Estimation of Distribution Overlap of Urn Models

Jerrad Hampton[1], Manuel E. Lladser[1,*]
1 Department of Applied Mathematics, University of Colorado, Boulder, Colorado, United States of America
∗ **E-mail: manuel.lladser@colorado.edu**

Abstract

A classical problem in statistics is estimating the expected coverage of a sample, which has had applications in gene expression, microbial ecology, optimization, and even numismatics. Here we consider a related extension of this problem to random samples of two discrete distributions. Specifically, we estimate what we call the dissimilarity probability of a sample, i.e., the probability of a draw from one distribution not being observed in k draws from another distribution. We show our estimator of dissimilarity to be a U-statistic and a uniformly minimum variance unbiased estimator of dissimilarity over the largest appropriate range of k. Furthermore, despite the non-Markovian nature of our estimator when applied sequentially over k, we show it converges uniformly in probability to the dissimilarity parameter, and we present criteria when it is approximately normally distributed and admits a consistent jackknife estimator of its variance. As proof of concept, we analyze V35 16S rRNA data to discern between various microbial environments. Other potential applications concern any situation where dissimilarity of two discrete distributions may be of interest. For instance, in SELEX experiments, each urn could represent a random RNA pool and each draw a possible solution to a particular binding site problem over that pool. The dissimilarity of these pools is then related to the probability of finding binding site solutions in one pool that are absent in the other.

Introduction

An inescapable problem in microbial ecology is that a sample from an environment typically does not observe all species present in that environment. In [1], this problem has been recently linked to the concepts of *coverage probability* (i.e. the probability that a member from the environment is represented in the sample) and the closely related *discovery* or *unobserved probability* (i.e. the probability that a previously unobserved species is seen with another random observation from that environment). The mathematical treatment of coverage is not limited, however, to microbial ecology and has found applications in varied contexts, including gene expression, microbial ecology, optimization, and even numismatics.

The point estimation of coverage and discovery probability seem to have been first addressed by Turing and Good [2] to help decipher the Enigma Code, and subsequent work has provided point predictors and prediction intervals for these quantities under various assumptions [1, 3–5].

Following Robbins [6] and in more generality Starr [7], an unbiased estimator of the expected discovery probability of a sample of size n is

$$\sum_{k=1}^{r} \frac{\binom{r-1}{k-1}}{\binom{n+r}{k}} \cdot N(k, n+r), \tag{1}$$

where $N(k, n+r)$ is the number of species observed exactly k-times in a sample with replacement of size $(n+r)$. Using the theory of U-statistics developed by Halmos [8], Clayton and Frees [9] show that the above estimator is the *uniformly minimum variance unbiased estimator* (UMVUE) of the expected discovery probability of a sample of size n based on an enlarged sample of size $(n+r)$.

A quantity analogous to the discovery probability of a sample from a single environment but in the context of two environments is *dissimilarity*, which we broadly define as the probability that a draw in one environment is not represented in a random sample (of a given size) from a possibly different environment. Estimating the dissimilarity of two microbial environments is therefore closely related to the problem

of assessing the species that are unique to each environment, and the concept of dissimilarity may find applications to measure sample quality and allocate additional sampling resources, for example, for a more robust and reliable estimation of the UniFrac distance [10,11] between pairs of environments. Dissimilarity may find applications in other and very different contexts. For instance, in SELEX experiments [12]—a laboratory technique in which an initial pool of synthesized random RNA sequences is repeatedly screened to yield a pool containing only sequences with given biological functions—the dissimilarity of two RNA pools corresponds to the probability of finding binding site solutions in one pool that are absent in the other.

In this manuscript, we study an estimator of dissimilarity probability similar to Robbins' and Starr's statistic for discovery probability. Our estimator is optimal among the appropriate class of unbiased statistics, while being approximately normally distributed in a general case. The variance of this statistic is estimated using a consistent jackknife. As proof of concept, we analyze samples of processed V35 16S rRNA data from the Human Microbiome Project [13].

Probabilistic Formulation and Inference Problem

To study dissimilarity probability, we use the mathematical model of a pair of urns, where each urn has an unknown composition of balls of different colors, and where there is no a priori knowledge of the contents of either urn. Information concerning the urn composition is inferred from repeated draws with replacement from that urn.

In what follows, $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are independent sequences of independent and identically distributed (i.i.d.) discrete random variables with probability mass functions $\mathbb{P}_x$ and $\mathbb{P}_y$, respectively. Without loss of generality we assume that $\mathbb{P}_x$ and $\mathbb{P}_y$ are supported over possibly infinite subsets of $\mathbb{N} = \{1, 2, 3, \ldots\}$, and think of outcomes from these distributions as "colors": i.e. we speak of color-1, color-2, etc. Let I_x denote the set of colors i such that $\mathbb{P}_x(i) > 0$, and similarly define I_y. Under this perspective, X_k denotes the color of the k-th ball drawn with replacement from urn-x. Similarly, Y_k is the color of the k-th ball drawn with replacement from urn-y. Note that based on our formulation, distinct draws are always independent.

The mathematical analysis that follows was motivated by the problem of estimating the fraction of balls in urn-x with a color that is absent in urn-y. We can write this parameter as

$$\theta_{x,y}(\infty) := \sum_{i \in (I_x \setminus I_y)} \mathbb{P}_x(i) = \lim_{k \to \infty} \theta_{x,y}(k), \tag{2}$$

where

$$\theta_{x,y}(k) := \sum_{i \in I_x} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^k = \mathbb{P}\left(X_1 \notin \{Y_1, \ldots, Y_k\}\right). \tag{3}$$

The parameter $\theta_{x,y}(\infty)$ measures the proportion of urn-x which is unique from urn-y. On the other hand, $\theta_{x,y}(k)$ is a measure of the effectiveness of k-samples from urn-y to determine uniqueness in urn-x. This motivates us to refer to the quantity in (2) as the *dissimilarity of urn-x from urn-y*, and to the quantity in (3) as the *average dissimilarity of urn-x relative to k-draws from urn-y*. Note that these parameters are in general asymmetric in the roles of the urns. In what follows, urns-x and -y are assumed fixed, which motivates us to remove subscripts and write $\theta(k)$ instead of $\theta_{x,y}(k)$.

Unfortunately, one cannot estimate unbiasedly the dissimilarity of one urn from another based on finite samples, as stated in the following result. (See the Materials and Methods section for the proofs of all of our results.)

Theorem 1. *(No unbiased estimator of dissimilarity.) There is no unbiased estimator of $\theta(\infty)$ based on finite samples from two arbitrary urns-x and -y.*

Furthermore, estimating $\theta(\infty)$ accurately without further assumptions on the compositions of urns-x and -y seems a difficult if not impossible task. For instance, arbitrarily small perturbations of urn-y are likely to be unnoticed in a sample of a given size from this urn but may drastically affect the dissimilarity of other urns from urn-y. To demonstrate this idea, consider a parameter $0 \leq \epsilon \leq 1$ and let $\mathbb{P}_x(1) := 1$, $\mathbb{P}_y(1) := \epsilon$ and $\mathbb{P}_y(2) := (1-\epsilon)$. If $\epsilon = 0$ then $\theta(\infty) = 1$ while, for each $\epsilon > 0$, $\theta(\infty) = 0$.

In contrast with the above, for fixed k, $\theta(k)$ depends continuously on $(\mathbb{P}_x, \mathbb{P}_y)$ e.g. under the metric

$$d\big((\mathbb{P}_x, \mathbb{P}_y), (\mathbb{P}_{x'}, \mathbb{P}_{y'})\big) := \|\mathbb{P}_x - \mathbb{P}_{x'}\| + \|\mathbb{P}_y - \mathbb{P}_{y'}\|,$$

where $\|\nu\| := \sup_{A \subset \mathbb{N}} |\nu(A)| = \sum_i |\nu(i)|/2$ denotes the total variation of a signed measure ν over $\mathbb{N}$ such that $\nu(\mathbb{N}) = 0$. This is the case because

$$\begin{aligned} \left| \sum_i \mathbb{P}_x(i)(1-\mathbb{P}_y(i))^k - \sum_i \mathbb{P}_{x'}(i)(1-\mathbb{P}_{y'}(i))^k \right| &\leq \sum_i |\mathbb{P}_x(i) - \mathbb{P}_{x'}(i)| + k \sum_i |\mathbb{P}_y(i) - \mathbb{P}_{y'}(i)|, \\ &\leq 2(k+1) \cdot d\big((\mathbb{P}_x, \mathbb{P}_y), (\mathbb{P}_{x'}, \mathbb{P}_{y'})\big). \end{aligned}$$

The above implies that $\theta(k)$ is continuous with respect to any metric equivalent to d. Many such metrics can be conceived. For instance, if $(\mathbb{P}_x^m \times \mathbb{P}_y^n)$ denotes the probability measure associated with m samples with replacement from urn-x that are independent of n samples with replacement from urn-y then $\theta(k)$ is also continuous with respect to any of the metrics $d_{m,n}\big((\mathbb{P}_x, \mathbb{P}_y), (\mathbb{P}_{x'}, \mathbb{P}_{y'})\big) := \|(\mathbb{P}_x^m \times \mathbb{P}_y^n) - (\mathbb{P}_{x'}^m \times \mathbb{P}_{y'}^n)\|$, with $m, n \geq 1$, because

$$d\big((\mathbb{P}_x, \mathbb{P}_y), (\mathbb{P}_{x'}, \mathbb{P}_{y'})\big)/2 \leq d_{m,n}\big((\mathbb{P}_x, \mathbb{P}_y), (\mathbb{P}_{x'}, \mathbb{P}_{y'})\big) \leq \max\{m, n\} \cdot d\big((\mathbb{P}_x, \mathbb{P}_y), (\mathbb{P}_{x'}, \mathbb{P}_{y'})\big).$$

Because of the above considerations, we discourage the direct estimation of $\theta(\infty)$ and focus on the problem of estimating $\theta(k)$ accurately.

Results

Consider a finite number of draws with replacement $X_1, \ldots, X_{n_x}$ and $Y_1, \ldots, Y_{n_y}$, from urn-x and urn-y, respectively, where $n_x, n_y \geq 1$ are assumed fixed. Using this data we can estimate $\theta(k)$, for $k = 1 : n_y$, via the estimator:

$$\hat{\theta}(k) := \frac{1}{n_x \binom{n_y}{k}} \sum_{j=0}^{n_y - k} \binom{n_y - j}{k} Q(j), \tag{4}$$

where

$$Q(j) := \left\{ \begin{array}{l} \text{number of indices } i = 1 : n_x \text{ such that} \\ \text{color } X_i \text{ occurs } j\text{-times in } Y_1, \ldots, Y_{n_y}. \end{array} \right. \tag{5}$$

We refer to $Q(0), \ldots, Q(n_y)$ as the Q-statistics summarizing the data from both urns. Due to the well-known relation: $\sum_{i=1}^{r} i = r(r+1)/2$, at most $(1 + \sqrt{2n_y})$ of these estimators are non-zero. This sparsity may be exploited in the calculation of the right-hand side of (4) over a large range of k's.

Our statistic in $\hat{\theta}(k)$ is the U-statistic associated with the kernel $[\![X_1 \notin \{Y_1, \ldots, Y_k\}]\!]$, where $[\![\cdot]\!]$ is used to denote the indicator function of the event within the brackets (Iverson's bracket notation). Following the approach by Halmos in [8], we can show that this U-statistic is optimal amongst the unbiased estimators of $\theta(k)$ for $k = 1 : n_y$. We note that no additional samples from either urn are necessary to estimate $\theta(k)$ unbiasedly over this range when $n_x \geq 1$. This contrasts with the estimator in equation (1), which requires sample enlargement for unbiased estimation of discovery probability of a sample of size n.

Theorem 2. *(Minimum variance unbiased estimator.) If $n_x \geq 1$ and $n_y \geq k$ then $\hat{\theta}(k)$ is the unique uniformly minimum variance unbiased estimator of $\theta(k)$. Further, no unbiased estimator of $\theta(k)$ exists for $n_x = 0$ or $n_y < k$.*

Our next result shows that $\hat{\theta}(k)$ converges uniformly in probability to $\theta(k)$ over the largest possible range where unbiased estimation of the later parameter is possible, despite the non-Markovian nature of $\hat{\theta}(k)$ when applied sequentially over k. The result asserts that $\hat{\theta}(k)$ is likely to be a good approximation of $\theta(k)$, uniformly for $k = 1 : n_y$, when n_x and n_y are large. The method of proof uses an approach by Hoeffding [14] for the exact calculation of the variance of a U-statistic.

Theorem 3. *(Uniform convergence in probability.) Independently of how n_x and n_y tend to infinity, it follows for each $\epsilon > 0$ that*

$$\lim_{n_x, n_y \to \infty} \mathbb{P}\left(\max_{k=1:n_y} |\hat{\theta}(k) - \theta(k)| > \epsilon \right) = 0. \tag{6}$$

We may estimate the variance of $\hat{\theta}(k)$ for $k = 1 : n_y$ via a leave-one-out or also called delete-1 jackknife estimator, using an approach studied by Efron and Stein [15] and Shao and Wu [16].

To account for variability in the x-data through a leave-one-out jackknife estimate, we require that $n_x \geq 2$ and let

$$S_x^2(k) \;:=\; \frac{1}{n_x(n_x-1)} \sum_{j=0}^{n_y-k} Q(j) \left(\frac{\binom{n_y-j}{k}}{\binom{n_y}{k}} - \hat{\theta}(k) \right)^2. \tag{7}$$

On the other hand, to account for variability in the y-data, consider for $i \geq 1$ and $j \geq 0$ the statistics

$$M(i,j) := \left\{ \begin{array}{l} \text{number of colors } c \text{ such that color } c \text{ occurs exactly} \\ i\text{-times in } (X_1, \ldots, X_{n_x}) \text{ and } j\text{-times in } (Y_1, \ldots, Y_{n_y}). \end{array} \right. \tag{8}$$

Clearly, $\sum_i i\, M(i,j) = Q(j)$; in particular, the M-statistics are a refinement of the Q-statistics. Define $S_y^2(n_y) := 0$ and, for $k < n_y$, define

$$S_y^2(k) \;:=\; \frac{n_y - 1}{n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y-k} j\, M(i,j) \left(i(c_{j-1}(k) - c_j(k)) + \hat{\theta}_y(k) - \hat{\theta}(k) \right)^2, \tag{9}$$

where

$$c_j(k) \;:=\; \frac{\binom{n_y-j-1}{k}}{n_x \binom{n_y-1}{k}}; \tag{10}$$

$$\hat{\theta}_y(k) \;:=\; \sum_{j=0}^{n_y-k-1} c_j(k)\, Q(j). \tag{11}$$

Our estimator of the variance of $\hat{\theta}(k)$ is obtained by summing the variance attributable to the x-data and the y-data and is given by

$$S^2(k) := S_x^2(k) + S_y^2(k), \tag{12}$$

for $k = 1 : n_y$; in particular, $S(k)$ is our jackknife estimate of the standard deviation of $\hat{\theta}(k)$.

To assess the quality of $S^2(k)$ as an estimate of the variance of $\hat{\theta}(k)$ and the asymptotic distribution of the later statistic, we require a few assumptions that rule out degenerate cases. The following conditions are used in the remaining theorems in this section:

(a) $|I_x \cap I_y| < \infty$.

(b) there are at least two colors in $(I_x \cap I_y)$ that occur in different proportions in urn-y; in particular, the conditional probability $\mathbb{P}_y(\cdot \mid I_x \cap I_y)$ is not a uniform distribution.

(c) urn-x contains at least one color that is absent in urn-y; in particular, $\theta(\infty) > 0$.

(d) n_x and n_y grow to infinity at a comparable rate i.e. $n_x = \Theta(n_y)$, which means that there exist finite constants $c_1, c_2 > 0$ such that $c_1 n_y \leq n_x \leq c_2 n_y$, as n_x, n_y tend to infinity.

Conditions (a-c) imply that $\hat{\theta}(k)$ has a strictly positive variance and that a projection random variable, intermediate between $\hat{\theta}(k)$ and $\theta(k)$, has also a strictly positive variance. The idea of projection is motivated by the analysis of Grams and Serfling in [17].

Condition (d) is technical and only used to show that the result in Theorem 5 holds for the largest possible range of values of k namely, for $k = 1 : n_y$. See [18] for results with uniformity related to Theorem 4, as well as uniformity results when condition (d) is not assumed.

Because the variance of $\hat{\theta}(k)$, from now on denoted $\mathbb{V}(\hat{\theta}(k))$, and its estimate $S^2(k)$ tend to zero as n_x and n_y increase, the unnormalized consistency result is unsatisfactory. As an alternative, we can show that $S^2(k)$ is a consistent estimator relative to $\mathbb{V}(\hat{\theta}(k))$, as stated next.

Theorem 4. *(Asymptotic consistency of variance estimation.) If conditions (a)-(c) are satisfied then, for each $k \geq 1$ and $\epsilon > 0$, it applies that*

$$\lim_{n_x, n_y \to \infty} \mathbb{P}\left(\left| \frac{S^2(k)}{\mathbb{V}(\hat{\theta}(k))} - 1 \right| > \epsilon \right) = 0. \tag{13}$$

Finally, under conditions (a)-(d), we show that $\hat{\theta}(k)$ is asymptotically normally distributed for all $k = 1 : n_y$, as n_x and n_y increase at a comparable rate.

Theorem 5. *(Asymptotic normality.) Let $Z \sim \mathcal{N}(0,1)$ i.e. Z has a standard normal distribution. If conditions (a)-(d) are satisfied then*

$$\lim_{n_x, n_y \to \infty} \max_{k=1:n_y} \left| \mathbb{P}\left(\frac{\hat{\theta}(k) - \theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k)}} \leq t \right) - \mathbb{P}(Z \leq t) \right| = 0, \tag{14}$$

for all real number t.

The non-trivial aspect of the above result is the asymptotic normality of $\hat{\theta}(k)$ when $k = \Theta(n_y)$, e.g. $\hat{\theta}(n_y)$, as the results we have found in the literature [14, 19, 20] only guarantee the asymptotic normality of our estimator of $\theta(k)$ for fixed k. We note that, due to Slutsky's theorem [21], it follows from (13) and (14) that the ratio

$$\frac{\hat{\theta}(k) - \theta(k)}{S(k)}$$

has, for fixed k, approximately a standard normal distribution when n_x and n_y are large and of a comparable order of magnitude.

Discussion

As proof of concept, we use our estimators to analyze data from the Human Microbiome Project (HMP) [13]. In particular, our samples are V35 16S rRNA data, processed by Qiime into an operational taxonomic unit (OTU) count table format (see File S1 in Supporting Information). Each of the

266 samples analyzed have more than 5000 successfully identified bacteria (see File S2 in Supporting Information). We sort these samples by the body location metadata describing the origin of the sample. This sorting yields the assignments displayed in Table 1.

We present our estimates of $\hat{\theta}(n_y)$ for all $266 \cdot 265$ possible sample comparisons in Figure 1, i.e., we estimate the average dissimilarity of sample-x relative to the full sample-y. Due to (4), observe that $\hat{\theta}(n_y) = Q(0)/n_y$. At the given sample sizes, we can differentiate four broad groups of environments: stool, vagina, oral/throat and skin/nostril. We differentiate a larger proportion of oral/throat bacteria found in stool than stool bacteria found in the oral/throat environments. We may also differentiate the throat, gingival and saliva samples, but cannot reliably differentiate between tongue and throat samples or between the subgingival and supragingival plaques. On the other hand, the stool samples have larger proportions of unique bacteria relative to other stool samples of the same type, and vaginal samples also have this property. In contrast the skin/nostril samples have relatively few bacteria that are not identified in other skin/nostril samples.

The above effects may be a property of the environments from which samples are taken, or an effect of noise from inaccurate estimates due to sampling. To rule out the later interpretation, we show estimates of the standard deviation of $\hat{\theta}(n_y)$ based on the jackknife estimator $S^2(n_y)$ from (12) in Figure 2. As $S^2_{n_y}(n_y)$ is zero, the error estimate is given by $S_x(n_y)$. We see from (7), with $k = n_y$, that

$$S(n_y) = \sqrt{\frac{\hat{\theta}(n_y) \cdot (1 - \hat{\theta}(n_y))}{n_x - 1}}.$$

Assuming a normal distribution and an accurate jackknife estimate of variance, $\theta(n_y)$ will be in the interval $\hat{\theta}(n_y) \pm 0.01$ with at least approximately 95% confidence, for any choice of sample comparisons in our data; in particular, on a linear scale, we expect at least 95% of the estimates in Figure 1 to be accurate in at least the first two digits.

As we mentioned earlier, estimating $\theta(\infty)$ accurately is a difficult problem. We end this section with two heuristics to assess how representative $\hat{\theta}(n_y)$ is of $\theta(\infty)$, when urn-y has at least two colors and at least one color in common with urn-x. First, observe that:

$$\theta(k) = \theta(\infty) + \sum_{i \in (I_x \cap I_y)} \mathbb{P}_x(i)(1 - \mathbb{P}_y(i))^k. \tag{15}$$

In particular, $\theta(k)$ is a strictly concave-up and monotonically decreasing function of the real-variable $k \geq 0$. Hence, if $\theta(n_y)$ is close to the asymptotic value $\theta(\infty)$, then $\theta(n_y) - \theta(n_y - 1)$ should be of small magnitude. We call the later quantity the *discrete derivative* of $\theta(k)$ at $k = n_y$. Since we may estimate the discrete derivative from our data, the following heuristic arises: *relatively large values of* $|\hat{\theta}(n_y) - \hat{\theta}(n_y - 1)|$ *are evidence that* $\hat{\theta}(n_y)$ *is not a good approximation of* $\theta(\infty)$.

Figure 3 shows the heat map of $|\hat{\theta}(n_y) - \hat{\theta}(n_y - 1)|$ for each pair of samples. These estimates are of order 10^{-5} for the majority of the comparisons, and spike to 10^{-4} for several sample-y of varied environment types, when sample-x is associated with a skin or vaginal sample. In particular, further sampling effort from environments associated with certain vaginal, oral or stool samples are likely to reveal bacteria associated with broadly defined skin or vaginal environments.

Another heuristic may be more useful to assess how close $\hat{\theta}(n_y)$ is to $\theta(\infty)$, particularly when the previous heuristic is inconclusive. As motivation, observe that $\theta(k) = \theta(\infty) + \Theta(\rho^k)$, because of the identity in (15), where

$$\rho := 1 - \min_{i \in (I_x \cap I_y)} \mathbb{P}_y(i).$$

Furthermore, $\log(\theta(k-1) - \theta(k)) = k(\ln \rho) + c + o(1)$, where c is certain finite constant. We can justify this approximation only when $\log(\theta(k-1) - \theta(k))$ is well approximated by a linear function

of k, in which case we let $\hat{\rho}$ denote the estimated value for ρ obtained from the linear regression. Since $0 \leq \theta(n_y) - \theta(\infty) \leq \rho^{n_y}$, the following more precise heuristic comes to light: *$\hat{\theta}(n_y)$ is a good approximation of $\theta(\infty)$ if the linear regression of $\log|\hat{\theta}(k-1) - \hat{\theta}(k)|$ for k near n_y gives a good fit, $S(n_y)$ is small relative to $\hat{\theta}(n_y)$, and $\hat{\rho}^{n_y}$ is also small.*

To fix ideas we have applied the above heuristic to three pairs of samples: $(255, 176)$, $(200, 139)$ and $(100, 10)$, with each ordered pair denoting urn-x and urn-y, respectively. As seen in Table 2 for these three cases, $\hat{\theta}(n_y)$ is at least 14-times larger than $S(n_y)$; in particular, due to the asymptotic normality of the later statistic, an appropriate use of the heuristic is reduced to a good linear fit and a small $\hat{\rho}^{n_y}$ value. In all three cases, $\hat{\rho}$ was computed from the estimates $\hat{\theta}(k)$, with $k = 5001 : n_y$.

For the $(255, 176)$-pair, $\hat{\rho}^{n_y}$ and the regression error, measured as the largest absolute residual associated with the best linear fit, are zero to machine precision, suggesting that $\hat{\theta}(n_y) = 0.9998$ is a good approximation of $\theta(\infty)$. This is reinforced by the blue plot in Figure 4. On the other hand, for the $(200, 139)$-pair, the regression error is small, suggesting that the linear approximation $\log(\hat{\theta}(k-1) - \hat{\theta}(k))$ is good for $k = 5001 : n_y$. However, because $\hat{\rho}^{n_y} = 0.9997$, we cannot guarantee that $\hat{\theta}(n_y)$ is a good approximation of $\theta(\infty)$. In fact, as seen in the red-plot in Figure 4, $\hat{\theta}(k)$, with $k = 1 : n_y$, exposes a steady and almost linear decay that suggests that $\theta(\infty)$ may be much smaller than $\hat{\theta}(n_y)$. Finally, for the $(100, 10)$-pair, the regression error is large and the heuristic is therefore inconclusive. Due to the green-plot in Figure 4, the lack of fit indicates that the exponential rate of decay of $\theta(k)$ to $\theta(\infty)$ has not yet been captured by the data from these urns. Note that the heuristic based on the discrete derivative shows no evidence that $\hat{\theta}(n_y)$ is far from $\theta(\infty)$.

Materials and Methods

Here we prove the theorems given in the Results section. The key idea to prove each theorem may be summarized as follows.

To show Theorem 1, we identify pairs of urns for which unbiased estimation of $\theta(\infty)$ is impossible for any statistic. To show Theorem 2, we exploit the diversity of possible urn distributions to show that there are relatively few unbiased estimators of $\theta(k)$ and, in fact, there is a single unbiased estimator $\hat{\theta}(k)$ that is symmetric on the data. The uniqueness of the symmetric estimator is obtained via a completeness argument: a symmetric statistic having expected value zero is shown to correspond to a polynomial with identically zero coefficients, which themselves correspond to values returned by the statistic when presented with specific data. The symmetric estimator is a U-statistic in that it corresponds to an average of unbiased estimates of $\theta(k)$, based on all possible sub-samples of size 1 and k from the samples of urn-x and -y, respectively. As any asymmetric estimator has higher variance than a corresponding symmetric estimator, the symmetric estimator must be the UMVUE.

To show Theorem 3 we use bounds on the variance of the U-statistic and show that, uniformly for relatively small k, $\hat{\theta}(k)$ converges to $\theta(k)$ in the $\mathcal{L}^2$-norm. In contrast, for relatively large values of k, we exploit the monotonicity of $\theta(k)$ and $\hat{\theta}(k)$ to show uniform convergence.

Finally, theorems 4 and 5 are shown using an approximation of $\hat{\theta}(k)$ by sums i.i.d. random variables, as well as results concerning the variance of both $\hat{\theta}(k)$ and its approximation. In particular, the approximation satisfies the hypotheses the Central Limit Theorem and Law of Large Numbers, which we use to transfer these results to $\hat{\theta}(k)$.

In what follows, $\mathcal{D}$ denotes the set of all probability distributions that are finitely supported over $\mathbb{N}$.

Proof of Theorem 1. Consider in $\mathcal{D}$ probability distributions of the form $\mathbb{P}_x(1) = 1$, $\mathbb{P}_y(1) = u$ and $\mathbb{P}_y(2) = (1-u)$, where $0 \leq u \leq 1$ is a given parameter. Any statistic $h(\cdot)$ which takes as input n_x draws from urn-x and n_y draws from urn-y has that $\mathbb{E}(h(X_1, \ldots, X_{n_x}, Y_1, \ldots, Y_{n_y}))$ is a polynomial of degree at most n_y in the variable u; in particular, it is a continuous function of u over the interval $[0, 1]$. Since

$\theta(\infty) = [\![u = 0]\!]$ has a discontinuity at $u = 0$ over this interval, there exists no estimator of $\theta(\infty)$ that is unbiased over pairs of distributions in $\mathcal{D}$. □

We use lemmas 6-11 to first show Theorem 2. The method of proof of this theorem follows an approach similar to the one used by Halmos [8] for single distributions, which we extend here naturally to the setting of two distributions.

Our next result implies that no uniformly unbiased estimator of $\theta(k)$ is possible when using less than one sample from urn-x and k samples from urn-y.

Lemma 6. *If $g(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ is unbiased for $\theta(k)$ for all $\mathbb{P}_x, \mathbb{P}_y \in \mathcal{D}$, then $m \geq 1$ and $n \geq k$.*

Proof. Consider in $\mathcal{D}$ probability distributions of the form $\mathbb{P}_x(1) = u$, $\mathbb{P}_x(2) = (1-u)$, $\mathbb{P}_y(1) = v$ and $\mathbb{P}_y(2) = (1-v)$, where $0 \leq u, v \leq 1$ are arbitrary real numbers. Clearly, $\mathbb{E}[g(X_1, \ldots, X_m, Y_1, \ldots, Y_n)]$ is a linear combination of polynomials of degree m in u and n in v and, as a result, it is a polynomial of degree at most m in u and n in v. Since $\theta(k) = u(1-v)^k + (1-u)v^k$ has degree 1 in u and k in v, and $g(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ is unbiased for $\theta(k)$, we conclude that $1 \leq m$ and $k \leq n$. □

The form of $\hat{\theta}(k)$ given in equation (4) is convenient for computation but, for mathematical analysis, we prefer its U-statistic form associated with the kernel function $(x, y_1, \ldots, y_k) \to [\![x \notin \{y_1, \ldots, y_k\}]\!]$.

In what follows, $S_{k,n}$ denotes the set of all functions $\sigma : \{1, \ldots, k\} \to \{1, \ldots, n_y\}$ that are one-to-one.

Lemma 7.

$$\hat{\theta}(k) = \frac{1}{n_x |S_{k,n_y}|} \sum_{i=1}^{n_x} \sum_{\sigma \in S_{k,n_y}} [\![X_i \notin \{Y_{\sigma(1)}, \ldots, Y_{\sigma(k)}\}]\!], \tag{16}$$

where $|S_{k,n_y}| = k!\binom{n_y}{k}$.

Proof. Fix $1 \leq i \leq n_x$ and suppose that color X_i occurs j-times in $Y_1, \ldots, Y_{n_y}$. If $j > (n_y - k)$ then any sublist of size k of $Y_1, \ldots, Y_{n_y}$ contains X_i, hence $[\![X_i \notin \{Y_{\sigma(1)}, \ldots, Y_{\sigma(k)}\}]\!] = 0$, for all $\sigma \in S_{k,n_y}$. On the other hand, if $j \leq (n_y - k)$ then $\sum_{\sigma \in S_{k,n_y}} [\![X_i \notin \{Y_{\sigma(1)}, \ldots, Y_{\sigma(k)}\}]\!] = k!\binom{n_y - j}{k}$. Since the rightmost sum only depends on the number of times that color X_i was observed in $Y_1, \ldots, Y_{n_y}$, we may use the Q-statistics defined in equation (5) to rewrite:

$$\frac{1}{n_x |S_{k,n_y}|} \sum_{i=1}^{n_x} \sum_{\sigma \in S_{k,n_y}} [\![X_i \notin \{Y_{\sigma(1)}, \ldots, Y_{\sigma(k)}\}]\!] = \frac{1}{n_x \binom{n_y}{k}} \sum_{j=0}^{n_y - k} \binom{n_y - j}{k} Q(j).$$

The right-hand side above now corresponds to the definition of $\hat{\theta}(k)$ given in equation (4). □

In what follows, we say that a function $f : \mathbb{N}^{n_x + n_y} \to \mathbb{R}$ is (n_x, n_y)-*symmetric* when

$$f(x_1, \ldots, x_{n_x}; y_1, \ldots, y_{n_y}) = f(x_{\sigma(1)}, \ldots, x_{\sigma(n_x)}; y_{\sigma'(1)}, \ldots, y_{\sigma'(n_y)}),$$

for all $x_1, \ldots, x_{n_x}, y_1, \ldots, y_{n_y} \in \mathbb{N}$ and permutations σ and σ' of $1, \ldots, n_x$ and $1, \ldots, n_y$, respectively. Alternatively, f is (n_x, n_y)-symmetric if and only if it may be regarded a function of $(x_{(1\ldots n_x)}, y_{(1\ldots n_y)})$, where $x_{(1\ldots n_x)}$ and $y_{(1\ldots n_y)}$ correspond to the order statistics $x_{(1)}, \ldots, x_{(n_x)}$ and $y_{(1)}, \ldots, y_{(n_y)}$, respectively. Accordingly, a statistic of $(X_1, \ldots, X_{n_x}, Y_1, \ldots, Y_{n_y})$ is called (n_x, n_y)-*symmetric* when it may be represented in the form $f(X_1, \ldots, X_{n_x}, Y_1, \ldots, Y_{n_y})$, for some (n_x, n_y)-symmetric function f. It is immediate from Lemma 7 that $\hat{\theta}(k)$ is (n_x, n_y)-symmetric.

The next result asserts that the variance of any non-symmetric unbiased estimator of $\theta(k)$ may be reduced by a corresponding symmetric unbiased estimator. The proof is based on the well-known fact that conditioning preserves the mean of a statistic and cannot increase its variance.

Lemma 8. *An asymmetric unbiased estimator of $\theta(k)$ that is square-integrable has a strictly larger variance than a corresponding (n_x, n_y)-symmetric unbiased estimator.*

Proof. Let $\mathcal{F}$ denote the sigma-field generated by the random vector $(X_{(1\ldots n_x)}; Y_{(1\ldots n_y)})$ and suppose that the statistic $T = f(X_1, \ldots, X_{n_x}, Y_1, \ldots, Y_{n_y})$ is unbiased for $\theta(k)$ and square-integrable. In particular, $U = \mathbb{E}[T \,|\, \mathcal{F}]$ is a well-defined statistic and there is an (n_x, n_y)-symmetric function $g : \mathbb{N}^{n_x+n_y} \to \mathbb{R}$ such that $U = g(X_1, \ldots, X_{n_x}; Y_1, \ldots, Y_{n_y})$. Clearly, U is unbiased for $\theta(k)$ and (n_x, n_y)-symmetric. Since $\mathbb{E}(T^2) < +\infty$, Jensen's inequality for conditional expectations [22] implies that $\mathbb{E}(U^2) \leq \mathbb{E}(T^2)$, with equality if and only if T is (n_x, n_y)-symmetric. □

Since $\hat{\theta}(k)$ is (n_x, n_y)-symmetric and bounded, the above lemma implies that if an UMVUE for $\theta(k)$ exists then it must be (n_x, n_y)-symmetric. Next, we show that there is a unique symmetric and unbiased estimator of $\theta(k)$, which immediately implies that $\hat{\theta}(k)$ is the UMVUE.

In what follows, $k_1, k_2 \geq 0$ denote integers. We say that a polynomial $Q(u_1, \ldots, u_m; v_1, \ldots, v_n)$ is (k_1, k_2)-*homogeneous* when it is a linear combination of polynomials of the form $\prod_{i=1}^m u_i^{m_i} \prod_{j=1}^n v_j^{n_j}$, with $\sum_{i=1}^m m_i = k_1$ and $\sum_{j=1}^n n_j = k_2$. Furthermore, we say that Q satisfies the *partial vanishing condition* if $Q(u_1, \ldots, u_m; v_1, \ldots, v_n) = 0$ whenever $u_1, \ldots, u_m, v_1, \ldots, v_n \geq 0$, $\sum_{i=1}^m u_i = 1$ and $\sum_{i=1}^n v_i = 1$.

The next lemma is an intermediate step to show that a (k_1, k_2)-homogeneous polynomial which satisfies the partial vanishing condition is the zero polynomial, which is shown in Lemma 10.

Lemma 9. *If Q is a (k_1, k_2)-homogeneous polynomial in the real variables $u_1, \ldots, u_m, v_1, \ldots, v_n$, with $m, n \geq 1$, that satisfies the partial vanishing condition, then $Q(u_1, \ldots, u_m; v_1, \ldots, v_n) = 0$ whenever $u_1, \ldots, u_m, v_1, \ldots, v_n \geq 0$, $\sum_{i=1}^m u_i > 0$ and $\sum_{i=1}^n v_i > 0$.*

Proof. Fix $u_1, \ldots, u_m, v_1, \ldots, v_n \geq 0$ such that $\sum_{i=1}^m u_i > 0$ and $\sum_{i=1}^n v_i > 0$ and observe that

$$Q(u_1, \ldots, u_m; v_1, \ldots, v_n) := \left(\sum_{i=1}^m u_i\right)^{k_1} \left(\sum_{i=1}^n v_i\right)^{k_2} Q\left(\frac{u_1}{\sum_{i=1}^m u_i}, \ldots, \frac{u_m}{\sum_{i=1}^m u_i}; \frac{v_1}{\sum_{i=1}^n v_i}, \ldots, \frac{v_n}{\sum_{i=1}^n v_i}\right),$$

because Q is a (k_1, k_2)-homogeneous polynomial. Notice now that the right hand-side above is zero because Q satisfies the partial vanishing condition. □

Lemma 10. *Let Q be a (k_1, k_2)-homogeneous polynomial in the real variables $u_1, \ldots, u_m, v_1, \ldots, v_n$, with $m, n \geq 1$. If Q satisfies the partial vanishing condition then $Q = 0$ identically.*

Proof. We prove the lemma using structural induction on (m, n) for all $k_1, k_2 \geq 0$.

If $m = n = 1$ then a (k_1, k_2)-homogeneous polynomial $Q(u_1, v_1)$ must be of the form $c u_1^{k_1} v_1^{k_2}$, for an appropriate constant c. As such a polynomial satisfies the partial-vanishing condition only when $c = 0$, the base case for induction is established.

Next, consider a (k_1, k_2)-homogeneous polynomial $Q(u_1, \ldots, u_m; v_1, \ldots, v_n, v_{n+1})$, with $m, n \geq 1$, that satisfies the partial vanishing condition, and let d denote its degree with respect to the variable v_{n+1}. In particular, there are polynomials $Q_0, \ldots, Q_d$ in the variables $u_1, \ldots, u_m, v_1, \ldots, v_n$ such that

$$Q(u_1, \ldots, u_m; v_1, \ldots, v_n, v_{n+1}) = \sum_{i=0}^d Q_i(u_1, \ldots, u_m; v_1, \ldots, v_n) v_{n+1}^i.$$

Now fix $u_1, \ldots, u_m, v_1, \ldots, v_n \geq 0$ such that $\sum_{i=1}^m u_i > 0$ and $\sum_{i=1}^n v_i > 0$. Because Q satisfies the partial vanishing condition, Lemma 9 implies that $\sum_{i=0}^d Q_i(u_1, \ldots, u_m; v_1, \ldots, v_n) v_{n+1}^i = 0$ for all $v_{n+1} > 0$. In particular, for each i, $Q_i(u_1, \ldots, u_m; v_1, \ldots, v_n) = 0$ whenever $u_1, \ldots, u_m, v_1, \ldots, v_n \geq 0$, $\sum_{i=1}^m u_i > 0$ and $\sum_{i=1}^n v_i > 0$. Thus each Q_i satisfies the partial vanishing condition. Since Q_i is a $(k_1, k_2 - i)$-homogeneous polynomial, the inductive hypothesis implies that $Q_i = 0$ identically and hence $Q = 0$ identically. The

same argument shows that if $Q(u_1,\ldots,u_m,u_{m+1};v_1,\ldots,v_n)$, with $m,n\geq 1$, is a (k_1,k_2)-homogeneous polynomial that satisfies the partial vanishing condition then $Q=0$ identically, completing the inductive proof of the lemma. □

Our final resultbefore proving Theorem 2 implies that $\theta(k)$ cannot admit more than one symmetric and unbiased estimator. Its proof depends on the variety of distributions in $\mathcal{D}$, and uses the requirement that our estimator must be unbiased for any pair of distributions chosen from $\mathcal{D}$.

Lemma 11. *If f is an (n_x,n_y)-symmetric function such that $\mathbb{E}[f(X_1,\ldots,X_{n_x},Y_1,\ldots,Y_{n_y})]=0$, for all $\mathbb{P}_x,\mathbb{P}_y\in\mathcal{D}$, then $f=0$ identically.*

Proof. Consider a point $\vec{z}=(x_1,\ldots,x_{n_x},y_1,\ldots,y_{n_y})\in\mathbb{N}^{n_x+n_y}$ and define m_1 and m_2 as the cardinalities of the sets $\{x_1,\ldots,x_{n_x}\}$ and $\{y_1,\ldots,y_{n_y}\}$, respectively. Furthermore, let $x'_1,\ldots,x'_{m_1}$ denote the distinct elements in the set $\{x_1,\ldots,x_{n_x}\}$ and define $m_{1,i}$ to be the number of times that x'_i appears in this set. Furthermore, let $\mathbb{P}_x\in\mathcal{D}$ be a probability distribution such that $\mathbb{P}_x(\{x'_1,\ldots,x'_{m_1}\})=1$ and define $p_{1,i}:=\mathbb{P}_x(x'_i)$. In a completely analogous manner define $y'_1,\ldots,y'_{m_2}$, $m_{2,j}$, $\mathbb{P}_y$ and $p_{2,j}$.

Notice that $\mathbb{E}[f(\vec{Z}_m)]$ is a polynomial in the real variables $p_{1,1},\ldots,p_{1,m_1},p_{2,1},\ldots,p_{2,m_2}$ that satisfies the hypothesis of Lemma 10; in particular, this polynomial is identically zero. However, because f is (n_x,n_y)-symmetric, the coefficient of $\prod_{i=1}^{2}\prod_{j=1}^{m_i}p_{i,j}^{m_{i,j}}$ in $\mathbb{E}[f(\vec{Z}_m)]$ is

$$f(\vec{z})\binom{n_x}{m_{1,1};\ldots;m_{1,m_1}}\binom{n_y}{m_{2,1};\ldots;m_{2,m_2}},$$

implying that $f(\vec{z})=0$. □

Proof of Theorem 2. From Lemma 8, as we mentioned already, if the UMVUE for $\theta(k)$ exists then it must be (n_x,n_y)-symmetric. Suppose there are two (n_x,n_y)-symmetric functions such that $f(X_1,\ldots,X_{n_x};Y_1,\ldots,Y_{n_y})$ and $g(X_1,\ldots,X_{n_x};Y_1,\ldots,Y_{n_y})$ are unbiased for $\theta(k)$. Applying Lemma 11 to $(f-g)$ shows that $f=g$, and $\theta(k)$ admits therefore a unique symmetric and unbiased estimator. From Lemma 7, $\hat{\theta}(k)$ is (n_x,n_y)-symmetric and unbiased for $\theta(k)$ hence it is the UMVUE for $\theta(k)$. From Lemma 6, it follows that no unbiased estimator of $\theta(k)$ exists for $n_x=0$ or $n_y<k$. □

Our next goal is to show Theorem 3, for which we prove first lemmas 12-13. We note that the later lemma applies in a much more general context than our treatment of dissimilarity.

Lemma 12. *If, for each $n\geq 1$, $k_n\geq 1$ is an integer such that $k_n^2=o(n)$ then*

$$\frac{\binom{k}{1}\binom{n-k}{k-1}}{\binom{n}{k}}=\frac{k^2}{n}+O\left(\frac{k^4}{n^2}\right); \tag{17}$$

$$\sum_{j=2}^{k}\frac{\binom{k}{j}\binom{n-k}{k-j}}{\binom{n}{k}}=O\left(\frac{k^4}{n^2}\right); \tag{18}$$

uniformly for $k=1:k_n$ as $n\to\infty$.

Proof. First observe that for all n sufficiently large and $k=1:k_n$, it applies that

$$\frac{\binom{k}{1}\binom{n-k}{k-1}}{\binom{n}{k}}=\frac{k^2}{n}\prod_{i=0}^{k-2}\left(1-\frac{k-1}{n-1-i}\right)=\frac{k^2}{n}\exp\left\{\sum_{i=0}^{k-2}\log\left(1-\frac{k-1}{n-1-i}\right)\right\}.$$

Note that $-x/(1-x) \leq \log(1-x) \leq -x$, for all $0 \leq x < 1$. As a result, we may bound the exponential factor on the right-hand side above as follows:

$$e^{-(k-1)^2/(n-2k+2)} \leq \exp\left\{\sum_{i=0}^{k-2} \log\left(1 - \frac{k-1}{n-1-i}\right)\right\} \leq e^{-(k-1)^2/(n-1)}.$$

Since $e^{-(k-1)^2/(n-2k+2)} = 1 + O(k^2/n)$ and $e^{-(k-1)^2/(n-1)} = 1 + O(k^2/n)$, uniformly for all $k = 1 : k_n$ as $n \to \infty$, (17) follows.

To show (18), first note the combinatorial identity

$$\sum_{j=0}^{k} \frac{\binom{k}{j}\binom{n-k}{k-j}}{\binom{n}{k}} = 1. \tag{19}$$

Proceeding in an analogous manner as we did to show (17), we see now that the term associated with the index $j = 0$ in the above summation satisfies that

$$e^{-k^2/(n-2k+1)} \leq \frac{\binom{k}{0}\binom{n-k}{k}}{\binom{n}{k}} \leq e^{-k^2/n},$$

for all n sufficiently large and $k = 1 : k_n$. Since $e^{-k^2/(n-2k+1)} = 1 - k^2/n + O(k^4/n^2)$ and $e^{-k^2/n} = 1 - k^2/n + O(k^4/n^2)$, the above inequalities together with (17) and (19) establish (18). □

Lemma 13. *Define $\lambda(k) := \mathbb{E}(h(X_1, Y_1, \ldots, Y_k))$, where $h(x_1, y_1, \ldots, y_k)$ is a bounded $(1,k)$-symmetric function, and let*

$$\hat{\lambda}(k) = \hat{\lambda}_{n_x,n_y}(k) := \frac{1}{n_x |S_{k,n_y}|} \sum_{i=1}^{n_x} \sum_{\sigma \in S_{k,n_y}} h(X_i, Y_{\sigma(1)}, \ldots, Y_{\sigma(k)})$$

be the U-statistic of $\lambda(k)$ associated with n_x draws from urn-x and n_y draws from urn-y; in particular, $\mathbb{E}(\hat{\lambda}(k)) = \lambda(k)$. Furthermore, assume that

(i) $0 \leq h \leq 1$,

(ii) there is a function $f : I_x \to [0,1]$ such that $\lim_{k\to\infty} h(X_1, Y_1, \ldots, Y_k) \overset{a.s}{=} f(X_1)$,

(iii) $\hat{\lambda}(k) \geq \hat{\lambda}(k+1)$; in particular, $\lambda(k) \geq \lambda(k+1)$.

Under the above assumptions, it follows that

$$\lim_{n_x,n_y\to\infty} \mathbb{E}\left(\max_{k=1:n_y} \left|\hat{\lambda}(k) - \lambda(k)\right|^2\right) = 0.$$

Proof. Define $k_n := 1 + \min\left\{\lfloor n_x^{1/2} \rfloor, \lfloor \log(n_y) \rfloor\right\}$; in particular, $1 \leq k_n \leq n_y$ and $k_n \to \infty$, $k_n = o(n_x)$ and $k_n^p = o(n_y)$, for any $p > 0$, as $n_x, n_y \to \infty$. The proof of the theorem is reduced to show that

$$\lim_{n_x,n_y\to\infty} \mathbb{E}\left(\max_{k=k_n:n_y} |\hat{\lambda}(k) - \lambda(k)|^2\right) = 0; \tag{20}$$

$$\lim_{n_x,n_y\to\infty} \mathbb{E}\left(\max_{k=1:k_n} |\hat{\lambda}(k) - \lambda(k)|^2\right) = 0. \tag{21}$$

Next, we compute the variance of $\hat{\lambda}(k)$ following an approach similar to Hoeffding [14]. Because h is $(1,k)$-symmetric, a tedious yet standard calculation shows that

$$\mathbb{V}\big(\hat{\lambda}(k)\big) = \frac{\sum_{j=0}^{k} \binom{k}{j}\binom{n_y-k}{k-j}((n_x-1)\xi_{0,j}(k) + \xi_{1,j}(k))}{n_x\binom{n_y}{k}}, \tag{22}$$

where

$$\xi_{0,j}(k) := \mathbb{V}\big(\mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|Y_1, \ldots, Y_j)\big); \tag{23}$$

$$\xi_{1,j}(k) := \mathbb{V}\big(\mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|X_1, Y_1, \ldots, Y_j)\big). \tag{24}$$

Clearly, $\xi_{1,j}(k) \leq 1$. On the other hand, if W is any random variable with finite expectation and $\mathcal{F}_1 \subset \mathcal{F}_2$ are sigma-fields then $\mathbb{V}(\mathbb{E}(W|\mathcal{F}_1)) \leq \mathbb{V}(\mathbb{E}(W|\mathcal{F}_2))$, due to well-known properties of conditional expectations [22]. In particular, for each $0 \leq j \leq k$, we have that

$$\xi_{0,j}(k) \leq \xi_{0,k}(k), \text{ and } \xi_{1,j}(k) \leq \xi_{1,k}(k). \tag{25}$$

Consequently, (22) implies that

$$\mathbb{V}\big(\hat{\lambda}(k)\big) \leq \xi_{0,k}(k) + \frac{1}{n_x}. \tag{26}$$

We claim that

$$\lim_{k\to\infty} \xi_{0,k}(k) = 0. \tag{27}$$

Indeed, using an argument similar as above, we find that

$$\begin{aligned}
\xi_{0,k}(k) &= \mathbb{V}\Big(\mathbb{E}\big(h(X_1, Y_1, \ldots, Y_k) - f(X_1) \mid Y_1, \ldots, Y_k\big)\Big), \\
&\leq \mathbb{V}\Big(\mathbb{E}\big(h(X_1, Y_1, \ldots, Y_k) - f(X_1) \mid Y_1, \ldots, Y_k, X_1\big)\Big), \\
&= \mathbb{V}\big(h(X_1, Y_1, \ldots, Y_k) - f(X_1)\big).
\end{aligned}$$

Due to assumptions (i)-(ii) and the Bounded Convergence Theorem, the right-hand side above tends to 0, and the claim follows.

It follows from (26) and (27) that

$$\lim_{k\to\infty} \mathbb{V}\big(\hat{\lambda}(k)\big) = 0.$$

Finally, because of assumption (iii),

$$\mathbb{E}\left(\max_{k=k_n:n_y} |\hat{\lambda}(k) - \lambda(k)|^2\right) \leq 2|\lambda(k_n) - \lambda(n_y)|^2 + 2\mathbb{V}(\hat{\lambda}(k_n)) + 2\mathbb{V}(\hat{\lambda}(n_y)).$$

Since each term on the right-hand side above tends to zero as $n_x, n_y \to \infty$, (20) follows.

We now show (21). As $\xi_{i,j}(k) \leq 1$ and $\xi_{0,0}(k) = 0$, it follows by (22) and Lemma 12 that

$$\mathbb{V}\left(\hat{\lambda}(k)\right) \leq 1 - \frac{\binom{n_y-k}{k}}{\binom{n_y}{k}} + \frac{1}{n_x} = \frac{1}{n_x} + \sum_{j=1}^{k} \frac{\binom{k}{j}\binom{n_y-k}{k-j}}{\binom{n_y}{k}} = \frac{1}{n_x} + \frac{k^2}{n_y} + O\left(\frac{k^4}{n_y^2}\right),$$

uniformly for $k = 1 : k_n$ as $n_x, n_y \to \infty$. In particular,

$$\mathbb{E}\left(\max_{k=1:k_n} |\hat{\lambda}(k) - \lambda(k)|^2\right) \leq \sum_{k=1}^{k_n} \mathbb{V}(\hat{\lambda}(k)) \leq \frac{k_n}{n_x} + \frac{k_n^3}{n_y} + O\left(\frac{k_n^5}{n_y^2}\right).$$

Due to the definition of the coefficients k_n, the right-hand side above tends to zero, and (21) follows. □

Proof of Theorem 3. Note that $\theta(k)=\mathbb{E}(h(X_1,Y_1,\ldots,Y_k))$, with $h(x_1,y_1,\ldots,y_k):=[\![x_1\notin\{y_1,\ldots,y_k\}]\!]$. We show that the kernel function h and the U-statistics $\hat{\theta}(k)$ satisfy the hypotheses of Lemma 13. From this the theorem is immediate because $\mathcal{L}^2$-convergence implies convergence in probability.

Clearly h is $(1,k)$-symmetric and $0\le h\le 1$, which shows assumption (i) in Lemma 13. On the other hand, due to the Law of Large Numbers, $\lim_{k\to\infty} h(X_1,Y_1,\ldots,Y_k) = [\![X_1\notin I_y]\!]$ almost surely, from which assumption (ii) in the lemma also follows.

Finally, to show assumption (iii), recall that $S_{k,n}$ is the set of one-to-one functions from $\{1,\ldots,k\}$ into $\{1,\ldots,n\}$; in particular, $|S_{k+1,n_y}| = (n_y-k)\cdot|S_{k,n_y}|$. Now note that for each indicator of the form $[\![X_1\notin\{Y_{\sigma(1)},\ldots,Y_{\sigma(k)}\}]\!]$, with $\sigma\in S_{k+1,n_y}$, there are (n_y-k) choices of $\sigma(k+1)$ outside the set $\{\sigma(1),\ldots,\sigma(k)\}$. Because $[\![X_1\notin\{Y_{\sigma(1)},\ldots,Y_{\sigma(k)}\}]\!]\ge[\![X_1\notin\{Y_{\sigma(1)},\ldots,Y_{\sigma(k+1)}\}]\!]$, it follows that $\hat{\theta}(k)\ge\hat{\theta}(k+1)$ for all $k=1:(n_y-1)$. This shows condition (iii) in Lemma 13, and Theorem 3 follows.□

Proof of equation (7). The jackknife estimate of the variance of $\hat{\theta}(k)$ obtained from removing a single x-data is, by definition, the quantity

$$S_x^2(k) := \frac{n_x-1}{n_x}\sum_{i=1}^{n_x}\left(\frac{1}{(n_x-1)|S_{k,n_y}|}\sum_{j\neq i}\sum_{\sigma\in S_{k,n_y}}[\![X_j\notin\{Y_{\sigma(1)},\ldots,Y_{\sigma(k)}\}]\!]-\hat{\theta}(k)\right)^2. \tag{28}$$

Note that removing a color from the x-data which would otherwise add to $Q(j)$, decrements this quantity by one unit. Let Q_i denote the Q-statistics associated with the data when observation X_i from urn-x is removed from the sample. Note that as each draw from urn-x contributes to exactly one $Q(j)$, $Q_i(j)=Q(j)$ for all j except for some $j_i^\star$ where $Q_i(j_i^\star)=Q(j_i^\star)-1$. We have therefore that

$$\begin{aligned}
S_x^2(k) &= \frac{n_x-1}{n_x}\sum_{i=1}^{n_x}\left(\sum_{j=0}^{n_y-k}\frac{\binom{n_y-j}{k}Q_i(j)}{(n_x-1)|S_{k,n_y}|}-\sum_{j=0}^{n_y-k}\frac{\binom{n_y-j}{k}Q(j)}{n_x|S_{k,n_y}|}\right)^2,\\
&= \frac{n_x-1}{n_x}\sum_{i=1}^{n_x}\left(\sum_{j=0}^{n_y-k}\frac{\binom{n_y-j}{k}Q(j)}{n_x(n_x-1)|S_{k,n_y}|}-\frac{\binom{n_y-j_i^\star}{k}}{|S_{k,n_y}|(n_x-1)}\right)^2,\\
&= \frac{1}{n_x(n_x-1)}\sum_{i=1}^{n_x}\left(\hat{\theta}(k)-\frac{\binom{n_y-j_i^\star}{k}}{|S_{k,n_y}|}\right)^2.
\end{aligned}$$

Since there are $Q(j)$ draws from urn-x which contribute to $Q(j)$, the above sum may be now rewritten in the form given in (7). □

Proof of equation (9). Similarly, $S_y^2(k)$ corresponds to the jackknife summed over each possible deletion of a single y-data, which is more precisely given by

$$S_y^2(k) = \frac{n_y-1}{n_y}\sum_{r=1}^{n_y}\left(\frac{1}{n_x|S_r|}\sum_{i=1}^{n_x}\sum_{\sigma\in S_r}[\![X_i\notin\{Y_{\sigma(1)},\ldots,Y_{\sigma(k)}\}]\!]-\hat{\theta}(k)\right)^2, \tag{29}$$

where S_r is the set of one-to-one functions from $\{1,\ldots,k\}$ into $\{1,\ldots,n_y\}\setminus\{r\}$.

Recall that $M(i,j)$ is the number of colors seen i times in draws from urn-x and j times in draws from urn-y, giving that $\sum_i iM(i,j)=Q(j)$.

Fix $1\le r\le n_y$ and suppose that Y_r is of a color that contributes to $M(i_r^\star,j_r^\star)$, for some $i_r^\star,j_r^\star$. Removing Y_r from the data decrements $M(i_r^\star,j_r^\star)$ and increments $M(i_r^\star,j_r^\star-1)$ by one unit. Proceeding

similarly as in the case for $S_x^2(k)$, if M_r is used to denote the M-statistics when observation Y_r is removed from sample-y, then

$$\begin{aligned} S_y^2(k) &= \frac{n_y-1}{n_y}\sum_{r=1}^{n_y}\left(\sum_{j=0}^{n_y-k-1}\frac{\binom{n_y-j-1}{k}}{n_x|S_r|}\sum_{i=1}^{n_x} iM_r(i,j)-\hat{\theta}(k)\right)^2, \\ &= \frac{n_y-1}{n_y}\sum_{r=1}^{n_y}\left(i_r^\star\frac{\binom{n_y-j_r^\star}{k}}{n_x|S_r|}-i_r^\star\frac{\binom{n_y-j_r^\star-1}{k}}{n_x|S_r|}+\sum_{j=0}^{n_y-k-1}\frac{\binom{n_y-j_r^\star-1}{k}}{n_x|S_r|}Q(j)-\hat{\theta}(k)\right)^2, \\ &= \frac{n_y-1}{n_y}\sum_{r=1}^{n_y}\Big(i_r^\star\big(c_{j_r^\star-1}(k)-c_{j_r^\star}(k)\big)+\hat{\theta}_y(k)-\hat{\theta}(k)\Big)^2, \end{aligned}$$

where $c_j(k)$ and $\hat{\theta}_y(k)$ are as defined in (10) and (11). Noting that for each i there are j draws from urn-y that contribute to $M(i,j)$, the form in (9) follows. □

In what follows, we specialize the coefficients in (23) and (24) to the kernel function of dissimilarity, $h(x_1,y_1,\ldots,y_k) := [\![x_1 \notin \{y_1,\ldots,y_k\}]\!]$. From now on, for each $j \geq 0$ and $k \geq 1$, define

$$\xi_{0,j}(k) := \mathbb{V}(\mathbb{P}(X_1 \notin \{Y_1,\ldots,Y_k\}|Y_1,\ldots,Y_j)); \tag{30}$$
$$\xi_{1,j}(k) := \mathbb{V}(\mathbb{P}(X_1 \notin \{Y_1,\ldots,Y_k\}|X_1,Y_1,\ldots,Y_j)). \tag{31}$$

Above it is understood that the sigma-field generated by $(Y_1,\ldots,Y_j)$ when $j=0$ is $\{\emptyset,\Omega\}$; in particular, $\xi_{0,0}(k)=0$, for all $k \geq 1$.

The following asymptotic properties of $\xi_{i,j}(k)$ are useful in the remaining proofs.

Lemma 14. *Assume that conditions (a)-(c) are satisfied and define* $c := \min_{i\in(I_x\cap I_y)}\mathbb{P}_y(i)$. *It follows that* $0<c<1$ *and* $0<\theta(\infty)<1$. *Furthermore*

$$\xi_{1,k}(k)-\xi_{1,0}(k) = \Theta\big((1-c)^k\big); \tag{32}$$
$$\xi_{1,0}(k) = \theta(\infty)\big(1-\theta(\infty)\big)+\Theta\big((1-c)^k\big); \tag{33}$$
$$\xi_{0,k}(k) = O\big((1-c)^k\big). \tag{34}$$

Proof. Observe that conditions (a)-(b) imply that $0<c<1$. In addition, condition (b) implies that $\theta(\infty)<1$, whereas condition (c) implies that $\theta(\infty)>0$.

Next, consider the set

$$I^\star := \big\{i \in (I_x\cap I_y) \text{ such that } \mathbb{P}_y(i)=c\big\},$$

i.e. $I^\star$ is the set of rarest colors in urn-y which are also in urn-x. Also note that

$$\theta(k) = \theta(\infty)+\sum_{i\in(I_x\cap I_y)}\mathbb{P}_x(i)\big(1-\mathbb{P}_y(i)\big)^k. \tag{35}$$

As an intermediate step before showing (32), we prove that

$$\xi_{1,j}(k) = \theta(2k-j)-\theta^2(k). \tag{36}$$

For this, first observe that

$$\mathbb{P}(X_1 \notin \{Y_1,\ldots,Y_k\} \mid X_1,Y_1,\ldots,Y_j) = [\![X_1 \notin \{Y_1,\ldots,Y_j\}]\!](1-\mathbb{P}_y(X_1))^{k-j}.$$

Hence

$$\begin{aligned}\mathbb{E}\big(\mathbb{P}(X_1 \notin \{Y_1,\ldots,Y_k\} \mid X_1, Y_1, \ldots, Y_j)^2\big) &= \mathbb{E}\big([\![X_1 \notin \{Y_1,\ldots,Y_j\}]\!](1-\mathbb{P}_y(X_1))^{2k-2j}\big),\\ &= \mathbb{E}\big(\mathbb{P}(X_1 \notin \{Y_1,\ldots,Y_{2k-j}\} \mid X_1, Y_1, \ldots, Y_j)\big),\\ &= \theta(2k-j),\end{aligned}$$

from which (36) now easily follows.

To show (32) note that (36) implies

$$\begin{aligned}\xi_{1,k}(k) - \xi_{1,0}(k) &= \theta(k) - \theta(2k),\\ &= \sum_{i\in I_x} \mathbb{P}_x(i)(1-\mathbb{P}_y(i))^k(1-(1-\mathbb{P}_y(i))^k),\\ &= \sum_{i\in (I_x\cap I_y)} \mathbb{P}_x(i)(1-\mathbb{P}_y(i))^k(1-(1-\mathbb{P}_y(i))^k),\\ &= (1-c)^k \sum_{i\in I^\star} \mathbb{P}_x(i) + o\left((1-c)^k\right),\end{aligned}$$

which establishes (32).

Now note that

$$\begin{aligned}\xi_{1,0}(k) &= \theta(2k) - \theta^2(k),\\ &= \sum_{i\in I_x} \mathbb{P}_x(i)\big(1-\mathbb{P}_y(i)\big)^{2k} - \left(\sum_{i\in I_x} \mathbb{P}_x(i)\big(1-\mathbb{P}_y(i)\big)^k\right)^2,\\ &= \theta(\infty) - \theta^2(\infty) - 2\theta(\infty)(1-c)^k\left(\sum_{i\in I^\star}\mathbb{P}_x(i)\right) + o(1-c)^k,\end{aligned}$$

which establishes (33).

Next we show (34), which we note gives more precise information than (27). Consider the random variable T defined as the smallest $n \geq 1$ such that $(I_x \cap I_y) \subset \{Y_1,\ldots,Y_n\}$. We may bound the probability of T being large by $\mathbb{P}(T > k) \leq n(1-c)^k$, where $n := |I_x \cap I_y|$ is finite because of condition (a). On the other hand, note that

$$\mathbb{P}\big(X_1 \notin \{Y_1,\ldots,Y_k\} \mid Y_1,\ldots,Y_k\big) = 1 - \mathbb{P}_x(\{Y_1,\ldots,Y_k\}).$$

Define $W_k := 1 - \mathbb{P}_x(\{Y_1,\ldots,Y_k\}) - \theta(k)$ and observe that, over the event $T \leq k$, $W_k = \theta(\infty) - \theta(k)$. Since $|W_k| \leq 1$, we obtain that

$$\begin{aligned}\xi_{0,k}(k) &= \mathbb{E}\Big(\mathbb{E}\big(W_k^2 \mid T > k\big)\Big)\cdot\mathbb{P}(T>k) + \mathbb{E}\Big(\mathbb{E}\big(W_k^2 \mid T \leq k\big)\Big)\cdot\mathbb{P}(T\leq k),\\ &\leq \mathbb{P}(T>k) + \mathbb{E}\Big(\mathbb{E}\big(W_k^2 \mid T\leq k\big)\Big),\\ &\leq n(1-c)^k + \big(\theta(\infty)-\theta(k)\big)^2.\end{aligned}$$

The identity in equation (34) is now a direct consequence of (35). □

Our next goal is to show Theorems 4 and 5. To do so we rely on the method of projection by Grams and Serfling [17]. This approach approximates $\hat{\theta}(k)$ by the random variable

$$\hat{\theta}_P(k) := \theta(k) + \sum_{i=1}^{n_x}(\mathbb{E}(\hat{\theta}(k)|X_i) - \theta(k)) + \sum_{j=1}^{n_y}(\mathbb{E}(\hat{\theta}(k)|Y_j) - \theta(k)).$$

The projection is the best approximation in terms of mean squared error to $\hat{\theta}(k)$ that is a linear combination of individual functions of each datapoint.

Under the stated conditions, $\hat{\theta}_P(k)$ is the sum of two independent sums of non-degenerate i.i.d. random variables and therefore satisfies the hypotheses of the classical central limit theorem. The variance of the projection is easier to analyze and estimate than the U-statistic directly, which is relevant in establishing consistency for the jackknife estimation of variance.

Let

$$R(k) := \hat{\theta}(k) - \hat{\theta}_P(k),$$

be the remainder of $\hat{\theta}(k)$ that is not accounted for by its projection. When $R(k)$ is small relative to $\hat{\theta}_P(k)$, $\hat{\theta}(k)$ is mostly explained by $\hat{\theta}_P(k)$ in relative terms.

The next lemma summarizes results about the asymptotic properties of $R(k)$, particularly with relation to the scale of $\hat{\theta}_P(k)$ as given by its variance.

Lemma 15. *We have that*

$$\mathbb{V}(\hat{\theta}_P(k)) = n_x^{-1}\xi_{1,0}(k) + k^2 n_y^{-1}\xi_{0,1}(k), \tag{37}$$

$$\mathbb{E}(R^2(k)) = \mathbb{V}(\hat{\theta}(k)) - \mathbb{V}(\hat{\theta}_P(k)). \tag{38}$$

Under assumptions (a)-(c), for a fixed $k \geq 1$, we have that

$$\lim_{n_x, n_y \to \infty} \left| \frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))} - 1 \right| = 0. \tag{39}$$

Furthermore, under assumptions (a)-(d) we have that

$$\lim_{n_x, n_y \to \infty} \max_{k=1:n_y} \left| \frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))} - 1 \right| = 0; \tag{40}$$

$$\lim_{n_x, n_y \to \infty} \max_{k=1:n_y} \mathbb{P}\left(\left| \frac{\hat{\theta}(k) - \hat{\theta}_P(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}} \right| > \epsilon \right) = 0; \tag{41}$$

for all $\epsilon > 0$.

Proof. A direct calculation from the form given in (16) gives that

$$\mathbb{E}(\hat{\theta}(k)|X_i)) = n_x^{-1}\mathbb{P}(X_i \notin \{Y_1, \ldots, Y_k\}|X_i) + \left(1 - n_x^{-1}\right)\theta(k); \tag{42}$$

$$\mathbb{V}(\mathbb{E}(\hat{\theta}(k)|X_i)) = \frac{\xi_{1,0}(k)}{n_x^2};$$

$$\mathbb{E}(\hat{\theta}(k)|Y_i)) = k n_y^{-1}\mathbb{P}(X_i \notin \{Y_i, \ldots, Y_{k+i-1}\}|Y_i) + \left(1 - k n_y^{-1}\right)\theta(k); \tag{43}$$

$$\mathbb{V}(\mathbb{E}(\hat{\theta}(k)|Y_i)) = \frac{k^2\xi_{0,1}(k)}{n_y^2}.$$

As $\mathbb{V}(\hat{\theta}_P(k)) = n_x\mathbb{V}(\mathbb{E}(\hat{\theta}(k)|X_i)) + n_y\mathbb{V}(\mathbb{E}(\hat{\theta}(k)|Y_i))$, (37) follows.

To show (38), first observe that

$$\mathbb{E}(R^2(k)) = \mathbb{V}(R(k)) = \mathbb{V}(\hat{\theta}(k)) + \mathbb{V}(\hat{\theta}_P(k)) - 2\,\mathrm{Cov}(\hat{\theta}(k), \hat{\theta}_P(k)). \tag{44}$$

Next, using the definition of the projection, we obtain that

$$\begin{aligned}\mathrm{Cov}(\hat{\theta}(k),\hat{\theta}_P(k)) &= \sum_{i=1}^{n_x}\mathrm{Cov}(\hat{\theta}(k),\mathbb{E}(\hat{\theta}(k)|X_i))+\sum_{j=1}^{n_y}\mathrm{Cov}(\hat{\theta}(k),\mathbb{E}(\hat{\theta}(k)|Y_j)),\\ &= \sum_{i=1}^{n_x}\mathbb{V}(\mathbb{E}(\hat{\theta}(k)|X_i))+\sum_{j=1}^{n_y}\mathbb{V}(\mathbb{E}(\hat{\theta}(k)|Y_j)),\\ &= \mathbb{V}\left(\sum_{i=1}^{n_x}\mathbb{E}(\hat{\theta}(k)|X_i)+\sum_{j=1}^{n_y}\mathbb{E}(\hat{\theta}(k)|Y_j)\right),\\ &= \mathbb{V}(\hat{\theta}_P(k)),\end{aligned}$$

from which (38) follows, due to the identity in (44). Note that the last identity implies that $\hat{\theta}_P(k)$ and $R(k)$ are uncorrelated.

Before continuing, we note that (41) is a direct consequence of (38), (40) and Chebyshev's inequality [22]. To complete the proof of the lemma all reduces therefore to show (41) under conditions (a)-(d). Indeed, if $b>1$ and we let $k_n=\log_b(n_y)$ then due to the identities in (22) and (37) and Lemma 12, we obtain under (a)-(c) that

$$\mathbb{V}(\hat{\theta}(k))=\mathbb{V}(\hat{\theta}_P(k))+O\left(\frac{k^2}{n_xn_y}+\frac{k^4}{n_y^2}\right),$$

uniformly for all $k=1:k_n$, as $n_x,n_y\to\infty$. Since $\mathbb{V}(\hat{\theta}_P(k))>0$ for all $k\geq 1$, we have thus shown (39). Furthermore, note that $\xi_{1,0}(k)>0$ for all $k\geq 1$; in particular, due to (33) and conditions (a)-(d), we can assert that $\inf_{k\geq 1}\xi_{1,0}(k)>0$. Since $\xi_{0,1}(k)\geq 0$, the above identity together with the one in (37) let us conclude that

$$\max_{k=1:k_n}\left|\frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))}-1\right|=O\left(\frac{k_n^2}{n_y}+\frac{n_xk_n^4}{n_y^2}\right),$$

as $n_x,n_y\to\infty$. Because of condition (d), the big-O term above tends to 0. As a result:

$$\lim_{n_x,n_y\to\infty}\max_{k=1:k_n}\left|\frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))}-1\right|=0. \tag{45}$$

On the other hand, (38) implies that $\mathbb{V}(\hat{\theta}(k))\geq\mathbb{V}(\hat{\theta}_P(k))$. Hence, using (19) and (25) to bound from above the variance of the U-statistic, we obtain:

$$1\leq\frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))}\leq\frac{\xi_{0,k}(k)+\xi_{1,k}(k)/n_x}{k^2\xi_{0,1}(k)/n_y+\xi_{1,0}(k)/n_x}\leq n_x\,\frac{\xi_{0,k}(k)}{\xi_{1,0}(k)}+\frac{\xi_{1,k}(k)}{\xi_{1,0}(k)}=1+n_x\cdot O\big((1-c)^k\big),$$

as $k\to\infty$, where for the last identity we have used (32) and (34). Since $n_x=\Theta(n_y)$, it follows from the above identity that

$$\max_{k=k_n:n_y}\left|\frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))}-1\right|=O\big(n_y(1-c)^{k_n}\big)=O\big(n_y^{1+\log_b(1-c)}\big).$$

In particular, if the base-b in the logarithm is selected to satisfy that $1<b<1/(1-c)$, then

$$\lim_{n_x,n_y\to\infty}\max_{k=k_n:n_y}\left|\frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))}-1\right|=0. \tag{46}$$

The identities in equation (45) and (46) show (41), which completes the proof of the lemma. □

Proof of Theorem 5. For a fixed k, note that $\hat{\theta}_P(k)$ is the sum of two independent sums of non-degenerate i.i.d. random variables and thus,

$$(\hat{\theta}_P(k) - \theta(k))/\sqrt{\mathbb{V}(\hat{\theta}_P(k))}$$

is asymptotically a standard Normal random variable as $n_x, n_y \to \infty$ by the classical Central Limit Theorem. We would like to show however that this convergence also applies if we let k vary with n_x and n_y. We do so using the Berry-Esseen inequality [23]. Motivated by this we define the random variables

$$X_i'(k) := \frac{\mathbb{E}(\hat{\theta}(k)|X_i) - \theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}};$$
$$Y_j'(k) := \frac{\mathbb{E}(\hat{\theta}(k)|Y_j) - \theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}}.$$

Note that $\mathbb{E}(X_i'(k)) = \mathbb{E}(Y_j'(k)) = 0$, and that

$$\sum_{i=1}^{n_x} \mathbb{E}(|X_i'(k)|^2) + \sum_{j=1}^{n_y} \mathbb{E}(|Y_j'(k)|^2) = 1.$$

We need to show that

$$\sum_{i=1}^{n_x} \mathbb{E}(|X_i'(k)|^3) + \sum_{j=1}^{n_y} \mathbb{E}(|Y_j'(k)|^3) = o(1), \tag{47}$$

uniformly for $k = 1 : n_y$, as $n_x, n_y \to \infty$.

Note that from (42) and (43),

$$\left|\mathbb{E}(\hat{\theta}(k)|X_i) - \theta(k)\right|^3 = \frac{|\mathbb{P}(X_i \notin \{Y_1, \ldots, Y_k\}|X_i) - \theta(k)|^3}{n_x^3};$$
$$\left|\mathbb{E}(\hat{\theta}(k)|Y_i) - \theta(k)\right|^3 = \frac{k^3|\mathbb{P}(X_1 \notin \{Y_i, \ldots, Y_{k+1-i}\}|Y_i) - \theta(k)|^3}{n_y^3}.$$

Let

$$\eta_{1,0}(k) := \mathbb{E}|\mathbb{P}(X_i \notin \{Y_1, \ldots, Y_k\}|X_i) - \theta(k)|^3;$$
$$\eta_{0,1}(k) := \mathbb{E}|\mathbb{P}(X_1 \notin \{Y_i, \ldots, Y_{k+1-i}\}|Y_i) - \theta(k)|^3.$$

It follows from (37) that

$$\sum_{i=1}^{n_x} \mathbb{E}(|X_i'(k)|^3) + \sum_{j=1}^{n_y} \mathbb{E}(|Y_j'(k)|^3) = \frac{\eta_{1,0}(k)/n_x^2 + k^3\eta_{0,1}(k)/n_y^2}{\left(\mathbb{V}(\hat{\theta}_P(k))\right)^{3/2}} = \frac{\eta_{1,0}(k)/n_x^2 + k^3\eta_{0,1}(k)/n_y^2}{\left(\xi_{1,0}(k)/n_x + k^2\xi_{0,1}(k)/n_y\right)^{3/2}}.$$

But note that $0 \le \eta_{0,1}(k) \le \xi_{0,1}(k)$. Since, according to Lemma 14, $\xi_{0,1}(k)$ decreases exponentially fast, we obtain

$$k^3\eta_{0,1}(k)/n_y^2 = O(n_y^{-2}),$$

uniformly for all $k = 1 : n_y$, as $n_y \to \infty$. On the other hand, $0 \le \eta_{1,0}(k) \le \xi_{1,0}(k) \le 1$. Furthermore, (33) implies that $\inf_{k\ge1} \xi_{1,0}(k) > 0$. Since $n_x = \Theta(n_y)$, for some finite constant $C > 0$ we find that

$$\sum_{i=1}^{n_x} \mathbb{E}(|X_i'(k)|^3) + \sum_{j=1}^{n_y} \mathbb{E}(|Y_j'(k)|^3) \le C\frac{1/n_x^2 + 1/n_y^2}{\left(\inf\limits_{k\ge1} \xi_{1,0}(k)/n_x\right)^{3/2}} = O\left(\frac{1}{\sqrt{n_x}}\right),$$

which shows (47).

The above establishes convergence in distribution of $(\hat{\theta}_P(k)-\theta(k))/\sqrt{\mathbb{V}(\hat{\theta}_P(k))}$ to a standard normal random variable uniformly for $k=1:n_y$, as $n_x, n_y \to \infty$. The end of the proof is an adaptation of the proof of Slutsky's Theorem [21]. Indeed, note that

$$\mathbb{P}\left(\frac{\hat{\theta}(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}} \leq t\right) = \mathbb{P}\left(\frac{\hat{\theta}_P(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}} + \frac{\hat{\theta}(k)-\hat{\theta}_P(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}} \leq t\sqrt{\frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))}}\right). \tag{48}$$

From this identity, it follows for any fixed $\epsilon > 0$ that

$$\mathbb{P}\left(\frac{\hat{\theta}(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}} \leq t\right) \leq \mathbb{P}\left(\frac{\hat{\theta}_P(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}} \leq t\sqrt{\frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))}} + \epsilon\right) + \mathbb{P}\left(\left|\frac{\hat{\theta}(k)-\hat{\theta}_P(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}}\right| \geq \epsilon\right).$$

The first term on the right-hand side of the above inequality can be made as close to $\mathbb{P}[Z \leq t+\epsilon]$ as wanted, uniformly for $k=1:n_y$, as $n_x, n_y \to \infty$, because of (40). On the other hand, the second term tends to 0 uniformly for $k=1:n_y$ because of (41). Letting $\epsilon \to 0^+$, shows that

$$\limsup_{n_x,n_y\to\infty} \max_{k=1:n_y} \mathbb{P}\left(\frac{\hat{\theta}(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}} \leq t\right) \leq \mathbb{P}[Z \leq t].$$

Similarly, using (48), we have:

$$\mathbb{P}\left(\frac{\hat{\theta}(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}} \leq t\right) \geq \mathbb{P}\left(\frac{\hat{\theta}_P(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}} \leq t\sqrt{\frac{\mathbb{V}(\hat{\theta}(k))}{\mathbb{V}(\hat{\theta}_P(k))}} - \epsilon\right) - \mathbb{P}\left(\left|\frac{\hat{\theta}(k)-\hat{\theta}_P(k)}{\sqrt{\mathbb{V}(\hat{\theta}_P(k))}}\right| \geq \epsilon\right),$$

and a similar argument as before shows now that

$$\liminf_{n_x,n_y\to\infty} \max_{k=1:n_y} \mathbb{P}\left(\frac{\hat{\theta}(k)-\theta(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}} \leq t\right) \geq \mathbb{P}[Z \leq t],$$

which completes the proof of the theorem. □

We finally show Theorem 4, for which we first show the following result.

Lemma 16. *Let $S^i_{k,n}$ be the set of one-to-one functions from $\{1,\ldots,k\}$ into $\{1,\ldots,n\}/\{i\}$. Consider the kernel $h(x_1,y_1,\ldots,y_k) := [\![x_1 \notin \{y_1,\ldots,y_k\}]\!]$, and define*

$$\hat{\theta}^i_x(k) := \frac{1}{|S_{k,n_y}|} \sum_{\sigma\in S_{k,n_y}} \frac{1}{n_x-1} \sum_{j=1,j\neq i}^{n_x} h(X_j, Y_{\sigma(1)},\ldots,Y_{\sigma(k)}); \tag{49}$$

$$\hat{\theta}^{i\prime}_x(k) := \frac{1}{|S_{k,n_y}|} \sum_{\sigma\in S_{k,n_y}} h(X_i, Y_{\sigma(1)},\ldots,Y_{\sigma(k)}); \tag{50}$$

$$\hat{\theta}^i_y(k) := \frac{1}{|S^i_{k,n_y}|} \sum_{\sigma\in S^i_{k,n_y}} \frac{1}{n_x} \sum_{j=1}^{n_x} h(X_j, Y_{\sigma(1)},\ldots,Y_{\sigma(k)}); \tag{51}$$

$$\hat{\theta}^{i\prime}_y(k) := \frac{1}{|S^i_{k-1,n_y}|} \sum_{\sigma\in S^i_{k-1,n_y}} \frac{1}{n_x} \sum_{j=1}^{n_x} h(X_j, Y_i, Y_{\sigma(1)},\ldots,Y_{\sigma(k-1)}). \tag{52}$$

Then, for each $k \geq 1$ and $\epsilon > 0$,

$$\lim_{n_x,n_y\to\infty} \mathbb{P}\left(\left|\sum_{i=1}^{n_x} \frac{\left(\hat{\theta}_x^i(k) - \hat{\theta}_x^{i\prime}(k)\right)^2}{n_x} - \xi_{1,0}(k)\right| > \epsilon\right) = 0; \tag{53}$$

$$\lim_{n_x,n_y\to\infty} \mathbb{P}\left(\left|\sum_{i=1}^{n_y} \frac{\left(\hat{\theta}_y^i(k) - \hat{\theta}_y^{i\prime}(k)\right)^2}{n_y} - \xi_{0,1}(k)\right| > \epsilon\right) = 0. \tag{54}$$

Proof. Fix $k \geq 1$. We first use a result by Sen [24] to show that, for each $i \geq 1$:

$$\lim_{n_x,n_y\to\infty} \hat{\theta}_x^i(k) = \theta(k); \tag{55}$$

$$\lim_{n_x,n_y\to\infty} \hat{\theta}_x^{i\prime}(k) = \mathbb{E}(h(X_i, Y_1, \ldots, Y_k)|X_i); \tag{56}$$

$$\lim_{n_x,n_y\to\infty} \hat{\theta}_y^i(k) = \theta(k); \tag{57}$$

$$\lim_{n_x,n_y\to\infty} \hat{\theta}_y^{i\prime}(k) = \mathbb{E}(h(X_1, Y_i, \ldots, Y_{i+k-1})|Y_i); \tag{58}$$

in an almost sure sense. Indeed, assume without loss of generality that $i = 1$. As the kernel functions found in (49) and (51) are bounded, the hypotheses of Theorem 1 in [24] are satisfied, from which (55) and (57) are immediate. Similarly, because X_1 and Y_1 are discrete random variables, (56) and (58) also follow from [24] .

Define

$$U(k) := \sum_{i=1}^{n_x} \frac{(\hat{\theta}_x^i(k) - \hat{\theta}_x^{i\prime}(k))^2}{n_x}; \tag{59}$$

$$V(k) := \sum_{i=1}^{n_y} \frac{(\hat{\theta}_y^i(k) - \hat{\theta}_y^{i\prime}(k))^2}{n_y}; \tag{60}$$

and observe that

$$\mathbb{V}(U(k)) = \frac{\mathbb{V}\left((\hat{\theta}_x^1(k) - \hat{\theta}_x^{1\prime}(k))^2\right)}{n_x} + \frac{n_x - 1}{n_x}\mathrm{Cov}\left((\hat{\theta}_x^1(k) - \hat{\theta}_x^{1\prime}(k))^2, (\hat{\theta}_x^2(k) - \hat{\theta}_x^{2\prime}(k))^2\right);$$

$$\mathbb{V}(V(k)) = \frac{\mathbb{V}\left((\hat{\theta}_y^1(k) - \hat{\theta}_y^{1\prime}(k))^2\right)}{n_y} + \frac{n_y - 1}{n_y}\mathrm{Cov}\left((\hat{\theta}_y^1(k) - \hat{\theta}_y^{1\prime}(k))^2, (\hat{\theta}_y^2(k) - \hat{\theta}_y^{2\prime}(k))^2\right).$$

Furthermore, due to (55)-(58), we have that

$$\lim_{n_x,n_y\to\infty} (\hat{\theta}_x^i(k) - \hat{\theta}_x^{i\prime}(k))^2 = (\theta(k) - \mathbb{E}(h(X_i, Y_1, \ldots, Y_k)|X_i))^2; \tag{61}$$

$$\lim_{n_x,n_y\to\infty} (\hat{\theta}_y^i(k) - \hat{\theta}_y^{i\prime}(k))^2 = (\theta(k) - \mathbb{E}(h(X_1, Y_i, \ldots, Y_{i+k-1})|Y_i))^2. \tag{62}$$

But note that, for $i \neq j$, $(\theta(k) - \mathbb{E}(h(X_i, Y_1, \ldots, Y_k)|X_i))^2$ and $(\theta(k) - \mathbb{E}(h(X_j, Y_1, \ldots, Y_k)|X_j))^2$ are independent and hence uncorrelated. Similarly, the random variables $(\theta(k) - \mathbb{E}(h(X_1, Y_i, \ldots, Y_{i+k-1})|Y_i))^2$ and $(\theta(k) - \mathbb{E}(h(X_1, Y_j, \ldots, Y_{j+k-1})|Y_j))^2$ are independent. Since $|\hat{\theta}_x^i(k) - \hat{\theta}_x^{i\prime}(k)| \leq 1$ and $|\hat{\theta}_y^i(k) - \hat{\theta}_y^{i\prime}(k)| \leq 1$, it follows from (61) and (62), and the Bounded Convergence Theorem [22] that

$$\mathbb{V}(U(k)) = o(1); \tag{63}$$

$$\mathbb{V}(V(k)) = o(1); \tag{64}$$

as $n_x, n_y \to \infty$.

Finally, by (30) and (31) it follows that

$$\xi_{1,0}(k) = \mathbb{E}\left(\theta(k) - \mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|X_1)\right)^2;$$
$$\xi_{0,1}(k) = \mathbb{E}\left(\theta(k) - \mathbb{E}(h(X_1, Y_1, \ldots, Y_k)|Y_1)\right)^2.$$

In particular, again by the Bounded Convergence Theorem, we have that $\lim_{n_x,n_y\to\infty} \mathbb{E}\left(U(k)\right) = \xi_{1,0}(k)$ and $\lim_{n_x,n_y\to\infty} \mathbb{E}\left(V(k)\right) = \xi_{0,1}(k)$. Since

$$U(k) - \xi_{1,0}(k) = \big(\mathbb{E}(U(k)) - \xi_{1,0}(k)\big) + \big(U(k) - \mathbb{E}(U(k))\big);$$
$$V(k) - \xi_{0,1}(k) = \big(\mathbb{E}(V(k)) - \xi_{0,1}(k)\big) + \big(V(k) - \mathbb{E}(V(k))\big);$$

the lemma is now a direct consequence of (63) and (64), and Theorem 1.5.4 of Durrett [22]. □

Proof of Theorem 4. Fix $k \geq 1$. Using (16) we have that

$$\hat{\theta}(k) = \left(1 - \frac{1}{n_x}\right)\hat{\theta}_x^i(k) + \frac{1}{n_x}\hat{\theta}_x^{i\prime}(k);$$
$$\hat{\theta}_x^i(k) - \hat{\theta}(k) = \frac{1}{n_x}\left(\hat{\theta}_x^i(k) - \hat{\theta}_x^{i\prime}(k)\right);$$
$$\hat{\theta}(k) = \left(1 - \frac{k}{n_y}\right)\hat{\theta}_y^i(k) + \frac{k}{n_y}\hat{\theta}_y^{i\prime}(k);$$
$$\hat{\theta}_y^i(k) - \hat{\theta}(k) = \frac{k}{n_y}\left(\hat{\theta}_y^i(k) - \hat{\theta}_y^{i\prime}(k)\right).$$

It follows by (28) and (29) that

$$S_x^2(k) = \frac{n_x - 1}{n_x} \cdot \frac{U(k)}{n_x}, \tag{65}$$
$$S_y^2(k) = \frac{n_y - 1}{n_y} \cdot \frac{k^2 V(k)}{n_y}, \tag{66}$$

where $U(k)$ and $V(k)$ are as in (59) and (60), respectively. Furthermore, observe that

$$S^2(k) = S_x^2(k) + S_y^2(k) = \frac{n_x - 1}{n_x} \cdot \frac{U(k)}{n_x} + \frac{n_y - 1}{n_y} \cdot \frac{k^2 V(k)}{n_y}.$$

In particular, due to (37), we obtain that

$$\left|\frac{S^2(k)}{\mathbb{V}(\hat{\theta}_P(k))} - 1\right| \leq \frac{|U(k) - \xi_{1,0}(k)|}{\xi_{1,0}(k)} + \frac{|V(k) - \xi_{0,1}(k)|}{\xi_{0,1}(k)} + \frac{|U(k)|}{n_x\,\xi_{1,0}(k)} + \frac{|V(k)|}{n_y\,\xi_{0,1}(k)}.$$

By Lemma 16, $U(k)$ converges in probability to $\xi_{1,0}(k)$, while similarly $V(k)$ converges in probability to $\xi_{0,1}(k)$; in particular, the first two terms on the right-hand side of the inequality converge to 0 in probability. Since $|U(k)| \leq 1$ and $|V(k)| \leq 1$, the same can be said about the last two terms of the inequality. Consequently, $S^2(k)/\mathbb{V}(\hat{\theta}_P(k))$ converges to 1 in probability, as $n_x, n_y \to \infty$. As stated in (39), however, conditions (a)-(c) imply that $\mathbb{V}(\hat{\theta}_P(k))$ and $\mathbb{V}(\hat{\theta}(k))$ are asymptotically equivalent as $n_x, n_y \to \infty$, from which the theorem follows. □

Acknowledgments

We thank Rob Knight for insightful discussions and comments about this manuscript, and Antonio Gonzalez for providing processed OTU tables from the Human Microbiome Project.

References

1. Lladser ME, Gouet R, Reeder J (2011) Extrapolation of urn models via poissonization: Accurate measurements of the microbial unknown. PLoS ONE 6: e21105.
2. Good IJ (1953) The population frequencies of species and the estimation of population parameters. Biometrika 40: 237-264.
3. Esty WW (1983) A Normal limit law for a nonparametric estimator of the coverage of a sample. Ann Stat 11: 905-912.
4. Mao CX (2004) Predicting the conditional probability of finding a new class. J Am Stat Assoc 99: 1108-1118.
5. Lijoi A, Mena RH, Prünster I (2007) Bayesian nonparametric estimation of the probability of discovering new species. Biometrika 94: 769-786.
6. Robbins HE (1968) Estimating the total probability of the unobserved outcomes of an experiment. Ann Math Statist 39: 256-257.
7. Starr N (1979) Linear estimation of discovering a new species. Ann Stat 7: 644-652.
8. Halmos PR (1946) The theory of unbiased estimation. Ann Math Statist 17: 34-43.
9. Clayton MK, Frees EM (1987) Linear estimation of discovering a new species. J Am Stat Assoc 82: 305-311.
10. Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71: 8228-8235.
11. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R (2011) UniFrac: An effective distance metric for microbial community comparison. ISME J 5: 169-172.
12. Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249: 505-510.
13. Huttenhower C, Gevers D, Sathirapongsasuti JF, Segata N, Earl AM, et al. (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207–214.
14. Hoeffding W (1948) A class of statistics with asymptotically normal distribution. Ann Math Statist 19: 293-325.
15. Efron B, Stein C (1981) The jackknife estimate of variance. Ann Stat 9: 586-596.
16. Shao J, Wu C (1989) A general theory for jackknife variance estimation. Ann Stat 17: 1176-1197.
17. Grams WF, Serfling RJ (1973) Convergence rate for U-statistics and related statistics. Ann Stat 1: 153-160.
18. Hampton JD (2012) Dissimilarity and Optimal Sampling in Urn Ensemble Model. Ph.D. thesis, University of Colorado, Boulder, Colorado.

19. Ahmad AA (1980) On the Berry-Esseen theorem for random U-statistics. Ann Stat 8: 1395-1398.

20. Callaert H, Janssen P (1978) The Berry-Esseen theorem for U-statistics. Ann Stat 6: 417-421.

21. Slutsky E (1925) Über stochastische asymptoten und grenzwerte. Metron 5: 3-89.

22. Durrett R (2010) Probability: Theory and Examples. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. URL `http://books.google.com/books?id=evbGTPhuvSoC`.

23. Shevtsova IG (2010) An improvement of convergence rate estimates in the Lyapunov theorem. Doklady Mathematics 82: 862-864.

24. Sen PK (1977) Almost sure convergence of generalized U-statistics. Ann Probab 5: pp. 287-290.

Figure Legends

Figure 1. Dissimilarity estimates. Heat map of $\hat{\theta}(n_y)$ sorted by site location metadata. Here, the x-axis denotes the sample from the environment corresponding to urn-x, and similarly for the y-axis. The entries on the diagonal are set to zero.

Figure 2. Error estimates. Heat map of $S(n_y)$ sorted by site location metadata. Here, the x-axis also denotes the sample from the environment corresponding to urn-x, and similarly for the y-axis, and the entries on the diagonal are again set to zero.

Figure 3. Discrete derivative estimates. Heat map of $|\hat{\theta}(n_y) - \hat{\theta}(n_y - 1)|$, sorted by site location metadata, following the same conventions as in the previous figures.

Figure 4. Sequential estimation. Plots of $\hat{\theta}(k)$, with $k = 1 : n_y$, for three pairs of samples of the HMP data.

Tables

Table 1. HMP data. Summary of V35 16S rRNA data processed by Qiime into an OTU table.

Body Supersite	Body Subsite	Assigned Labels
Airways	Anterior Nares	1-5
	Throat	6-17
Gastrointestinal Tract	Stool	18-47
Oral	Attached/Keratinized Gingiva	48-59
	Buccal Mucosa	60-76
	Hard Palate	77-90
	Palatine Tonsils	91-112
	Saliva	113-122
	Subgingival Plaque	123-144
	Supragingival Plaque	145-167
	Tongue Dorsum	168-191
Skin	Left Antecubital Fossa	192-195
	Left Retroauricular Crease	196-217
	Right Antecubital Fossa	218-222
	Right Retroauricular Crease	223-242
Urogenital Tract	Mid Vagina	243-248
	Posterior Fornix	249-259
	Vaginal Introitus	260-266

Table 2. Sample comparisons. Summary of estimates for three pairs of samples of the HMP data.

Urn-x	Urn-y	n_x	n_y	$\hat{\theta}(n_y)$	$\hat{\theta}(n_y) - \hat{\theta}(n_y - 1)$	Regression Error	$S(n_y)$	$\hat{\rho}^{n_y}$
255	176	5054	6782	0.9998	0.0	0.0	1.9892×10^{-4}	0.0
200	139	12747	5739	0.0499	-1.6533×10^{-4}	6.8306×10^{-6}	1.9286×10^{-3}	0.9997
100	10	6206	8655	0.0324	-2.9416×10^{-6}	0.0438	2.2477×10^{-3}	0.5130