Distinct Value Estimators For Zipfian Distributions

Sergei Vassilvitskii
Rajeev Motwani

Stanford University

Problem Statement

Given a large multiset X with n elements, count the number of distinct elements in X.

$$X = \{a, b, c, a, a, c, b, a\} \Longrightarrow Distinct(X) = 3$$

Alternatively, given samples from a distribution $\mathcal{P}$, estimate the 0-th frequency moment.

Why Do We Care?

Good Planning for SQL Queries. Consider:

$$\text{select } * \text{ from } R, S \text{ where } R.A = S.B \text{ and } f(S.C) > k$$

where f is expensive to compute.

If $S.C$ has few distinct elements, compute f first, cache results, then join.

If $S.C$ has many elements, compute the join first, then check the f condition.

Orders of Magnitude Improvements

Classical problem

Different approaches:

Streaming Input - Minimize space used.

Sample from Input - Guarantee on approximations?

Given a sample of size r from X, find $\hat{D}$ an approximation to $Distinct(X)$.

Previous Work

Good-Turing Estimator: “The Population Frequencies of Species, and the estimation of Population Parameters,” 1953.

Other Heuristic Estimators:

- Smoothed Jackknife Estimator (Haas et. al)
- Adaptive Estimator (Charikar et. al)
- Many Others

Previous Work - Theory

Given r samples from a set of size n

Guaranteed Error Estimator(GEE) [CCMN]

Approximation Ratio: $O(\sqrt{n/r})$

Lower Bound:

There exist inputs such that with constant probability any estimator will have approximation ratio at least:

$$\sqrt{\frac{n-r}{2r}}$$

Lower Bound Detail

Scenario 1: $S = \{x, x, \ldots, x\}, |S| = n$

Scenario 2: $S' = \{x, x, \ldots, x, y_1, \ldots, y_k\}, |S'| = n$

With $k = \frac{n-r}{2r} \ln \frac{1}{\delta}$, after r samples cannot distinguish between the two scenarios with probability at least δ.

So Why Are We Here?

Many large datasets are not worst-case. In fact, many follow Zipfian Distributions.

$$Zipf_\theta(i) \propto \frac{1}{i^\theta}$$

Examples:

- In/Out-Degrees of the Web Graph
- Word frequencies in many languages
- many, many more.

Problem Definition

Suppose $X \sim Zipf_\theta$ on D elements.

θ is known, D is unknown

Estimate D by sampling from X .

Two Kinds of Results:

- Adaptive Sampling: Will sample from X until a stopping condition is met.

- Best-you-can Estimation: Given a sample from X, return best estimate of D .

Results

Let p^* be the probability of the least likely element.

Adaptive sampling will return D after at most $O(\frac{\log D}{p^*})$ samples with constant probability.

Given $r = \frac{(1+2\epsilon)^{1+\theta}}{p^*}$ samples, can return an $1+\epsilon$ estimate to D with probability at least $1 - \exp(-\Omega(D\epsilon^2))$

Outline

Introduction

Techniques

Experimental Results

Conclusion

Approximation Techniques

For a sample of size r let f_r be the number of distinct values in the sample.

Suppose D and θ are known, then we can compute $E_{D,\theta}[f_r]$ the expected number of distinct values in the sample.

If f_r^* is the number of distinct values observed, the estimator returns $\hat{D}$ such that $E_{\hat{D},\theta}[f_r] = f_r^*$.

Analysis

Lemma: Tight Distribution of f_r.

For large enough r,

$$Pr\Big[|E[f_r] - f_r| \geq \epsilon E[f_r]\Big] \leq \exp(-\epsilon^2 \Omega(D))$$

Proof: Parallels the sharp threshold coupon collector arguments for uniform distributions.

Analysis (2)

Lemma: MLE preserves approximation

Given: $f_r \leq (1+\epsilon) E_{D,\theta}[f_r]$, observed f_r^* elements

Let $\hat{D}$ such that $f_r^* = E_{\hat{D},\theta}[f_r]$, and $r \geq 1/p^*$.

Then: $(1-2\epsilon)\hat{D} \leq D \leq (1+2\epsilon)\hat{D}$

Outline

Introduction

Techniques

Experimental Results

Conclusion

The Competition

Zipfian Estimator (ZE): Performance guarantees only for Zipfian Distributions.

Guaranteed Error Estimator (GEE): $O(\sqrt{n/r})$ error guarantee. (Works for all distributions)

Analytic Estimator (AE): Best performing heuristic - no theoretical guarantees.

Datasets

Synthetic Data:

- Vary number of distinct elements $D \in \{10k, 50k, 100k\}$

- Vary the Database size $n \in \{100k, 500k, 1000k\}$

- Vary the skew of the distribution $\theta \in \{0, 0.5, 1\}$

Real Datasets

- "Router" dataset - Packet trace from the Internet Traffic Archive. $\theta \approx 1.6$, $n \approx 4M$, $D \approx 250k$

Estimating θ

Recall: $Zipf_\theta(i) \propto \frac{1}{i^\theta}$

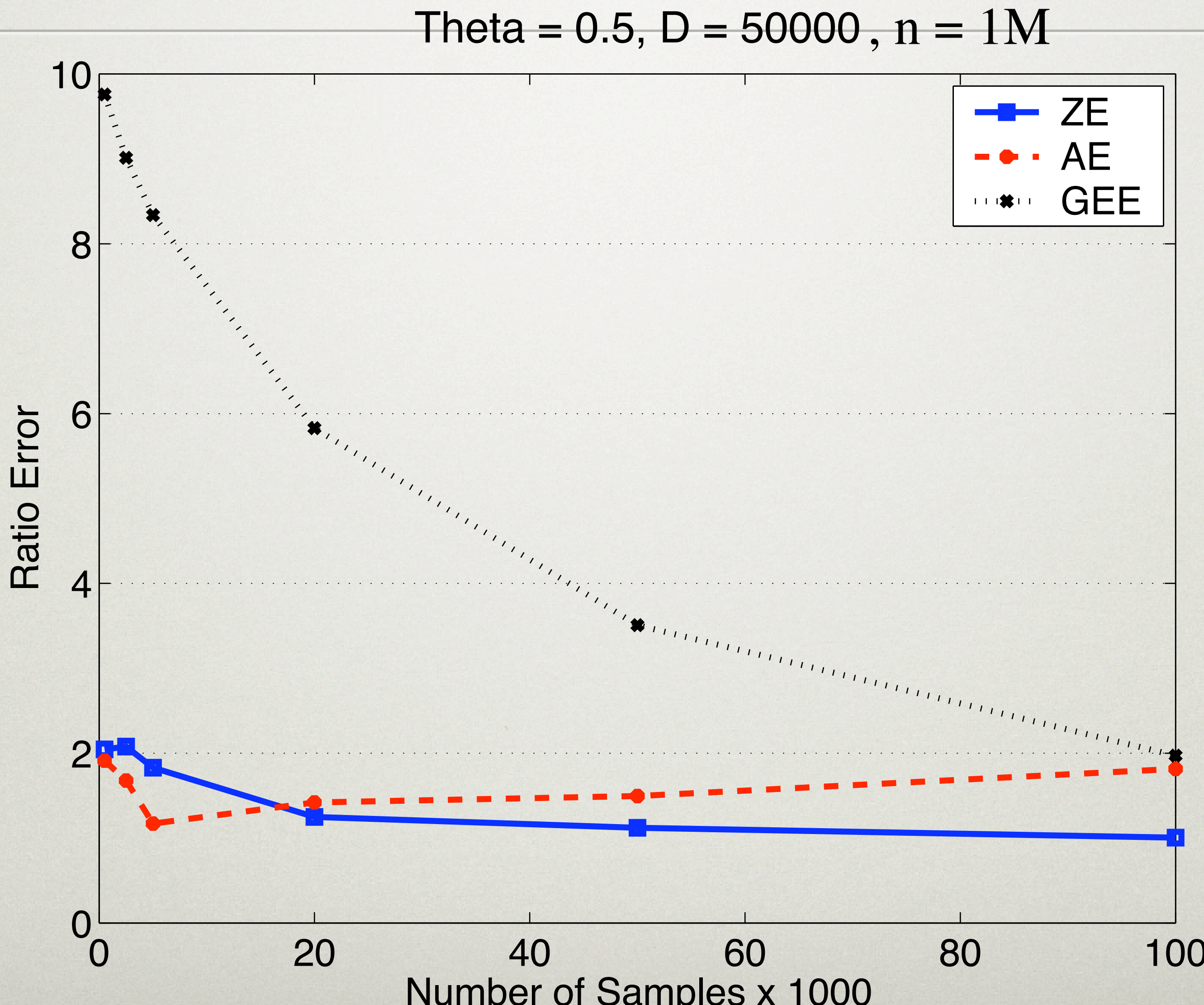
Let f_i be the frequency of the i-th element.

$E[f_i] = cri^{-\theta} \implies \log E[f_i] = \log cr - \theta \log i$

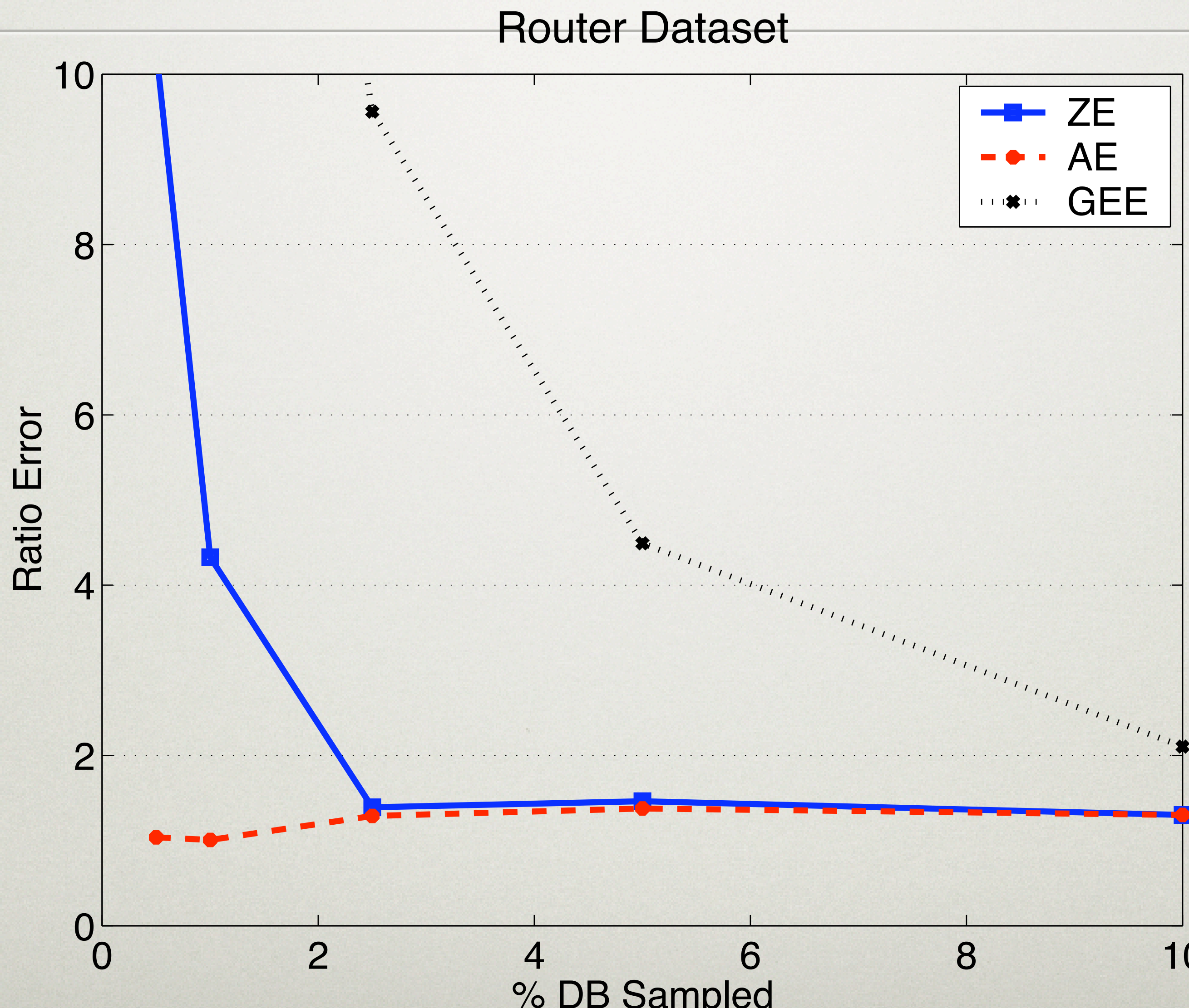
Estimate θ by doing linear regression on

$\log f_i$ vs $\log i$ plot.

Experimental Results



Experimental Results (2)



Outline

- Introduction
- Techniques
- Experimental Results
- Conclusion

Conclusion

Can have error guarantees if the family of distributions is known ahead of time.

How does the approximation of θ affect error guarantees?

Subtle problem: disk reads occur in blocks. Time to sample 10% is equivalent to reading the whole DB.

Thank You