Abstract—The problem of distinct value estimation has many applications. Being a critical component of query optimizers in databases, it also has high commercial impact. Many distinct value estimators have been proposed, using various statistical approaches. However, characterizing the errors incurred by these estimators is an open problem: existing analytical approaches are not powerful enough, and extensive empirical studies at large scale do not exist. We conduct an extensive large-scale empirical study of 11 distinct value estimators from four different approaches to the problem over families of Zipfian distributions whose parameters model real-world applications. Our study is the first that scales to the size of a billion-rows that today’s large commercial databases have to operate in. This allows us to characterize the error that is encountered in real-world applications of distinct value estimation. By mining the generated data, we show that estimator error depends on a key latent parameter — the average uniform class size — that has not been studied previously. This parameter also allows us to unearth error patterns that were previously unknown. Importantly, ours is the first approach that provides a framework for visualizing the error patterns in distinct value estimation, facilitating discussion of this problem in enterprise settings. Our characterization of errors can be used for several problems in distinct value estimation, such as the design of hybrid estimators. This work aims at the practitioner and the researcher alike, and addresses questions frequently asked by both audiences.

I. Introduction

Consider the following problem: estimate the number of distinct values (or classes) in a population by statistically analyzing its sample. This is the problem of distinct value estimation, and it arises in a surprisingly large variety of applications, where it is used to estimate a number of interest: in census studies, the number of individuals in lists having duplications [20]; in economics, the number of investors from samples of share registers of companies [23]; in ecology, the number of species from some sampling scheme [29]; in numismatics, the number of coins produced by a mint when a selection of such coins has been found [7].

However, the application that has arguably the most commercial impact arises in databases1. Large data warehouses rely on complex query optimizers to formulate their query plans. A query optimizer needs an estimate of the number of distinct values in its attributes in order to formulate its query plan [19]. On attributes that are not indexed, single pass (or scan) estimators provide fairly accurate estimates of distinct values while using a small memory footprint [14, 2]. This is a highly valuable field of study, and has vibrant literature. However, databases have witnessed explosive growth over the past decade; today’s large commercial data warehouses have billions of rows. Therefore, accessing all the data in this manner is not always possible. This makes sampling based estimators of distinct values a highly practical solution for distinct value estimation in many large data warehouses. All the large scale commercial data warehouses that we are aware of use sampling based estimators of distinct values as part of their query optimizer. This includes those data warehouses listed in Gartner’s “magic quadrant” for enterprise data warehousing platforms. Our work was motivated by, and tested on, sampling based distinct value estimation on a data warehouse from a Fortune-10 company.

A great range of sampling based estimators for distinct values have been proposed in both the statistics and database literature, see [3] for a survey. Distinct value estimation is a hard problem, and estimators generally incur (significant) errors in performing the estimates. Therefore, users with some knowledge about the datasets they will encounter in applications, are interested in the following questions.

Q1 From this plethora of distinct value estimators, which estimator will give the least error on their dataset which is of the size of a typical database — (millions to a billion rows)?

Q2 For a choice of estimator, can one characterize the errors that will be incurred on such a large dataset?

Q3 How high should the sampling fraction be in order to keep error within a tolerable margin? Conversely, given a sampling fraction and a desired error margin, what choice of estimator will restrict errors to that margin?

Q4 Do the errors occur in patterns that can be effectively visualized?

These problems have serious practical ramifications in database design: a poor estimate of distinct values can result in a considerably more expensive query plan [19]. Furthermore, a query estimator may have more tolerance for a certain region of bias, as opposed to other regions (depending on the region where the query plan changes from a good one to an inefficient one). Unfortunately, due to the difficulty of the problem, and the lack of an adequate characterization of errors, commercial systems often suffer from unexpected poor estimates that seem to occur “at random”. These cause the query optimizer to formulate highly inefficient query plans, resulting in intolerable delays to the system user, especially when the database has in excess of hundred million rows. Simply put, our state of knowledge about error in distinct value estimators is untenable due to the large sizes of today’s databases.

One might think that a way out would be to increase the sampling fraction. However, the cost of even 1% sampling in a commercial database with billions of rows, is significant, and a larger than 2% sample is often simply not possible. Also, because data is stored in blocks, generating a 10% random sample is sometimes as expensive as scanning all the data. Therefore, there is great practical interest in the least sample size that can deliver a reasonable estimate of distinct values.

Finally, visual descriptions of errors, in addition to facilitating one’s own understanding, permit effective communication to non-specialists, which is an important aspect of working with statistical technologies in an enterprise setting.

Unfortunately, there do not yet exist analytical techniques that characterize error satisfactorily (or answer any of (Q1)-(Q4)) for any distinct value estimator in literature (except, to a degree, for the \(D_{\text{DEE}}\); see related work. Instead, there is a powerful result in the “opposite direction” in [10, 9] that says that every estimator will give a large error on some dataset. This suggests that new analytical approaches may have to be developed for specific datasets/distributions.

In the absence of analytical approaches to characterizing error, we must turn to empirical characterizations performed through a large-scale extensive study over distributions that represent real-world applications. However, here, the situation is best described by [19]:

Work done when first author was with Hewlett Packard.

1 The high-end large data warehouse market for 2022 is estimated to be $22B [30].
Unfortunately, analysis of distinct value estimators is non-trivial, and few analytic results are available. To make matters worse, there has been no extensive empirical testing or comparison of estimators either in the database or statistical literature—the testing that has been done in the statistical literature has generally involved small sets of data in which the frequencies of different attribute values are fairly uniform; real data is seldom so well-behaved.

Our work arose from an effort to design a distinct value estimator for a large commercial database product that had, among its customers, a Fortune-10 company. We did not find an extensive, large-scale, comparative study on distinct value estimators that would allow us to answer questions (Q1)-(Q4) reliably for critical database applications. Several studies in literature show different estimators to be the best over the datasets in the purview of the particular study. These are valuable data points, but they are almost never comparable, and it is not clear how far the results generalize (see related works section online).

This brings us to our approach. Our study characterizes error in distinct value estimators, and provides answers to (Q1)-(Q4). Our approach can be described in the following steps.

1) Conduct an extensive empirical study of the relative performance of various estimators on a well-chosen parameter space.
2) Mine the generated data for stable patterns to the relative performance of estimators on datasets,
3) Find parameters that organize these stable patterns.
4) Present these stable patterns in a manner that is easy to visualize and communicate among practitioners.
5) Characterize error through bias, ratio-error, and RMSE, for each region of the parameter space that is delineated by stable patterns of behavior.

A. Our Contributions

Our extensive empirical study allowed us to construct a detailed large-scale characterization of the relative behavior of important families of distinct value estimators. Our study is the first that scales to the billion-row size that today’s large commercial databases operate on. Our characterization describes both inter- and intra-family behavior over a parameter space that models the variation of real-world data. We identify stable patterns of behavior that allow us to “make sense” of the huge amount of data generated in our empirical study. We identify a critical latent parameter — the average uniform class size — that estimators are sensitive to, and that allows us to find patterns in our results. This variable has not yet been studied in literature.

Our study allows us to answer the following types of questions.

1) What are scale effects on each estimator?
2) What are skewness effects on each estimator?
3) How does the actual distinct value count affect an estimator?
4) What is the “best” estimator under various definitions?
5) What sampling percentage is adequate for various error requirements?

Finally, our study allows us to comparatively evaluate, at large-scale, the four different approaches to distinct value estimation that are within our purview.

Limitations and Scope of our Study. Our study is most useful for distributions that approximate a power law, and with population parameters that resemble those of our study. However, both these are intended to reflect real-world problems.

II. Related Work

Due to space constraints, we refer the reader to the section in the long version of the paper on arXiv, which also contains a description of the ideas behind the 11 estimators that are under study.

III. Methodology

A. Datasets and Protocol

Characteristics of real-world data are numerous, and often interacting. We can never be certain which characteristic of the data is causing or influencing the performance of the estimator. In this way, the performance of the estimator becomes relative, and not absolute as a function of a chosen set of parameters.

This points us to the use of artificially generated data from a well-chosen parameter space, as the means to our desired characterization. The potential hazard in doing so is that the artificial data may not represent any real-world application. So while desiderata (1) and (2) are met, we might fail on (3).

Recall that the family of Zipfian distributions $Z_{A,\theta}$ parametrized by their skewness parameter $\theta$, and the size of the alphabet $A$, have probability masses $P_{A,\theta}(i)$ satisfying $P_{A,\theta}(i) \propto \frac{1}{i^\theta}$. Normalizing so that we get a distribution, we obtain the probability mass

$$P_{A,\theta}(i) = \left( \sum_{i=0}^{D} \frac{1}{i^\theta} \right)^{-1} \frac{1}{i^\theta}.$$

It is by now accepted that several naturally occurring distributions of importance are power laws, see [22] for diverse examples. Equally importantly, several distributions approximate power laws once their outliers are removed. Several quantities stored in our commercial database systems also follow power law distributions after similar processing. Therefore, a study on a Zipfian parameter space is important in and of itself. Secondly, the Zipfian distribution allows us to vary just the right parameters — namely skewness and the size of the unique alphabet — that are most germane to the performance of distinct value estimators.

In light of the discussion above, we choose to characterize our estimators over a Zipfian parameter space, by varying the parameters of the Zipfian population over a wide range of values that reflect real-world applications. In this way, we get the “best of both worlds” — we are able to vary parameters, understand estimator behavior as an absolute function of these parameters, and model real-world applications.

In order to obtain a fine grained characterization, we use the design of (population size, alphabet size) pairs as shown in Table I. Each such pair will be called a regime. Therefore, there are 20 regimes. As depicted, the regimes were organized into five $N \times 7$ values in $[10, 20, 100, 500, 1000]$.

| 1B | 100M | 1B | 50M | 1B | 1M | 1B | 50K |
|---|---|---|---|---|---|---|---|
| 100M | 10M | 100M | 50M | 100M | 1M | 100K | 100M |
| 100K | 1M | 500K | 1M | 100K | 10K | 50K | 10K |
| 1M | 10K | 1M | 5K | 1M | 1K |

For each regime, we generated 5 Zipfian populations by varying the Zipfian skewness parameter $\theta$ through the range $[0, 0.5, 1, 1.5, 2]$ that covers most real-world applications. In this way, we obtain 100 Zipfian populations. Note that at high skewness, the number of distinct classes $D$ in the population is less than the size of the alphabet $A$.

We refer to $\frac{N}{A}$ as the average uniform class size since it is the average class size when $\theta = 0$ (and up till the time that each alphabet occurs in the population). We will see that it is a critical parameter in characterizing estimator error.

For each of the 100 Zipfian populations, we varied the sampling percentage through the values $[0.1, 1, 2.5, 10]$. For each data-point, we drew 10 random samples without replacement. We ran our 11 distinct value estimators on each of the 10 samples, thereby generating 10 estimates for each of the 11 estimators. Finally, for
each estimator, we computed the average bias of the 10 estimates, as well as the variance across the 10 estimates. In this way, we report a total of \(100 \times 5 \times 10 \times 11 = 55,000\) experiments. We should note that in our study, we experimented with a strictly larger range of parameter values than what is reported in this paper. However, to conserve space, we “compress” to a subset of our range of parameter values that we feel adequately described the error patterns.

**IV. Results**

Our extensive empirical study generated a large amount of data. Our goal is to provide a thorough characterization of both, the individual and relative performance of the estimators, and to organize and understand the error patterns that emerged from our study. By mining the generated data, we found that the parameter \(\frac{N}{A}\) provides us with this organizing principle: *when the results of the experiments are arranged in a grid whose X-axis is \(\frac{N}{A}\), and Y-axis is \(N\), we can see regularity in the patterns.* Accordingly, we provide grids of 2D plots of estimate vs. actual distincts for all estimators in a family, as well as grids of 3D surfaces showing normalized bias \((\hat{D} - D)/D\) for each individual estimator. In both cases varying the parameters of the underlying population as well as the sampling fraction. For the 3D surfaces, normalized biases of \(-1, 0,\) and 0 are marked on the vertical axes, and 1 and 2 can be seen in the form of dotted lines.

We then point out the salient features of both the relative and the individual behaviors as we vary the parameters of the population. We observe patterns that arise when we go from left to right on each row of the 2D plots. This gives us the variations with the parameter \(\frac{N}{A}\). Likewise, we report variations with \(N\), within each plot with \(\theta\), and, finally, with \(q\). This analysis is available in the online version on arXiv.

To save space, we show only a single 2D plot for each family: the one at which the maximum ratio error for the most accurate estimator in the family is at most 5 (Table V). The remaining 2D plots are all included in supplementary material in the online version on arXiv.

We suggest that when reading the results, the reader begin with the 3D surfaces for each estimator to understand its individual performance, followed by inspection of the 2D plots (including those in the supplementaries) to complete the relative picture.

We begin with the jackknife family. This is described in Fig 1 and 2. Next, the Schlosser estimators are described in Fig 3, 4, and 5. The \(\hat{D}_{\text{GEE}}\) and \(\hat{D}_{\text{AE}}\) estimators are described in Fig. 6. Finally, Fig 7 shows estimators having relatively flatter bias surfaces.

**V. Discussion**

The discussion is organized as follows. First we describe the relative sensitivity of each estimator to the various parameters. Then we address the question “which are the best estimators?” in terms of accuracy over the entire parameter space of our characterization. Next, we identify regions of the parameter space where certain estimators do well, even though they may not perform well over the entire parameter space. Finally, we address the question “how much sampling do we need?”

**A. Sensitivity to Parameters**

**a) Sensitivity to \(\frac{N}{A}\):**

The grids of 3D bias surfaces clearly show that every family—the Schlossers, jackknives, \(\hat{D}_{\text{GEE}}\) and \(\hat{D}_{\text{AE}},\) and Chao-Lee — is sensitive to changes in \(\frac{N}{A}\), and all except the last show regularity in their behaviour as a function of \(\frac{N}{A}\). The parameter \(\frac{N}{A}\) emerges as the single most important organizing parameter for estimator behavior overall. The sensitivity to \(\frac{N}{A}\) is high in \(\hat{D}_{\text{GEE}}, \hat{D}_{\text{AE}}, \hat{D}_{\text{GEE}}, \hat{D}_{\text{DLJ}}, \hat{D}_{\text{GEE}}, \hat{D}_{\text{CLJ}},\) and \(\hat{D}_{\text{CLJ}}\). Compared to the above, there was mild sensitivity in \(\hat{D}_{\text{aj1}}, \hat{D}_{\text{aj2}}, \hat{D}_{\text{aj3}},\) and \(\hat{D}_{\text{Aj}}\). Therefore, even within a family, some members are highly sensitive to changes in \(\frac{N}{A}\), while others are not.

**b) Sensitivity to Scale \(N\):**

Among the 11 estimators we tested, \(\hat{D}_{\text{aj2}},\) and the two Chao-Lee estimators were highly sensitive to scale. Namely, as we go up vertically along their grids, their 3D bias surfaces showed significant changes. Of these, the change was quite regular and predictable in \(\hat{D}_{\text{aj2}},\) but irregular and unpredictable in the Chao-Lee families. In the case of \(\hat{D}_{\text{aj2}},\) the bias increases as we increase the scale, with the increase being most prominent around mid-skew as can be seen from the 3D bias surfaces for \(\hat{D}_{\text{aj2}}\). (In the online version). See supplementaries for complete Chao-Lee estimator grids.

**c) Sensitivity to Sampling Fraction \(q\):**

The 2D plots (see supplementaries) are useful to illustrate the effect of sampling fraction. The estimators that improve most as \(q\) increases are \(\hat{D}_{\text{aj1}}, \hat{D}_{\text{aj3}}, \hat{D}_{\text{CLJ}}\). The estimators that are relatively less sensitive to increases in \(q\) in our range, for some values of other parameters, are \(\hat{D}_{\text{aj2a}}\) and \(\hat{D}_{\text{AE}}\). For example, both of them remain relatively accurate for low skew even at low sampling fractions (see Tables II and III).
The estimators that show anomalous or irregular behavior as \( q \) is increased are \( \hat{D}_{0.02} \), \( \hat{D}_{0.02} \), and the Chao-Lee family. In the case of \( \hat{D}_{0.02} \) and \( \hat{D}_{0.02} \), we see anomalous degradation in performance as we go from \( q = 0.001 \) to 0.005, especially when \( \frac{N_q}{N} < 200 \), see 2D plots in supplements.

**d) Sensitivity to Zipfian Skew \( \theta \):**

Sensitivity to \( \theta \) can be seen more finely in the 2D plots (see supplements). The estimators \( \hat{D}_{0.02} \), \( \hat{D}_{0.02} \), \( \hat{D}_{0.02} \), \( \hat{D}_{0.02} \), and \( \hat{D}_{0.02} \) are highly sensitive to changes in Zipfian skew. In particular, \( \hat{D}_{0.02} \) and \( \hat{D}_{0.02} \) are quite inaccurate at low skew, \( \hat{D}_{0.02} \) and \( \hat{D}_{0.02} \) perform poorly in mid-skew, while \( \hat{D}_{0.02} \) also performs poorly for smaller populations and high skew, even at high sampling fractions.

The estimators \( \hat{D}_{0.02} \), \( \hat{D}_{0.02} \), \( \hat{D}_{0.02} \), and \( \hat{D}_{0.02} \) are relatively insensitive to changes in Zipfian skew. Finally, \( \hat{D}_{0.02} \) and \( \hat{D}_{0.02} \) respond irregularly to changes in Zipfian skew.

**B. The Best Estimators Overall and their Relative Performance**

From our extensive study, we can conclude that three estimators provide relatively strong performance across variations in all underlying parameters. These three are \( \hat{D}_{0.02} \), \( \hat{D}_{0.02} \), \( \hat{D}_{0.02} \), and \( \hat{D}_{0.02} \) (cf. the choice of the provisional estimator in [3], which is \( \hat{D}_{0.02} \)). It is perhaps easiest to see this from Table II, and inspect the low and high sampling cases separately. Of course, which estimator is to be used for an application depends on the sampling fraction that is available (we return to this question in § V-E), and the skewness of the population in case it is known. A subtler issue is whether it is the ratio error that is critical to the application, or the actual value and sign of the bias, and the role played by \( \frac{N}{N} \). For example, when the number of distincts is relatively low, even large ratio errors do not result in high absolute value of bias. Therefore, we cannot speak meaningfully about “best estimator” in terms of just ratio error or bias — we do need to include the \( \frac{N}{N} \) factor as well. Since in certain database query optimizers, it may be the value and sign of bias that is the critical factor in change of a query plan, while in others, it may be the ratio error, practitioners will find it useful to have an analysis along both metrics. In both cases, we discuss the low sampling scenario (\( q < 0.01 \)) below since that is where differences may be most manifest.

**1) By Ratio Error**

See Table II. For low sampling fractions, for low \( \frac{N_q}{N} \), \( \hat{D}_{0.02} \) and \( \hat{D}_{0.02} \) are the best estimators when \( 0 \leq \theta \leq 1 \). For higher skew, \( \hat{D}_{0.02} \) continues to perform well. At high \( \frac{N_q}{N} \) and low-mid skew, \( \hat{D}_{0.02} \) and \( \hat{D}_{0.02} \) are the best, which \( \hat{D}_{0.02} \) again does very well as \( \frac{N_q}{N} \) increases.

**C. Regions of Good Performance**

The three estimators discussed above do reasonably well in all regions. However, there are other estimators that actually do better than these, but in small sub-domains. However, these sub-domains are clearly delineated, and therefore we can potentially use these estimators when our data lies in the corresponding domains.

**1) Regions Defined by Zipfian Skew**

The best example of such estimators are \( \hat{D}_{0.02} \) and \( \hat{D}_{0.02} \). Whether sampling fractions are low or high, these estimators absolutely shine...
in the high skew region of \( \theta > 1.5 \) (see Table II). Indeed, their ratio errors are an order of magnitude lower than other estimators in this region. Interestingly, \( \hat{D}_{Sh2} \), which is, on the average, a far better estimator than \( \hat{D}_{Sh} \) and \( \hat{D}_{Sh3} \) due to its reasonable performance at low-mid \( \theta \), does not offer as good of a accuracy gain in this high skew region. We also note that for high \( \frac{N}{\theta} \sim 1000 \), the bias profile of \( \hat{D}_{Sh} \) and \( \hat{D}_{Sh3} \) turns from a "slope" to a "hat", and then they also offer reasonably accurate estimates at low-skew (see § IV for a discussion of this phenomenon).

2) Regions Defined by Coefficient of Class Variation

Earlier studies [19, 20] have reported some trends by coefficient of class variation. We validate some of these at higher scale and dimensionality of parameter space. On the other hand, other reported trends no longer continue to hold in our large-scale study. Note that we vary the sampling percentage through a wider range of values than previous studies.

a) \( 0 \leq \gamma^2 \leq 1 \):

For low sampling fractions (\( \leq 0.005 \)), \( \hat{D}_{aj2a} \) is the best estimator in the region \( 0 \leq \gamma^2 \leq 1 \); however, \( \hat{D}_{aj1} \), \( \hat{D}_{Sh2} \), and \( \hat{D}_{AE} \) are comparable (cf. [20], where \( \hat{D}_{aj2} \) was declared the best estimator in this region).

For high sampling fraction (\( > 0.005 \)), the picture remains the same, except for high \( \frac{N}{\theta} \) (\( > 100 \)), where \( \hat{D}_{AE} \) emerges as the best estimator. For sizes below 1B, \( \hat{D}_{Sh2} \) is the best estimator in this region.

b) \( 1 \leq \gamma^2 \leq 50 \):

For low and high sampling fractions, \( \hat{D}_{aj2a} \) is the best among the jackknives, but comparable to \( \hat{D}_{aj1} \) and \( \hat{D}_{aj2} \). However, it is the Schlossers \( \hat{D}_{Sh} \) and \( \hat{D}_{Sh3} \) that are the best, by a comfortable margin in this region. Finally \( \hat{D}_{Sh2} \), \( \hat{D}_{GEE} \) and \( \hat{D}_{AE} \) are comparable to \( \hat{D}_{aj2a} \). Again, this shows that the optimal estimator for this region, which was \( \hat{D}_{aj2a} \) in the study of [20], is no longer optimal as we increase the scale and dimensionality of the underlying characterization (this includes reducing the average \( q \)).

We also note that at our scale and dimensionality, \( \hat{D}_{Sh2} \) shows similar accuracy to the jackknife families, but with the exception of \( \hat{D}_{aj2} \), for low to medium \( \gamma^2 \) (cf. [20], who do not exclude \( \hat{D}_{aj2} \)).

c) \( \gamma^2 > 50 \):

For both low and high sampling fractions, the best estimators are the same as the three best estimators overall, namely \( \hat{D}_{aj2a} \), \( \hat{D}_{GEE} \), and \( \hat{D}_{AE} \); with \( \hat{D}_{aj2} \) being comparable. Among these, for low \( \frac{N}{\theta} \), \( \hat{D}_{GEE} \) is the best, while for high \( \frac{N}{\theta} \), \( \hat{D}_{aj2a} \) is the best. The reasoning...
Fig. 7: The “flatter” bias surfaces of $\hat{D}_{ij1}$, $\hat{D}_{ij2a}$, $\hat{D}_{Shh}$, and $\hat{D}_{AE}$ shown with more detail. Only $N = 1B$, and $\frac{N}{\sqrt{T}} \in [10, 100, 1000]$ shown. Note also that these estimators are relatively agnostic to $N$, and show regularity in variation with $\frac{N}{\sqrt{T}}$.

Fig. 8: The three most consistent estimators at $q = 0.02$, which is the lowest $q$ at which $D_{AE}$ has maximum ratio error at most 5, see Table V (although $\hat{D}_{ij2a}$ offers the same at $q = 0.01$). Note that at mid-high $\frac{N}{\sqrt{T}}$ there is a region of significant low-mid skew bias for $\hat{D}_{GEE}$, described in text.

that $\hat{D}_{Shh}$ is a good estimator when $\gamma^2$ is large since its derivation does not depend on a Taylor-series expansion in $\gamma^2$ [19] does not find evidence. Indeed, in the very high $\gamma^2$ regions of mid-Zipfian skew, all the Schlosser estimators do very poorly.

$\hat{D}_{Shh}$ is the most accurate among the Schlossers in this region, by a considerable margin, as opposed to $\hat{D}_{Shh}$ which was declared the best estimator in this range in [20].

D. Smoothing versus Stabilization in Jackknives

Our study also validates, at a higher scale, some observations made in [20]. Namely, stabilization works far more effectively than smoothing. In the mid-skew regions, the second order jackknives exhibit poor performance, while the stabilized $\hat{D}_{ij2a}$ retains acceptable performance.

E. Sampling Percentages Required

In today’s large commercial databases, hundreds of millions of rows are standard, and billions of rows are frequently encountered. Therefore, the cost of sampling is significant, and is a major design consideration. In our experience, it is the second question asked by designers behind the choice of estimator. The “default” value of sampling fraction in industrial databases is 0.02, but there is increasing pressure to reduce this as database sizes increase.

From our experience of working closely with query optimizer designers, we observed a wide gulf between the accuracy that distinct value estimators can provide, and the accuracy that query optimizer designers expect. It is important to understand that without essentially scanning the entire relation, we cannot hope to achieve the accuracies that are expected for arbitrary datasets. We feel that this is a communication gap that should be addressed. The published literature that deals with required accuracies [1] is now fairly old, and was suitable to the small tables encountered then. Today’s query optimizers should be designed with the understanding that obtaining ratio errors of less than 10% consistently, with the sampling fractions that are feasible for such large tables, is itself a non-trivial problem. For instance, the ratio error bound on the GEE — the only estimator to have error bounds — at even 10% sampling, is $\sqrt{10} \sim 3.2$. At the more feasible sampling rate of 2%, this bound is $\sqrt{50} \sim 7.1$ and at the sampling rate of 1%, it is 10. Note that even a ratio error of 3.2 is enough at large database sizes to cause the query optimizer to formulate highly inefficient plans.

In Table V, we provide the best estimator as a function of both maximum and average ratio error, for the ratio error values of two and five. Table V indicates that if $q = 0.01$, then $\hat{D}_{GEE}$ or $\hat{D}_{ij2a}$ may be the better choice over $\hat{D}_{AE}$. However, we should note that $\hat{D}_{GEE}$ has a region of relatively poor performance at very low sampling fractions, and so we should not use $\hat{D}_{GEE}$ for low-mid skew data if the sampling fraction is to be dropped below 0.005. Fig. 8 shows the three most consistent estimators at $q = 0.02$.

F. Ease of Implementation

There is no significant difference in the ease of implementation among the estimators (besides our simplification for $\hat{D}_{ij2a}$ described in the related work section online). While some of the estimators require storing values of $f_i$, $i \geq 2$, the $\hat{D}_{GEE}$ requires only the storage of $f_1$. However, this is not a factor in today’s systems. Likewise, the iterations required for the Newton-Raphson method in the $\hat{D}_{AE}$ are far too few to be a design factor. In all our experiments, Newton-Raphson converged in less than ten iterations. In summary, the choice of estimator depends only on accuracy, and not on implementation considerations.

VI. Conclusion and Future Work

A. Conclusions

Our high-level conclusion is that there exist stable patterns of relative behavior of distinct value estimators over populations of real-world size and frequently occurring distributions. This provides us with the best characterization yet of the answer to “which estimators do well on which datasets?” and therefore also sheds light on the question of “what properties of datasets allow certain estimators to do well on them?” We have proposed a systematic methodology, which integrates visualization, for characterizing errors of distinct value estimators.

Some of the conclusions that arise from our study are below.

1) The parameter $\frac{N}{\sqrt{T}}$ is critical in characterizing datasets.
### TABLE II: Ratio-error vs. $\theta$ and $\frac{N}{X}$, for low and high sampling fractions. Note that the high and low ranges for $\frac{N}{X}$ overlap on the mid-value of $\frac{N}{X} = 100$.

| Skew | $\frac{N}{X}$ | $0.001 \leq q \leq 0.005$ | $0.01 \leq q \leq 0.1$ |
|------|---------------|-----------------------------|---------------------------|
| Skew | $\frac{N}{X}$ | $D_{a1j}$ | $D_{a2j}$ | $D_{a3j}$ | $D_{a4j}$ | $D_{a5j}$ | $D_{a1a}$ | $D_{a2a}$ | $D_{a3a}$ | $D_{a4a}$ | $D_{a5a}$ |
| 0 $\leq \theta \leq 2$ | $\frac{N}{X} \leq 100$ | 12.99 | 8.24 | 3.33 | 2.43 | 2.69 | 28.23 | 4.16 | 1.23 |
| 1.5 $\leq \theta \leq 2$ | $\frac{N}{X} \leq 100$ | 38.56 | 5.8 | 38.69 | 9.24 | 8.06 | 8.88 | 8.14 | 2.07 |
| 0 $\leq \theta \leq 1$ | $\frac{N}{X} \geq 100$ | 3.04 | 19.89 | 3.46 | 1.28 | 1.22 | 19.26 | 1.63 | 1.13 |
| 1.5 $\leq \theta \leq 2$ | $\frac{N}{X} \geq 100$ | 47.8 | 6.57 | 28.5 | 7.23 | 6.23 | 9.46 | 4.27 | 1.83 |

### TABLE III: Percentage bias vs. $\theta$ and $\frac{N}{X}$ for low and high sampling fractions. Shaded cells indicate a change in the sign of the bias in the individual values within that region.

| Skew | $\frac{N}{X}$ | $0.001 \leq q \leq 0.005$ | $0.01 \leq q \leq 0.1$ |
|------|---------------|-----------------------------|---------------------------|
| Skew | $\frac{N}{X}$ | $D_{b1j}$ | $D_{b2j}$ | $D_{b3j}$ | $D_{b4j}$ | $D_{b5j}$ | $D_{b1a}$ | $D_{b2a}$ | $D_{b3a}$ | $D_{b4a}$ | $D_{b5a}$ |
| 0 $\leq \theta \leq 1$ | $\frac{N}{X} \leq 100$ | -50.01 | 680.77 | 181.94 | -34.67 | 34.07 | 2,719.94 | 310.84 | 6.31 |
| 1.5 $\leq \theta \leq 2$ | $\frac{N}{X} \leq 100$ | -96.19 | 68.14 | -96.22 | -86.06 | 83.27 | 610.59 | -83.36 | -37.33 |
| 0 $\leq \theta \leq 1$ | $\frac{N}{X} \geq 100$ | 32.34 | 1,886.06 | 241.56 | 1.38 | 10.24 | 1,825.79 | 57.9 | 12.77 |
| 1.5 $\leq \theta \leq 2$ | $\frac{N}{X} \geq 100$ | -94.36 | 159.33 | -94.41 | 79.67 | 77.14 | 661.2 | -77.24 | -28.33 |

### TABLE IV: Percentage RMSE vs. $\gamma^2$ and $\frac{N}{X}$ for low and high sampling fractions

| $\gamma^2$ | $\frac{N}{X}$ | $0.001 \leq q \leq 0.005$ | $0.01 \leq q \leq 0.1$ |
|------------|---------------|-----------------------------|---------------------------|
| $\gamma^2$ | $\frac{N}{X}$ | $D_{c1j}$ | $D_{c2j}$ | $D_{c3j}$ | $D_{c4j}$ | $D_{c5j}$ | $D_{c1a}$ | $D_{c2a}$ | $D_{c3a}$ | $D_{c4a}$ | $D_{c5a}$ |
| 0 $\leq \gamma^2 < 1$ | $\frac{N}{X} \leq 100$ | 63.68 | 2,293.07 | 575.77 | 48.91 | 45.78 | 5,816.36 | 705.15 | 26.77 |
| 1 $\gamma^2 > 50$ | $\frac{N}{X} \geq 100$ | 96.21 | 228.61 | 96.24 | 86.3 | 83.7 | 1,571.14 | 83.79 | 47.92 |
| $\gamma^2 > 50$ | $\frac{N}{X} \geq 100$ | 95.63 | 3,300.52 | 30.57 | 80.5 | 80.49 | 3,956.07 | 708.07 | 45.69 |
| $\gamma^2 > 50$ | $\frac{N}{X} \geq 100$ | 94.48 | 467.77 | 94.52 | 81.3 | 78.61 | 1,577.23 | 78.7 | 43.14 |
| $\gamma^2 > 50$ | $\frac{N}{X} \geq 100$ | 89.84 | 5,083.31 | 640.39 | 70.98 | 68.54 | 4,775.26 | 242.68 | 42.17 |

### TABLE V: Sampling percentage required for maximum/average ratio error of 2 and 5

| Error | $D_{a1j}$ | $D_{a2j}$ | $D_{a3j}$ | $D_{a4j}$ | $D_{a5j}$ | $D_{b1j}$ | $D_{b2j}$ | $D_{b3j}$ | $D_{b4j}$ | $D_{b5j}$ | $D_{c1j}$ | $D_{c2j}$ | $D_{c3j}$ | $D_{c4j}$ | $D_{c5j}$ | $D_{d1j}$ | $D_{d2j}$ |
|-------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Max 5 | 0.1 | na | 0.01 | 0.1 | 0.05 | 0.01 | 0.02 | na | na |
| Max 2 | na | na | 0.05 | na | na | na | 0.1 | 0.1 | na |
| Avg 5 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| Avg 2 | na | na | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
2) Scale effects cannot be ignored: conclusions drawn through studies on small datasets can lead to erroneous choices for large real world datasets.

3) Three distinct value estimators — \( \hat{\theta}_{GEE}, \hat{\theta}_{AB}, \hat{\theta}_{GEE, GEE} \) are the best estimators across a wide range of parameters, and at large scales. Each represents a different approach to estimation. Moreover, the choice of estimator should be informed by finer-grained behavior w.r.t. parameters.

4) Estimators obtained through “second order” methods that require estimation of \( \gamma^2 \) are highly inaccurate, especially as \( \gamma^2 \) increases.

5) A sampling fraction of \( 2\% \) may be considered optimum in the sense that it is at the high end of what may be considered feasible for today’s large commercial databases, and at the low end of obtaining acceptable ratio errors provided good estimators are chosen.

6) Visualization of error patterns is a powerful methodology to gain insight into the behavior of distinct value estimators.

This paper was written for both the practitioner and the researcher. The practitioner seeks to answer questions such as sampling percentage, choice of estimator for his dataset, etc. The researcher will find the relative performance of estimators a source of challenging problems: why do certain estimators behave in certain ways relative to one another for certain population parameters.

B. Future Work

It has been remarked before [19] that perhaps the reason that there is relatively less literature on the distinct values problem in the database community is that the problem is hard, and our understanding of it is limited. We hope that with the characterization we provide in this work, there will be more clarity on the accuracies we can expect for various datasets of commercial importance. An important question is: how well can we estimate \( \hat{\alpha}^2 \) for a dataset? Can we then use the resulting better understanding of errors incurred to make the query optimizer more robust?

We also hope that the understanding of the relative performances of estimator families that emerges from this study should lead to better hybrid estimators.

Finally, the empirical characterization of this work could be used to improve the theory and methods for existing families of estimators. For example, why is \( \hat{\alpha}^2 \) such a critical parameter for error patterns? Can we design estimators that operate very well for specified ranges of \( \hat{\alpha} \)? Can a modified form of stabilization be used on other estimators in light of its effectiveness in \( \hat{\theta}_{GEE, GEE} \)? Can we understand the “slope” to “hat” transitions that happen in multiple estimators?

VII. Supplementary Material

A large amount of data was generated in our study. Due to space considerations, the following material is not included in the main body of the paper, but will be provided in the supplementary material made available online via arXiv.

1) Grids for 2D bias plots for the following families: jackknife, Schlosser, \( \hat{\theta}_{GEE} \) and \( \hat{\theta}_{AB} \) at each sampling fraction \( q \in [0.001, 0.005, 0.01, 0.02, 0.1] \) excluding those that appear in the main body of the paper.

2) Grids for the top three estimators — \( \hat{\theta}_{GEE, GEE}, \hat{\theta}_{AB}, \hat{\theta}_{GEE, GEE} \) for each sampling fraction \( q \in [0.001, 0.005, 0.01, 0.1] \).

3) Grid of 3D bias surfaces and 2D plots for \( \hat{\alpha}_{CL2} \).

Dedication

This study was carried out in 2011: the 40th year of the genocide of two million Hindus in 1971. This work is dedicated to their sacred memory, and especially to the women violated during that genocide; and also to Flt. Lt. Vijay Vasant Tambay.

References

[1] M. M. Astrahan, M. Schkolnick, and K.-Y. Whang. Approximating the number of unique values of an attribute without sorting. Inf. Syst., 12(1):11–15, 1987.

[2] K. Beyer, R. Gennis, P. J. Haas, B. Reinwald, and Y. Sismanis. Distinct-value synopses for multiset operations. Commun. ACM, 59(7):625–633, 1978.

[3] K. P. Burnham and W. S. Overton. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrics, 65(3):pp. 625–633, 1979.

[4] A. Chao. Nonparametric estimation of the number of classes in a population. Scandinavian Journal of Statistics, 14(4):pp. 265–270, 1984.

[5] A. Chao and S.-M. Lee. Estimating the number of classes via sample coverage. Journal of the American Statistical Association, 87(417):pp. 210–217, 1992.

[6] A. Chao, M.-C. Ma, and M. C. K. Yang. Stopping rules for recapture debugging with unequal failure rates. Biometrics, 52(1):pp. 193–201, 1996.

[7] M. Charikar, S. Chaudhuri, R. Motwani, and V. R. Narasayya. Towards estimation error guarantees for distinct values. In PODS, pages 268–279, 2000.

[8] S. Chaudhuri, R. Motwani, and V. R. Narasayya. Random sampling for histogram construction: How much is enough? In SIGMOD Conference, pages 436–447, 1998.

[9] J. N. Darroch and D. Ratcliff. A note on capture-recapture estimation. Biometrics, 36(1):pp. 149–153, 1980.

[10] W. W. Esty. Confidence intervals for the coverage of low coverage samples. The Annals of Statistics, 10(1):pp. 190–196, 1982.

[11] W. W. Esty. The efficiency of good’s nonparametric coverage estimator. The Annals of Statistics, 14(3):pp. 1257–1260, 1986.

[12] P. Flajolet and G. N. Martin. Probabilistic counting algorithms for data base applications. J. Comput. System Sci., 31(2):182–209, 1985. Special issue: Twenty-fourth annual symposium on the foundations of computer science (Tucson, Ariz., 1983).

[13] L. J. Good. The population frequencies of species and the estimation of population parameters. Biometrika, 40(3/4):pp. 237–264, 1953.

[14] L. J. Good and G. H. Toumlin. The number of new species, and the increase in population coverage, when a sample is increased. Biometrika, 43(1/2):pp. 45–63, 1956.

[15] L. A. Goodman. On the estimation of the number of classes in a population. Ann. Math. Statistics, 20:572–579, 1949.

[16] H. L. Gray and W. R. Schucany. The generalized jackknife statistic. Marcel Dekker Inc., New York, 1972. Statistics Textbooks and Monographs, Vol. 1.

[17] P. J. Haas, J. F. Naughton, S. Seshadri, and L. Stokes. Sampling-based estimation of the number of distinct values of an attribute. In VLDB, pages 311–322, 1995.

[18] P. J. Haas and L. Stokes. Estimating the number of classes in a finite population. Journal of the American Statistical Association, 93(444):pp. 1475–1487, 1998.

[19] W.-C. Hou, G. Osoyoglu, and B. K. Taneja. Statistical estimators for relational algebra expressions. In PODS, pages 276–287, 1988.

[20] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. Internet Math., 12:226–251, 2004.

[21] F. Mosteller. Questions and answers. American Statistician, 3:12–13, 1949.

[22] J. F. Naughton and S. Seshadri. On estimating the size of projections. In ICDT, pages 499–513, 1990.

[23] G. Osoyoglu, K. Du, A. Tjahjana, W.-C. Hou, and D. Y. Rowland. On estimating count, sum, and average. In DEIA, pages 406–412, 1991.

[24] H. E. Robbins. Estimating the total probability of the unobserved outcomes of an experiment. The Annals of Mathematical Statistics, 39(1):pp. 256–257, 1968.

[25] A. Schlosser. On estimation of the size of the dictionary of a long text on the basis of a sample. Engineering Cybernetics, 19:pp. 97–102, 1981.

[26] H. S. Sichel. Anatomy of the generalized inverse gaussian-poisson distribution with special applications to bibliometric studies. Inf. Process. Manage., 28(1):5–18, 1992.

[27] E. P. Smith and G. v. Belle. Nonparametric estimation of species richness. Biometrics, 40(1):pp. 119–129, 1984.

[28] Tabular Analysis. Data warehouse market forecast 2017-2022.