencies (over the entire text) as the true histogram. We sampled a small number of words (equal in all books) either randomly or from a contiguous block of text and used $\hat{H}_{counts}$ to predict the total number of distinct words in total in all three books. In Figure 2(a), the red line is the true value, and blue and green lines are predictions based on $\hat{H}_{count}$ derived from samples of either random words, or words occurring in a random contiguous block of text, respectively. We obtain accurate estimates using a fraction of words (10K from each book). The estimates based on independent samples of words is more accurate than that based on contiguous blocks of text—likely due to correlation in words that occur near each other.

Optimizing discovery rate. Given the estimated histogram $\hat{H}_{count}$ or $\hat{H}_{ll}$, we can optimize the allocation of new samples across the populations to maximize the number of unseen elements we can expect to discover given a bound on $\sum_j t_j n_j$. To illustrate, we obtained genome sequencing data of 45K individuals from the Exome Aggregation Consortium (Lek et al., 2016). The individuals come from four ancestries: Europeans, Africans, East Asians and Latinos. We used all the observed mutations from the 45K samples to construct a four population frequency distribution. For the experiment, we treat this as the ground truth and sampled 10^5 mutations from each population to obtain "seen" data. Suppose we have budget to sample 3×10^6 variants (10 fold extrapolation from current sample size), how should we allocate these new samples across the four populations in order to maximize the number of new variants discovered? We use $\hat{H}_{count}$ to predict the extrapolation curves for three scenarios: 1. all the samples are allocated to Europeans (current genomic studies are heavily enriched of Europeans); 2. the samples are evenly allocated across the four populations; 3) we explicitly optimize the factors t_j using $\hat{H}_{count}$. The dotted curves in Fig. 2(b) correspond to the predictions, and the solid curves are the actual numbers using the true distribution, showing good agreement. Optimization using $\hat{H}_{count}$ led to 10 % increase in the number of new variants discovered. This is a simplistic example (there are many other factors in the design of real cohorts) but it illustrate the potential power in having multi-population histogram estimates. In Appendix Fig. 3, we also show that $\hat{H}_{count}$ gives accurate predictions for a different statistic—the number of new variants we expect to find at least twice in the new samples.

7. Discussion

We introduce and formalize the problem of multi-population unseen estimation. We provide a weighted linear estimator for the number of new elements and a general optimization algorithm to estimate the multi-population histogram. These two approaches have complementary strength. The weighted linear estimator $\hat{U}^W$ specifically estimates the number of unseen elements. It's accuracy is independent of the number of populations, m, and it is worst-case optimal. This can be a good method especially when m is large and the extrapolation factor is small compared to $\log$ of the number of observed samples. When the extrapolation is larger, however, $\hat{U}^W$ is consistently downward biased due to its variance-reducing weights. For relatively small number of populations ($m = 2, 3, 4$) and larger extrapolation factors, the unseen predictions of our histogram estimators, $\hat{H}_{count}$ and $\hat{H}_{ll}$ are significantly more accurate than $\hat{U}^W$. While both likely have comparable worst-case performance, the linear estimator nearly always incurs this worst-case loss and is largely incapable of extrapolating beyond this worst-case logarithmic factor. In contrast, the histogram-based estimators seem to perform well for much larger extrapolation factors on all of the distributions that we considered. $\hat{H}_{count}$ and $\hat{H}_{ll}$ are computationally more expensive than $\hat{U}^W$, but are still tractable for many applications—each run of our experiments took less than 20 minutes on a single laptop.

Acknowledgments

Gregory Valiant's contributions were supported by NSF CAREER CCF-1351108 and a Sloan Research Fellowship. James Zou is a Chan Zuckerberg Biohub investigator and is supported by NSF CISE-1657155.

References

Acharya, Jayadev, Das, Hirakendu, Orlitsky, Alon, and Suresh, Ananda Theertha. A unified maximum likelihood approach for optimal distribution property estimation. *arXiv preprint arXiv:1611.02960*, 2016.

Efron, Bradley and Thisted, Ronald. Estimating the number of unsen species: How many words did shakespeare know? *Biometrika*, pp. 435–447, 1976.

Fisher, Ronald A, Corbet, A Steven, and Williams, Carrington B. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pp. 42–58, 1943.

Good, IJ and Toulmin, GH. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

Good, Irving J. The population frequencies of species and the estimation of population parameters. *Biometrika*, pp. 237–264, 1953.

Lek, Monkol, Karczewski, Konrad J, Minikel, Eric V, Samocha, Kaitlin E, Banks, Eric, Fennell, Timothy, ODonnell-Luria, Anne H, Ware, James S, Hill, Andrew J, Cummings, Beryl B, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536 (7616):285–291, 2016.

Orlitsky, Alon, Suresh, Ananda Theertha, and Wu, Yihong. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, pp. 201607774, 2016.

Valiant, Gregory and Valiant, Paul. Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 685–694. ACM, 2011.

Valiant, Gregory and Valiant, Paul. Instance optimal learning of discrete distributions. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 142–155. ACM, 2016.

Valiant, Paul and Valiant, Gregory. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pp. 2157–2165, 2013.

Zou, James, Valiant, Gregory, Valiant, Paul, Karczewski, Konrad, Chan, Siu On, Samocha, Kaitlin, Lek, Monkol, Sunyaev, Shamil, Daly, Mark, and MacArthur, Daniel G. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications*, 7, 2016.