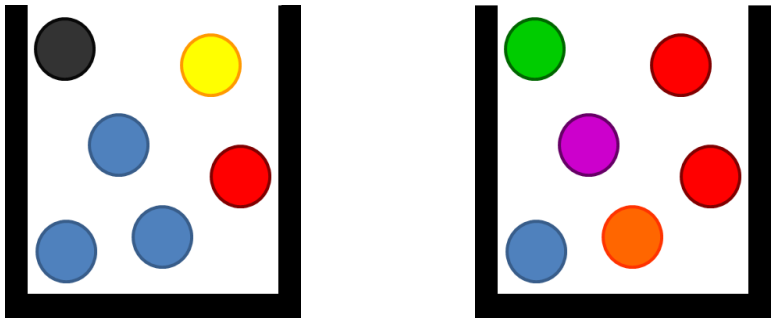
Estimation of Distribution Overlap for Urn Models

J. Hampton
Applied Mathematics
University of Colorado

work with
M. Lladser

Problem Description:

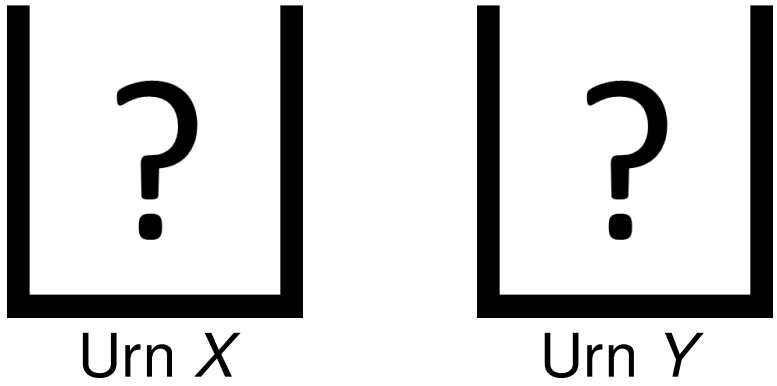
Consider two **urns** with **colored balls**:



Let $\mathbb{P}_X(i) \geq 0$ denote the **proportion** of color i in urn X and $\mathbb{P}_Y(i) \geq 0$ denote the **proportion** of color i in urn Y.

Problem Description:

Consider two **urns** with **colored balls**:

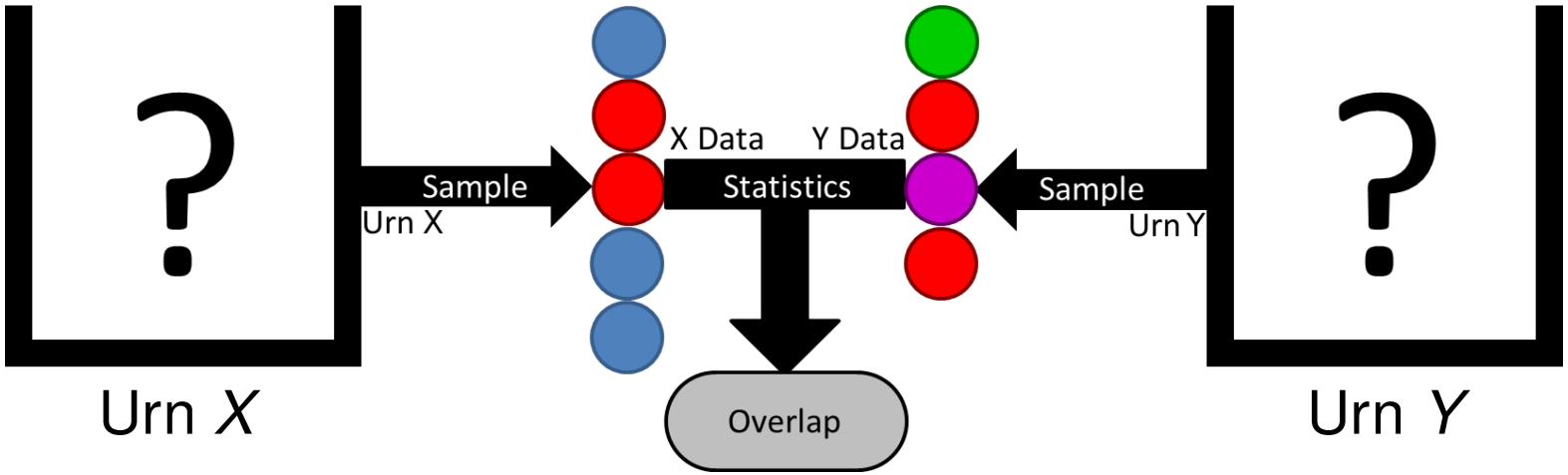


Let $\mathbb{P}_X(i) \geq 0$ denote the **proportion** of color i in urn X and $\mathbb{P}_Y(i) \geq 0$ denote the **proportion** of color i in urn Y.

Problem Description:

Goal:

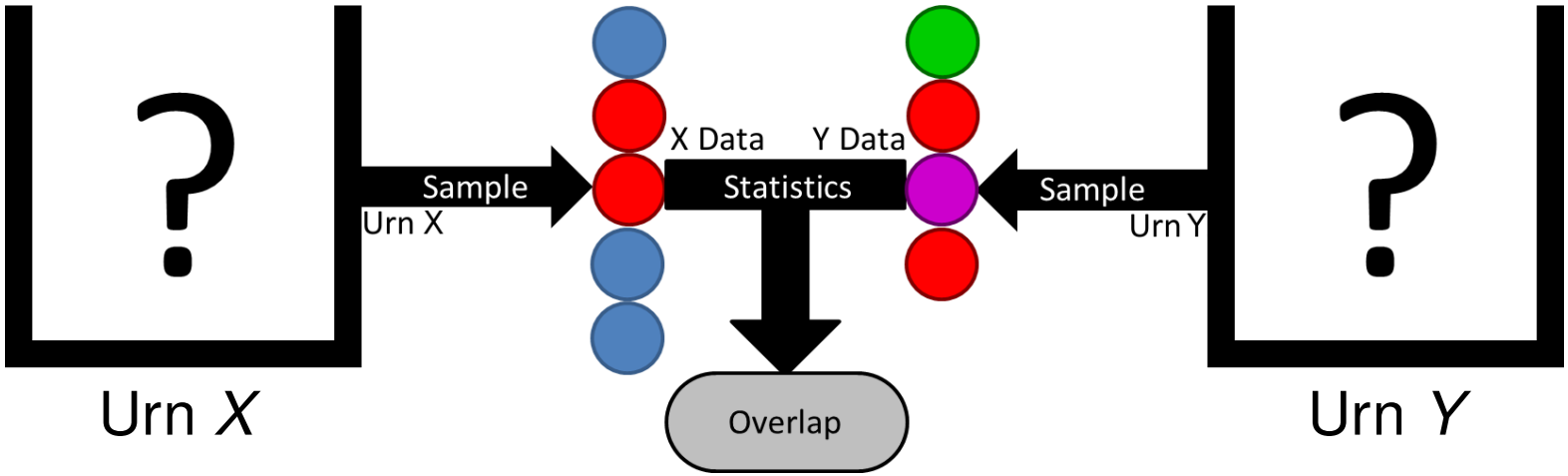
To estimate a measure of overlap based on **samples with replacement** from each urn.



Problem Description:

Goal:

To estimate a measure of overlap based on **samples with replacement** from each urn.

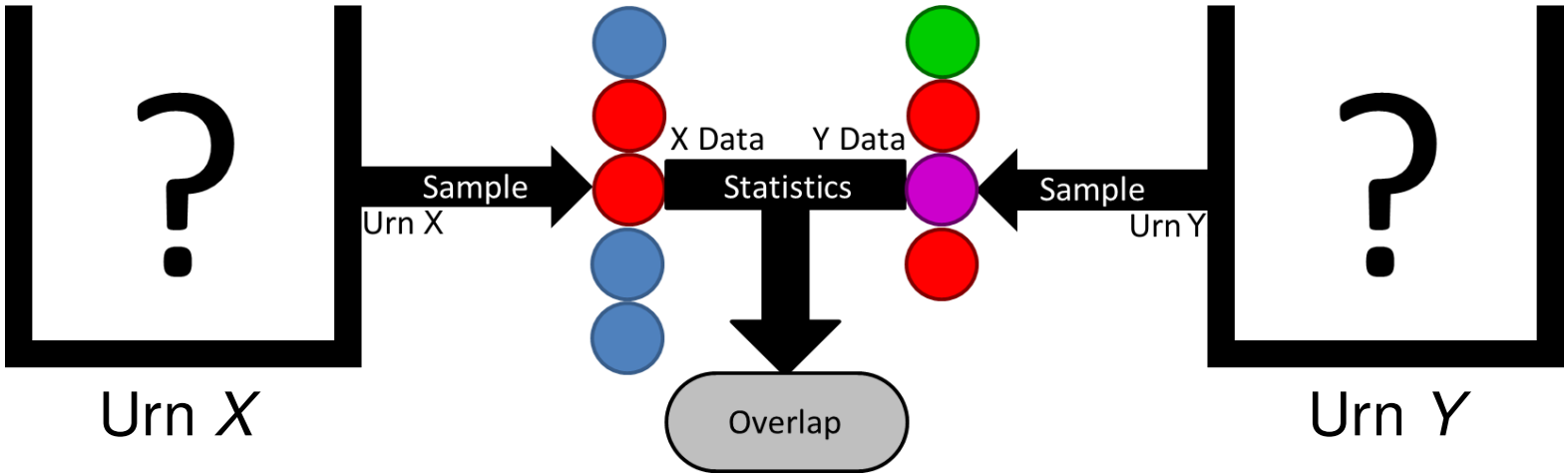


In particular, let $X_1, \cdots, X_{n_x}$ denote our draws from Urn X, and $Y_1, \cdots, Y_{n_y}$ denote our draws from Urn Y.

Measure of overlap:

Overlap:

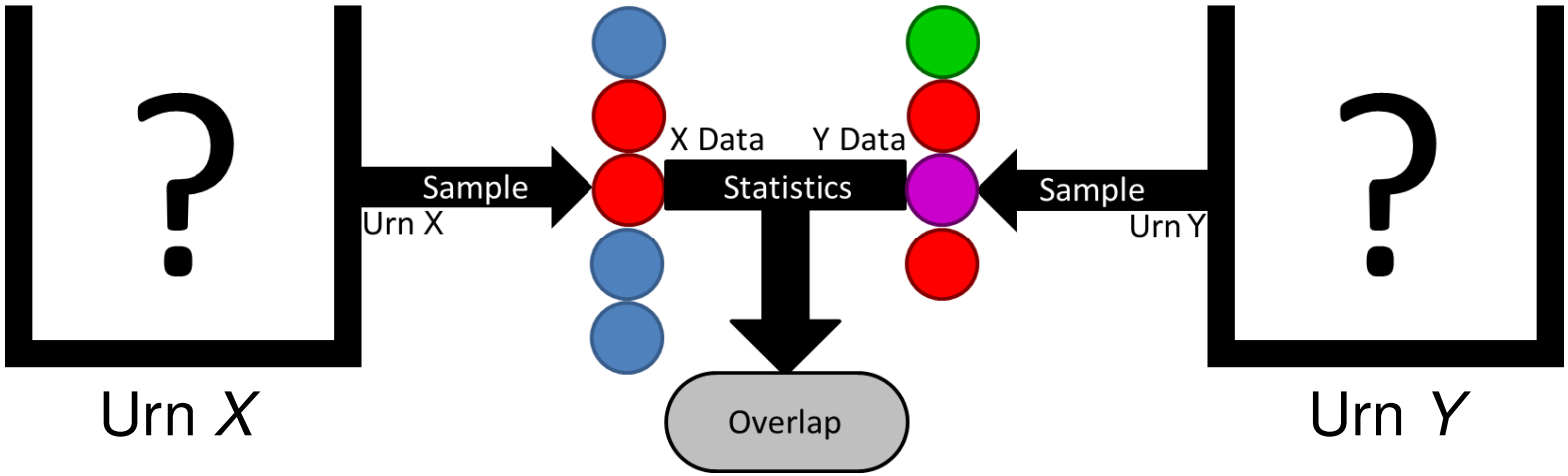
Consider $\theta(k) := \mathbb{P}(X_1 \notin \{Y_1, \cdots, Y_k\})$ as a basis for understanding **overlap**.



Measure of overlap:

Overlap:

Consider $\theta(k) := \mathbb{P}(X_1 \notin \{Y_1, \cdots, Y_k\})$ as a basis for understanding **overlap**.



$\theta(k)$ is **well-posed**, and relevant to the concept of **overlap**.
We will estimate $\theta(k)$ **unbiasedly**.

Summarizing the Data:

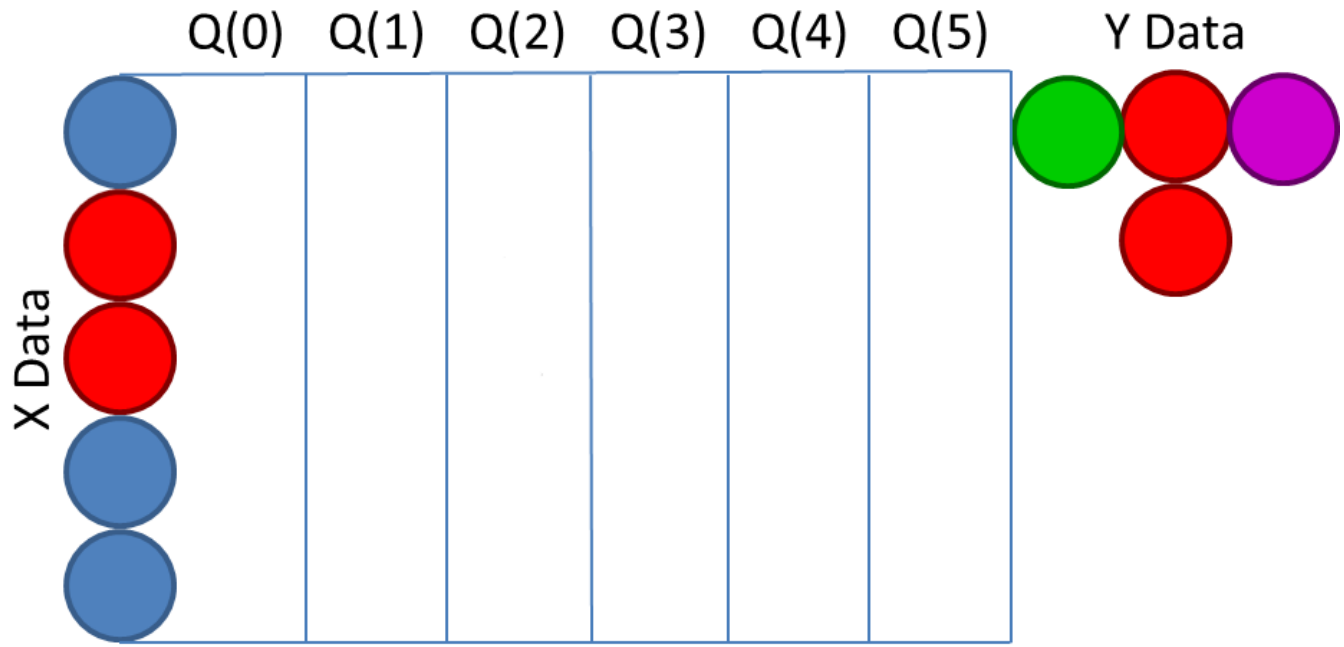
We wish to use $X_1, \cdots, X_{n_x}$, and $Y_1, \cdots, Y_{n_y}$ to **estimate** $\theta(k)$.

Summarizing the Data:

We wish to use $X_1, \cdots, X_{n_x}$, and $Y_1, \cdots, Y_{n_y}$ to **estimate** $\theta(k)$.

Q-Statistics:

Let $Q(j)$ be the number of draws from the X data with colors seen j times in the Y data.

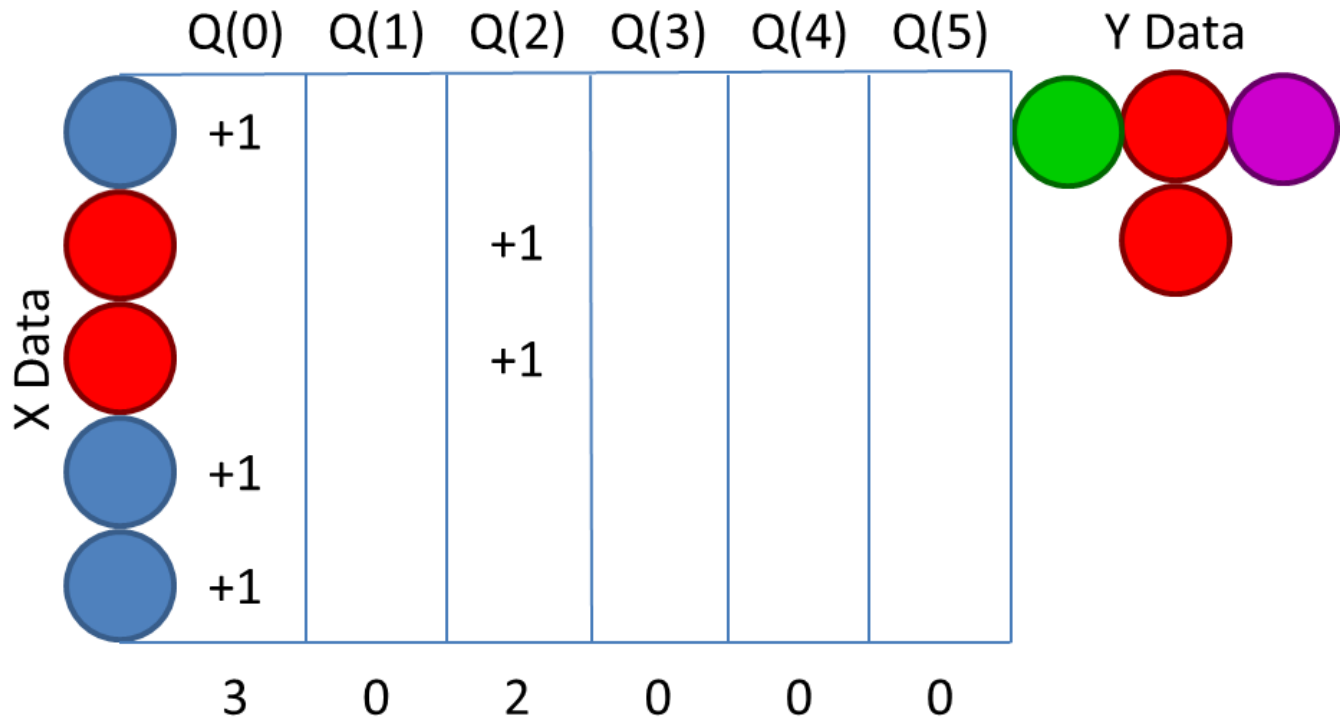


Summarizing the Data:

We wish to use $X_1, \cdots, X_{n_x}$, and $Y_1, \cdots, Y_{n_y}$ to **estimate** $\theta(k)$.

Q-Statistics:

Let $Q(j)$ be the number of draws from the X data with colors seen j times in the Y data.



First Side of an Estimator:

$$\theta(k) = \mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\})$$

Motivated by Robbins' (1968) and Starr's (1979) work on $\mathbb{P}(X_{k+1} \notin \{X_1, \cdots, X_k\})$, we seek an **unbiased** estimator for $\theta(k)$ that is a **linear combination** of the previous Q-statistics.

First Side of an Estimator:

$$\theta(k) = \mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\})$$

Motivated by Robbins' (1968) and Starr's (1979) work on $\mathbb{P}(X_{k+1} \notin \{X_1, \cdots, X_k\})$, we seek an **unbiased** estimator for $\theta(k)$ that is a **linear combination** of the previous Q-statistics.

Theorem: [Hampton-Lladser'11]

Let $\hat{\theta}_Q(k) = \sum_{i=0}^{n_y-1} \alpha_{i,k} Q(i)$, where $\alpha_{i,k} = \frac{\binom{n_y-i}{k}}{\binom{n_y}{k} n_x}$. Then $\hat{\theta}_Q(k)$ is an **unbiased** estimator for $\theta(k)$.

First Side of an Estimator:

$$\theta(k) = \mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\})$$

Motivated by Robbins' (1968) and Starr's (1979) work on $\mathbb{P}(X_{k+1} \notin \{X_1, \cdots, X_k\})$, we seek an **unbiased** estimator for $\theta(k)$ that is a **linear combination** of the previous Q-statistics.

Theorem: [Hampton-Lladser'11]

Let $\hat{\theta}_Q(k) = \sum_{i=0}^{n_y-1} \alpha_{i,k} Q(i)$, where $\alpha_{i,k} = \frac{\binom{n_y-i}{k}}{\binom{n_y}{k} n_x}$. Then $\hat{\theta}_Q(k)$ is an **unbiased** estimator for $\theta(k)$.

$\hat{\theta}_Q(k)$ is **easy to compute**, but **difficult to analyze**.

Second Side of an Estimator:

$$\theta(k) = \mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\})$$

A *U*-statistic is a **perfect boostrap** of a **kernel** statistic.

Second Side of an Estimator:

$$\theta(k) = \mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\})$$

A U-statistic is a **perfect boostrap** of a **kernel** statistic.

Estimating by a U-statistic:

Let $\hat{\theta}_U(k)$ be the U-statistic associated with **kernel** $[\![x_1 \notin \{y_1, \cdots, y_k\}]\!]$.

Second Side of an Estimator:

$$\theta(k) = \mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\})$$

A U-statistic is a **perfect boostrap** of a **kernel** statistic.

Estimating by a U-statistic:

Let $\hat{\theta}_U(k)$ be the U-statistic associated with **kernel** $[\![x_1 \notin \{y_1, \cdots, y_k\}]\!]$.

Let S_{k,n_y} be the set of **one-to-one** functions from $\{1, \cdots, k\}$ into $\{1, \cdots, n_y\}$. The U-statistic $\hat{\theta}_U(k)$ is given by

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

Second Side of an Estimator:

$$\theta(k) = \mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\})$$

A U-statistic is a **perfect boostrap** of a **kernel** statistic.

Estimating by a U-statistic:

Let $\hat{\theta}_U(k)$ be the U-statistic associated with **kernel** $[\![x_1 \notin \{y_1, \cdots, y_k\}]\!]$.

Let S_{k,n_y} be the set of **one-to-one** functions from $\{1, \cdots, k\}$ into $\{1, \cdots, n_y\}$. The U-statistic $\hat{\theta}_U(k)$ is given by

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

$\hat{\theta}_U(k)$ is **easy to analyze**, but **difficult to compute**.

A Key Equality:

$$\theta(k) = \mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\})$$

Both $\hat{\theta}_Q(k)$, and $\hat{\theta}_U(k)$ are symmetric and unbiased estimators for $\theta(k)$. Further, Clayton and Frees (1987) were able to equate Starr's Estimator of $\mathbb{P}(X_{k+1} \notin \{X_1, \cdots, X_k\})$ to a U-statistic form.

A Key Equality:

$$\theta(k) = \mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\})$$

Both $\hat{\theta}_Q(k)$, and $\hat{\theta}_U(k)$ are symmetric and unbiased estimators for $\theta(k)$. Further, Clayton and Frees (1987) were able to equate Starr's Estimator of $\mathbb{P}(X_{k+1} \notin \{X_1, \cdots, X_k\})$ to a U-statistic form.

Similarly in our problem,

Theorem: [Hampton-Lladser'11]

$\hat{\theta}_Q(k) = \hat{\theta}_U(k)$.

We remove the subscripts from our estimator.

A Key Equality:

$$\theta(k) = \mathbb{P}(X_1 \notin \{Y_1, \ldots, Y_k\})$$

Both $\hat{\theta}_Q(k)$, and $\hat{\theta}_U(k)$ are symmetric and unbiased estimators for $\theta(k)$. Further, Clayton and Frees (1987) were able to equate Starr's Estimator of $\mathbb{P}(X_{k+1} \notin \{X_1, \cdots, X_k\})$ to a U-statistic form.

Similarly in our problem,

Theorem: [Hampton-Lladser'11]

$\hat{\theta}_Q(k) = \hat{\theta}_U(k)$.

We remove the subscripts from our estimator.

$\hat{\theta}(k)$ is easy to **compute** and **analyze**.

Results from the *U*-statistics form:

With a proof similar to one given by Halmos (1946) we can prove the following theorem.

Theorem: [Hampton-Lladser'11]

When $n_x > 1, n_y > k$, $\hat{\theta}(k)$ is the **UMVUE** for $\theta(k)$. (The UMVUE is the estimator with the minimum variance among all unbiased estimators for our model.)

Idea of Proof:

1. Show symmetric estimators are optimal compared to asymmetric estimators.
2. Show the uniqueness of a symmetric estimator.

Results from the *U*-statistics form:

Many results can be written in terms of the **kernel**, $[\![x_1 \notin \{y_1, \cdots, y_k\}]\!]$.
For $i \in \{0, 1\}$, $j \in \{0, \cdots, k\}$, let

$$\xi_{i,j} = \mathbb{V}(\mathbb{E}([\![X_1 \notin \{Y_1, \cdots, Y_k\}]\!] | X_1, \cdots, X_i, Y_1, \cdots, Y_j)).$$

Results from the *U*-statistics form:

Many results can be written in terms of the **kernel**, $[\![x_1 \notin \{y_1, \cdots, y_k\}]\!]$. For $i \in \{0, 1\}$, $j \in \{0, \cdots, k\}$, let

$$\xi_{i,j} = \mathbb{V}(\mathbb{E}([\![X_1 \notin \{Y_1, \cdots, Y_k\}]\!] | X_1, \cdots, X_i, Y_1, \cdots, Y_j)).$$

An approach by Hoeffding (1948) gives that

$$\mathbb{V}(\hat{\theta}(k)) = \frac{\sum_{i=0}^{1} \binom{n_x - 1}{1-i} \sum_{j=0}^{k} \binom{n_y - k}{k-j} \binom{k}{j} \xi_{i,j}}{n_x \binom{n_y}{k}}.$$

Results from the *U*-statistics form:

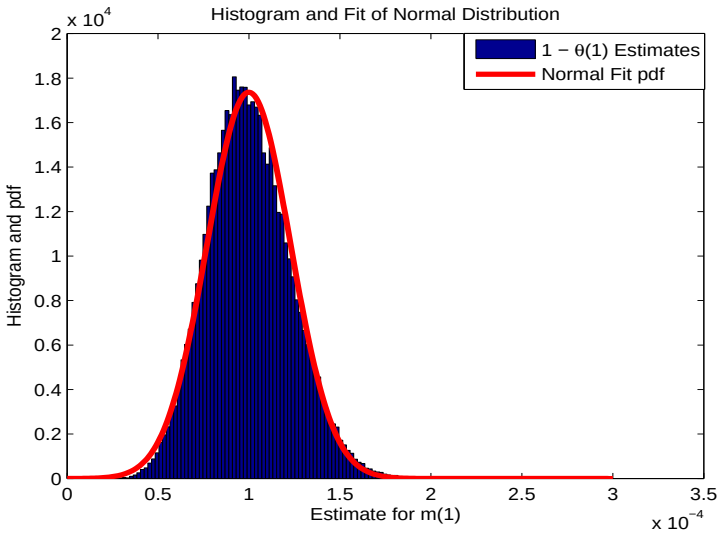
With a goal of creating **confidence intervals**, we compare $\xi_{i,j}$ to $\theta(k)$ to bound the variance.

Theorem: [Hampton-Lladser'11]

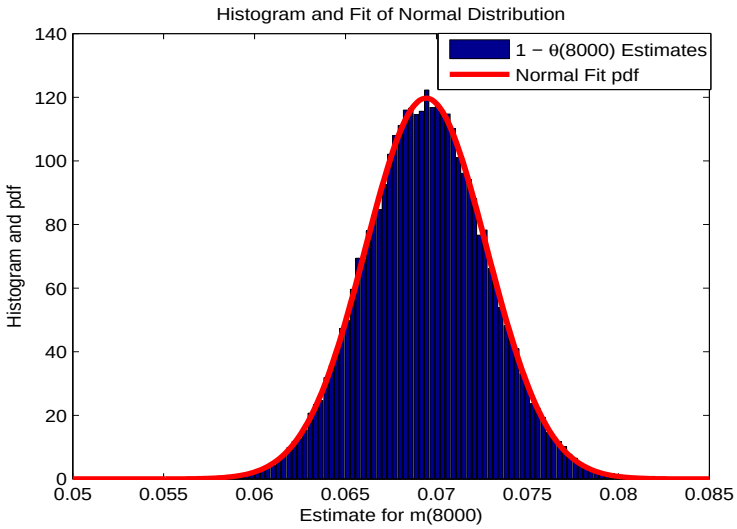
$$\begin{aligned}
\xi_{0,0} &= 0. \\
\xi_{1,0} &= \theta(2k) - \theta^2(k). \\
\xi_{1,j} &\leq \theta(2(k-j)) - \theta^2(k). \\
\xi_{0,j} &\leq \theta^2(k-j) - \theta^2(k).
\end{aligned}$$

An example of centrality:

Let $\mathbb{P}_X$ be **uniform** on a 1000 colors, and $\mathbb{P}_Y$ be approximately **geometric** on the same 1000 colors with $p = 0.9$. Here $n_x = n_y = 10,000$.



Distribution of $(1 - \hat{\theta}(1))$



Distribution of $(1 - \hat{\theta}(8000))$

A Central Limit Theorem:

The **projection** of $\hat{\theta}(k)$ explored by Grams and Serfling (1973) is given by

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

A Central Limit Theorem:

The **projection** of $\hat{\theta}(k)$ explored by Grams and Serfling (1973) is given by

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

$\hat{\theta}_P(k)$ is usually asymptotically **Normal**, and $\mathbb{V}(\hat{\theta}_P(k)) = \frac{\xi_{1,0}}{n_x} + \frac{k^2\xi_{0,1}}{n_y}$.
We can show that $\hat{\theta}(k)$ is asymptotically Normal by comparing $\hat{\theta}(k)$ to $\hat{\theta}_P(k)$.

A Central Limit Theorem:

Theorem: [Hampton-Lladser'11]

For each k, as $n_x, n_y \to \infty$,

$$\frac{\hat{\theta}(k) - \hat{\theta}_P(k)}{\sqrt{\mathbb{V}(\hat{\theta}(k))}} \xrightarrow{\mathcal{L}^2} 0.$$

Idea of Proof:

Let $U(k) := \hat{\theta}(k) - \hat{\theta}_P(k)$. U_k is a U-statistic with expectation 0. $\mathbb{E}(U^2(k))$ can be shown to decrease faster than $\mathbb{V}(\hat{\theta}(k))$ by a counting argument.

If $\lim_{k\to\infty} \theta(k) > 0$, then this convergence can be shown to be uniform for $k \in \{1, \cdots, n_y\}$.

Application:

Suppose we have an initial sample from microbial environments, and are interested in maximizing a **sum of overlap** between environments after our second sample.

Application:

Suppose we have an initial sample from microbial environments, and are interested in maximizing a **sum of overlap** between environments after our second sample.

$\theta(k)$ is a **sum of decaying exponentials**. We approximate $\hat{\theta}(k)$ with such a sum. An **extrapolation** gives an estimate of $\theta(k)$ for $k > n_y$.

Application:

Suppose we have an initial sample from microbial environments, and are interested in maximizing a **sum of overlap** between environments after our second sample.

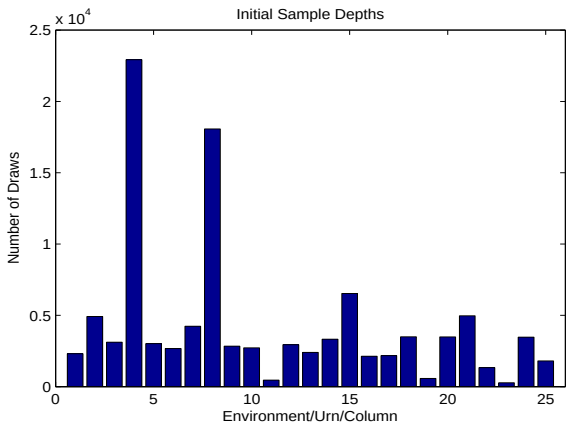
$\theta(k)$ is a **sum of decaying exponentials**. We approximate $\hat{\theta}(k)$ with such a sum. An **extrapolation** gives an estimate of $\theta(k)$ for $k > n_y$.

Algorithm:

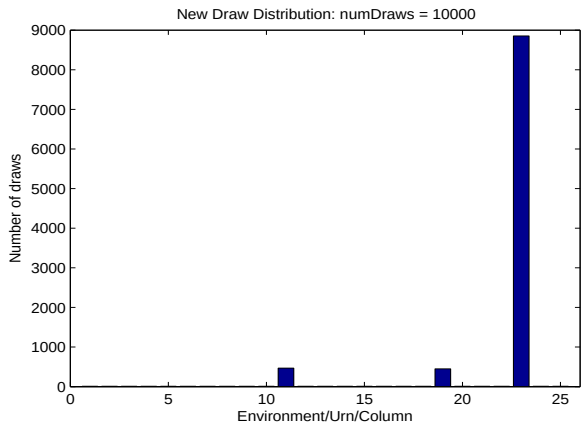
1. Compute $\hat{\theta}(k)$ for each Urn X and Urn Y combination.
2. Approximate each $\{\hat{\theta}(k)\}_{k=1}^{n_y}$ by a sum of exponentials.
3. Find an optimal sample strategy for the second sample using Newton's Method and Log-Barriers.

Application:

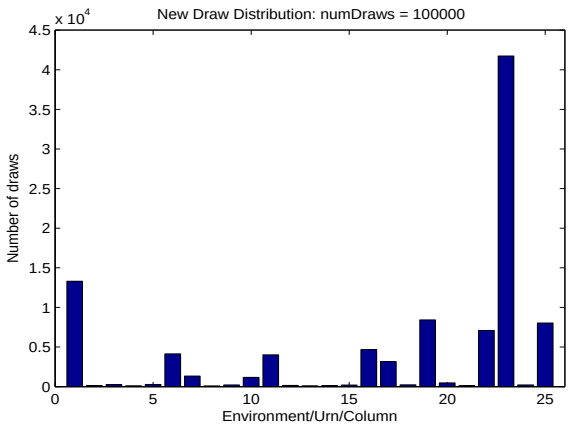
Data from 25 human microbial environments produces the following sample strategies.



Initial Sampling



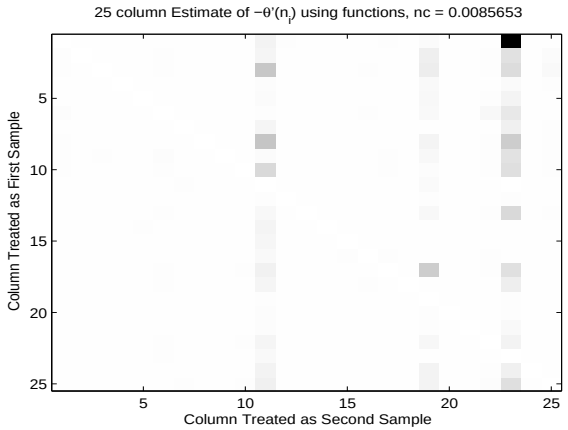
Optimal Sampling for 10K Draws



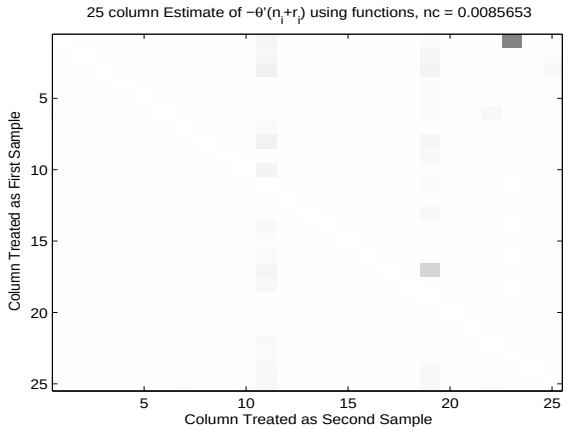
Optimal Sampling for 100K Draws

Application:

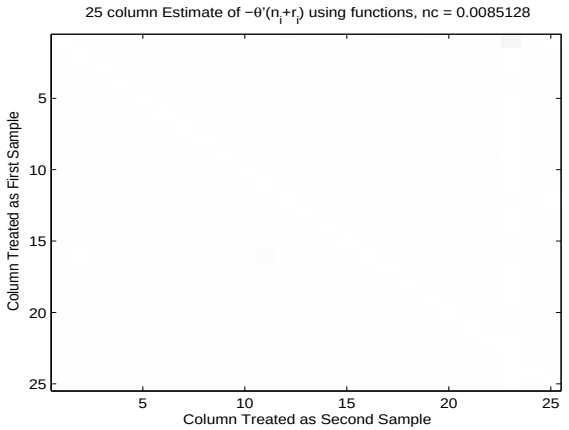
Here we see the magnitude of the **derivative estimates** for the $\theta(k)$ at different sampling depths.



Initial Sampling



Optimal Sampling for 10K Draws



Optimal Sampling for 100K Draws

Thank You!

Questions?