

Towards Estimation Error Guarantees for Distinct Values

Moses Charikar*
Stanford University

Surajit Chaudhuri †
Microsoft Research

Rajeev Motwani‡
Stanford University

Vivek Narasayya§
Microsoft Research

ABSTRACT

We consider the problem of estimating the number of distinct values in a column of a table. For large tables without an index on the column, random sampling appears to be the only scalable approach for estimating the number of distinct values. We establish a powerful negative result stating that no estimator can guarantee small error across all input distributions, unless it examines a large fraction of the input data. In fact, any estimator must incur a significant error on at least some of a natural class of distributions. We then provide a new estimator which is provably optimal, in that its error is guaranteed to essentially match our negative result. A drawback of this estimator is that while its worst-case error is reasonable, it does not necessarily give the best possible error bound on any given distribution. Therefore, we develop heuristic estimators that are optimized for a class of typical input distributions. While these estimators lack strong guarantees on distribution-independent worst-case error, our extensive empirical comparison indicate their effectiveness both on real data sets and on synthetic data sets.

1. INTRODUCTION

Efficient processing of complex queries over large volumes of data has taken on increased importance with the growing interest in decision support systems. A principled choice

*Department of Computer Science, Gates 4B, Stanford University, Stanford, CA 94305-9045. E-mail: moses@cs.stanford.edu. Supported by the Pierre and Christine Lamond Fellowship, an ARO MURI Grant DAAH04-96-1-0007 and NSF Grant IIS-9811904.

†Microsoft Research, One Microsoft Way, Redmond, WA 98052. E-mail: surajitc@microsoft.com

‡Department of Computer Science, Gates 4B, Stanford University, Stanford, CA 94305-9045. E-mail: rajeev@cs.stanford.edu. Supported by an ARO MURI Grant DAAH04-96-1-0007 and NSF Grant IIS-9811904.

§Microsoft Research, One Microsoft Way, Redmond, WA 98052. E-mail: viveknar@microsoft.com

of an execution plan by an optimizer depends heavily on the availability of statistical summaries such as *histograms* and the *number of distinct values* in a column for the tables referenced in a query. In particular, accuracy of distinct values estimation greatly impacts the ability of the query optimizer’s ability to generate good plans for SQL queries.

This paper is concerned with efficient and accurate distinct-values estimation. The traditional approach for distinct-values estimation in the absence of an index would be to scan the table, followed by a sort or a hash. However, in large datawarehouses, these traditional techniques can be prohibitively expensive. This leaves sampling-based estimation of the number of distinct values as the only remaining alternative that scales well with increasing data size. Unfortunately, while other statistical parameters such as histograms can be fairly accurately computed from small random samples, accurate distinct-values estimation has proved to be an extremely challenging task.

In this paper, we focus on the problem of accurately estimating the number of distinct values from *samples* of large tables. We begin by establishing a powerful negative result stating that no estimator can guarantee small error across all input distributions, unless it examines a large fraction of the input data. We then provide a new estimator which is provably optimal, in that its error is guaranteed to match our negative result within a small constant factor. A drawback of this estimator is that although its worst-case error (over *all* distributions) is fairly small, it does not leverage the knowledge of input distributions and therefore does not give the best possible error bound for that *specific* distribution. We develop heuristic estimators which, while lacking strong error guarantees on distribution-independent worst-case error, address the above concern. A key contribution of the paper is an extensive experimental comparison of the estimators that complements its theoretical contributions.

1.1 Related Work

This first thing to note about the distinct-values estimation problem is its apparent difficulty. In fact, as Haas et al [15] observe: “Identification of improved estimators is difficult. ... Unfortunately, analysis of distinct-value estimators is highly non-trivial, and few analytic results are available.”

In the statistics literature, the problem of estimating the number of species (or classes) is a well-studied problem that is equivalent to our distinct-values estimation problem; refer to Bunge and Fitzpatrick [3] for a survey.

This problem has also been studied extensively in the database literature, e.g., Hou, Ozsoyoglu, and Taneja [18];

19], Naughton and Seshadri [22], and Ozsoyoglu, Du, Tjahjana, Hou, and Rowland [25] present results applying the statistical estimators to data arising in database applications, with relatively poor results. The state-of-the-art is the work of Haas, Naughton, Seshadri, and Stokes [15], who devise several new estimators including a *hybrid estimator* that appear to outperform the aforementioned estimators. In recent work, Haas and Stokes [16] do an extensive study of several generalized jackknife estimators, relating previously known estimators and proposing new ones.

The results of [15] and [16] indicate that the estimators they propose typically outperform other known estimators, both for real and synthetic data sets. Consequently, we will compare our new estimators against the best estimators proposed in these works. The *hybrid estimator* of [15] first uses a chi-squared test to decide whether the data has low or high skew, applying a *smoothed jackknife* estimator in the former case and the *Shlosser* estimator in the latter case. In this paper, we will refer to it as HYBSKEW. The estimator \hat{D}_{hybrid} recommended by [16] is similar in spirit to HYBSKEW. \hat{D}_{hybrid} chooses between one of three estimators (one of them being a modified Shlosser estimator) based on an estimate of a certain coefficient of variation of class sizes. In this paper, we will refer to this estimator as HYBVAR. The estimator \hat{D}_{duj2a} is recommended by [16] as an alternative to hybrid estimators, i.e. estimators that choose from one of several different estimators, depending on some estimated parameter. We refer to this as DUJ2A. Based on their evaluation in [16], they recommend the estimators HYBVAR and DUJ2A.

In an earlier work (Chaudhuri, Motwani, and Narasayya [8]), we had touched upon the problem of distinct-values estimation, and indeed the current paper has its genesis in the results presented therein. The first of these results established the inherent difficulty of accurate distinct-values estimation based purely on *random samples*; the second was a heuristic estimator but *without any formal analysis*. Among the results presented in the current paper is a proof of optimality of a variant of the heuristic estimator proposed in [8]; in addition, we present significantly improved estimators and extensive experimental results demonstrating the superior performance of our new estimators.

We note that there are some hashing techniques called “probabilistic counting” which can help alleviate the memory requirements. While these methods reduce memory requirements at the cost of introducing imprecision, they still involve a full scan of the table and remain infeasible for large tables from the point of view of computation time.

1.2 Overview and Discussion of Main Results

Before describing our results, we present a general framework for studying this problem by enumerating the features desired of an estimator.

Average Bias: The average value of the estimator should be close to the number of distinct values.

Variance: The estimator should be consistent in terms of having low variance.

Error Guarantee: There should be an analytic guarantee on the degree of error of the estimator.

Low Sampling: The estimator should guarantee small error for relatively low sample sizes. It should converge

consistently and rapidly to the true value as the sample size is increased.

Distribution-independence: The estimator should not rely on any assumptions about the data.

Confidence: An estimator should indicate the confidence in its estimate and its variance.

Scalability: Error should not increase with increased data size, for a fixed sampling fraction.

Most estimators that have been proposed earlier fail completely on all counts. Almost all of them are designed for data satisfying questionable assumptions (e.g., uniformity), and certainly do not provide any error guarantees. Even worse, they do not have low average bias or variance. Some notable exceptions are the estimators of [15] and [16]. We observe the following positive aspects of their estimators HYBSKEW, HYBVAR and DUJ2A: they exhibit much lower average bias than the previous estimators; and, they perform reasonably well for both low-skew and high-skew data distributions. On the other hand, they fall short on the other counts. There are no analytic guarantees on their errors. Due to the hybrid nature of HYBSKEW and HYBVAR, their performance is inconsistent in several ways: they do not give consistent error-reduction as the sample size increases; and, they have high variance, particularly for low sampling rates. Finally, none of the estimators provide any indication of the likely error in their estimate.

After some preliminaries in Section 2, we begin in Section 3 by strengthening the negative result mentioned in Chaudhuri, Motwani, and Narasayya [8]. The earlier result stated that if a *random sample* of size r is chosen from a table of size n , then any estimator provided *only with the random sample* must incur error proportional to $\sqrt{n/r}$, for some input distributions. However, it did not rule out the possibility of good estimators based on a strategy more sophisticated than oblivious random sampling, e.g., an adaptive strategy that chooses a set of values, examines the values and then decides where to look next, and so on. Our new result shows that the same error must be incurred by *the most general possible* family of estimators which are allowed to use any (possibly randomized) strategy to select a sequence of r values to examine in the input table, where the choice of the values examined could be based on the values seen so far. We also discuss the proof, elucidating intuition behind the main difficulty in the distinct-values estimation problem.

In Section 4, we consider a variant of the heuristic estimator from Chaudhuri, Motwani, and Narasayya [8]. We provide an analytic result establishing that this estimator, called GEE (Guaranteed-Error Estimator), is *optimal* in that it achieves an error bound proportional to $\sqrt{n/r}$, thereby matching the lower bound. An interesting side-effect of the proof is the derivation of lower and upper bounds on the number of distinct values. Thus, the GEE estimator has a novel property that not only does it return an estimate, but it provides also a confidence interval surrounding the estimate that with high probability, contains the actual number of distinct values. The width of this interval reflects the confidence in the estimate as well as its likely variance. We feel that such measures of confidence should be required of all estimators.

Our GEE estimator must cater to all possible input distributions to achieve analytical guarantees on its error. Consequently, it is possible that GEE does not outperform HYBSKEW on certain specific input distributions. Indeed, our experiments show that while GEE outperforms HYBSKEW in many cases, it compares poorly with HYBSKEW in the particular case of low-skew data with a large number of distinct values. We observe that in the latter distributions HYBSKEW never invokes the Shlosser estimator, preferring instead the smoothed jackknife estimator. Based on this insight, we present in Section 5 a heuristic estimator called HYBGEE (for HYBSKEW with GEE) which substitutes the GEE estimator for the Shlosser estimator in HYBSKEW. The HYBGEE estimator consistently outperforms HYBSKEW across all data distributions and compares favorably with HYBVAR and DUJ2A, but unfortunately it has no error guarantees unlike in the case of GEE. We then apply the intuition behind the hybrid estimator in a more principled manner to devise a new estimator called AE (Adaptive Estimator). This is a refinement of GEE in that it attempts to adapt certain parameters in GEE to the input distribution to avoid the bad performance on low-skew data with a large number of distinct values. We give an analytical justification for the formulation of AE in Section 5. We conjecture that an analytical error guarantee is possible for AE, and leave it as an open problem for future work. Modulo this unproven guarantee, the AE estimator has all the features we desire of an estimator, including the confidence interval that carries over from the GEE estimator.

In Section 6 we provide experimental comparison of the estimations by these estimators on real and synthetic data sets. In particular, the experiments validate the claimed behavior of the new estimators GEE, HYBGEE, and AE.

2. PRELIMINARIES

Let n be the number of rows in the table. We focus on a specific column of the table. Suppose there are D distinct values in this column, say $\{1, 2, \dots, D\}$. Let n_i denote the number of times value i occurs in the column and define $p_i = n_i/n$; then, $\sum_{i=1}^D n_i = n$ and $\sum_{i=1}^D p_i = 1$. When estimating the number of distinct values in a specific column, we can ignore the remaining columns in the table and simply view it as a (multi)set of values from the column of interest.

The estimators we consider are based on examining a random sample of r tuples chosen uniformly at random from the table. The sampling could be with or without replacement; we will consider both possibilities and assume use of standard efficient schemes for sampling from a table [23; 24; 9]. Let d be the number of distinct values in the sample, and let f_i be the number of values that exactly occur i times in the sample; then, $d = \sum_{i=1}^r f_i$. Clearly, $d \leq D \leq n$. We will use \hat{D} to denote an estimate of the number of distinct values. We will apply sanity bounds to all estimators to ensure that $d \leq \hat{D} \leq n$; i.e., if $\hat{D} > n$ we set \hat{D} to n , and if $\hat{D} < d$ we set \hat{D} to d .

To evaluate the performance of estimators, we use the *multiplicative error* of \hat{D} with respect to D . We refer to it as the *ratio error* and define it as follows:

$$\text{error}(\hat{D}) = \begin{cases} \frac{D}{\hat{D}} & \text{if } D \geq \hat{D} \\ \frac{\hat{D}}{D} & \text{if } D < \hat{D} \end{cases}$$

That is, the error is simply the *ratio* of the estimator and the true number of distinct values D , where the ratio is inverted if necessary to ensure that the error is always greater than 1. Equivalently, the estimator \hat{D} has ratio error at most α if and only if $D/\alpha \leq \hat{D} \leq \alpha D$.

A common error measure is the *relative error* which is defined to be $(\hat{D} - D)/D$. This measures the *fractional additive* error of the estimator and is similar to the error measure used by Haas et al [15]. We do not favor the use of relative error as it is fairly misleading in comparing an overestimate with an underestimate. In any case, it is fairly easy to translate between the two error measures, and the choice does not affect the qualitative interpretation of our analysis or experiments.

3. NEGATIVE RESULT FOR ESTIMATION

Olken [23] points out that all known estimators give exceedingly large errors on at least some input data. Chaudhuri et al [8] explained this observation by showing that large errors are unavoidable for estimates derived only from random samples. They considered the problem of estimating the number of distinct values from a uniform random sample of the data and showed that in this restricted setting it is impossible for an estimator to guarantee a small error unless the sample size is very close to the size of the table itself. We generalize this result and prove a similar negative result for the the most general possible class of estimators that, instead of being restricted to merely a random sampling, are allowed to use *any* (possibly randomized) strategy to select a sequence of r values to examine in the the input table. Such estimators could even pick a sequence of r tuples *adaptively*, i.e., pick the next tuple to examine based on the values of the previously-examined tuples.

Previously, Haas et al [15] observed that no known estimator behaves well on all data distributions primarily because their performance is sensitive to the skew of the data. This situation can be easily explained in light of our results. Our lower bound on the error stems in part from the difficulty in distinguishing between low-skew and high-skew data. In fact, the proof hinges on the inability to distinguish between two specific scenarios: one where there are few distinct values of very high frequency, and another where there are a few high-frequency values together with a large number of very low-frequency values (thus having a large number of distinct values). The first scenario has low skew with few distinct values and the second one has high skew with a large number of distinct values. We prove that any possible estimator must incur large error on at least one of these two scenarios.

THEOREM 1. *Consider any (possibly adaptive and randomized) estimator \hat{D} for the number of distinct values D that examines at most r rows in a table with n rows. Then, for any $\gamma > e^{-r}$, there exists a choice of the input data such that with probability at least γ ,*

$$\text{error}(\hat{D}) \geq \sqrt{\frac{n-r}{2r} \ln \frac{1}{\gamma}}.$$

PROOF. Consider the following two distinct scenarios for the column C of interest:

Scenario A: There is only one distinct value in C , say x .

Scenario B: There are $k+1$ distinct values $\{x, y_1, \dots, y_k\}$, of which x appears $n-k$ times and each y_i appears once. Further, the y_i 's are placed in k rows chosen uniformly at random.

Consider Scenario B and let $C(X)$ denote the value appearing in column C at row X . Suppose now that an estimator E examines a sequence of r distinct rows, say X_1, \dots, X_r , where the choice of X_i could depend arbitrarily on $C(X_1), \dots, C(X_{i-1})$, and may be randomized. We claim:

$$\begin{aligned} \text{Prob}[C(X_i) = x \mid C(X_1) = x, C(X_2) = x, \dots, C(X_{i-1}) = x] \\ = \frac{n-i-k+1}{n-i+1}. \end{aligned}$$

This is because so long as the first $i-1$ values are all x , the estimator gets no information about whether it is in Scenario A or B, and the y_i values are randomly distributed in the remaining $n-i+1$ rows of C . Let \mathcal{E} denote the event that $C(X_1) = x, C(X_2) = x, \dots, C(X_r) = x$. Then, using a standard conditional probability computation [21], we obtain that:

$$\begin{aligned} \text{Prob}[\mathcal{E}] &= \prod_{i=1}^r \text{Prob}[C(X_i) \mid C(X_1) = x, \dots, C(X_{i-1}) = x] \\ &= \prod_{i=1}^r \frac{n-i-k+1}{n-i+1} \geq \left(\frac{n-r-k}{n-r} \right)^r \\ &= \left(1 - \frac{k}{n-r} \right)^r \geq e^{-\frac{2kr}{n-r}}. \end{aligned}$$

The last inequality follows from the claim that $1-z \geq e^{-2z}$ for $0 \leq z \leq 1/2$, which is easily verified. Note that the only constraint on k is that $k+1 \leq n$, hence for $\gamma \geq 1/e^r$, we may select

$$k = \frac{n-r}{2r} \ln \frac{1}{\gamma}.$$

For this choice of k , we obtain that $\text{Prob}[\mathcal{E}] \geq \gamma$. Note that we also need the condition that $z = k/(n-r) \leq 1/2$, which is valid for our choice of k and γ .

Thus, when C is as in Scenario B, with probability at least γ the estimator will obtain r copies of the same value x . On the other hand, in Scenario A the set of rows examined by the estimator will always contain r copies of the same value. Assume that $C(X_1) = x, C(X_2) = x, \dots, C(X_r) = x$, let α be the value returned by the estimator E in that case. Then, the error with respect to Scenario A is α and the error with respect to Scenario B is k/α . If $\alpha \geq \sqrt{k}$, then the error with respect to Scenario A is at least \sqrt{k} ; otherwise, the error with respect to Scenario B is at least \sqrt{k} . In either case, the estimator E incurs an error \sqrt{k} , implying the desired result. \square

Let us compare our worst case bounds with the maximum error of the estimators considered in Haas et al [15]. For a sampling fraction of 20%, i.e., $r = 0.2n$, they observed a maximum error of 1.58 for Shlosser's estimator, 2.86 for the smoothed jackknife estimator and 1.42 for the Hybrid estimator. Setting $\gamma = 0.5$ in our lower bound, we obtain that there exists a scenario in which the error is at least 1.18 with probability 1/2. Thus, our error bounds seem to be close to the errors observed in real experiments, even though our results give worst-case bounds and the data in

the experiments may have had special structure that could have been used to obtain better estimates.

4. AN OPTIMAL ESTIMATOR

We analyze a variant of an estimator proposed by Chaudhuri et al [8] and prove that, in fact, it has optimal error, matching the lower bound proved in Theorem 1 within a small constant factor. Our estimator is called GEE (for Guaranteed-Error Estimator) and is defined as follows for a random sample of size r :

$$\hat{D} = \sqrt{\frac{n}{r}} f_1 + \sum_{j=2}^r f_j.$$

where f_i 's are as defined in Section 2.

We begin by giving the intuition behind GEE. We can view the set that is sampled as being composed of high frequency and low frequency values; here "high" and "low" are relative terms that depend on the sampling fraction. The high frequency values get picked up in the random sample almost surely. We can think of the values that appear more than once in the random sample as being the values of high frequency; we need to count them only once each in our estimate of the number of distinct values. We can think of the singleton values in the sample as representing the low frequency values; however, the random sample contains only some of these low frequency values. Roughly, the f_1 singleton values represent a fraction f_1/r of the whole set, i.e., a total of nf_1/r values. This set contains at least f_1 values and could contain as many as nf_1/r values. We take the contribution of f_1 to our estimate to be the geometric mean of these two extreme bounds, in order to minimize the ratio error.

We now formally analyze the estimator GEE. For clarity of exposition, we present the proof only for the case where GEE samples with replacement, but our proof can be extended to sampling without replacement. We also restrict the analysis to showing only that the *expected* error of GEE is small; however, we can extend our proofs to establishing the stronger result that GEE has small error *with high probability*. Observe that the ratio error of GEE in the following theorem is within a small constant (roughly e) of the lower bound in Theorem 1.

THEOREM 2. *The expected ratio error of GEE is $O(\sqrt{n/r})$ when it samples r values from any possible input of size n .*

PROOF. When we pick a random value from a column C , the probability that we pick i is p_i . When we pick r values at random from C , the probability that we do not pick i is $(1-p_i)^r$. Hence the probability that i belongs to the random sample is $1-(1-p_i)^r$. The probability that i occurs exactly once in the random sample is $p_i r (1-p_i)^{r-1}$. Now, the number of distinct elements d in a random sample of size r is a random variable with expectation $E[d] = \sum_{i=1}^D 1 - (1-p_i)^r$. Let $x_i = 1 - (1-p_i)^r$, then $E[d] = \sum_{i=1}^D x_i$. Also f_1 , the number of singleton values in a random sample of size r , is a random variable with expectation $E[f_1] = \sum_{i=1}^D p_i r (1-p_i)^{r-1}$. Let $y_i = p_i r (1-p_i)^{r-1}$, then $E[f_1] = \sum_{i=1}^D y_i$. The estimator GEE returns the value $d + (\sqrt{\frac{n}{r}} - 1) \cdot f_1$. Thus, the expected value returned by this estimator is $E[\text{GEE}] = \sum_{i=1}^D [x_i + (\sqrt{\frac{n}{r}} - 1) y_i]$. Note that the actual number of distinct values is $\sum_{i=1}^D 1$. We will relate the individual terms

in the two summations by showing that $x_i + (\sqrt{\frac{n}{r}} - 1) y_i$ is within a multiplicative factor of 1, in particular between $\frac{1}{e}\sqrt{\frac{n}{r}}(1 - o(1))$ and $\sqrt{\frac{n}{r}}$. This will prove that, the expected value of estimator GEE is within a factor $e\sqrt{\frac{n}{r}}(1 + o(1))$ of the correct answer. We consider two cases:

Case 1: $p_i \geq \frac{1}{r}$.

In this case $1 > x_i > 1 - \frac{1}{e}$ and $y_i < \frac{1}{e}$. Thus, $x_i + (\sqrt{\frac{n}{r}} - 1) y_i$ is at least $1 - \frac{1}{e}$ and at most $\sqrt{\frac{n}{r}}$.

Case 2: $p_i < \frac{1}{r}$.

Note that $p_i \geq \frac{1}{n}$. In this case $0 < x_i < 1 - \frac{1}{e}$ and $\frac{x_i}{en} \leq y_i \leq \frac{1}{e}$. Again, $x_i + (\sqrt{\frac{n}{r}} - 1) y_i$ is within the required bounds. \square

Together with the estimator GEE, we can also provide upper and lower bounds on the number of distinct values. This gives some idea about the interval in which the true number of distinct values lies and the confidence with which we can estimate it. Our lower bound called LOWER is simply $d = \sum f_i$, the number of distinct values in the sample; clearly, this is always a valid lower bound. The upper bound, called UPPER is $\sum_{i>1} f_i + \frac{n}{r} f_1$. The analysis of the estimator GEE can be used to show that with high probability, the value of UPPER is greater than D , the actual number of distinct values. We omit the details in this extended abstract. The experiments confirm that the actual number of distinct values always lies in the interval [LOWER,UPPER]; in fact, the size of the interval [LOWER,UPPER] decreases sharply with increasing r , indicating that the estimates rapidly converge to D with increasing sample sizes (see Section 6 for details on this experiment).

5. HEURISTIC ESTIMATORS

The HYBSKEW estimator [15] is a combination of two estimators, the smoothed jackknife estimator and Shlosser's estimator. It was observed that the smoothed jackknife estimator performed well on data with low skew, while Shlosser's estimator performed well on data with high skew. To exploit this complementarity, HYBSKEW first uses the standard χ^2 test on the random sample to probabilistically estimate whether the data has high skew or low skew, resorting to Shlosser's estimator in the former case and the smoothed jackknife estimator in the latter case. We first show how to improve HYBSKEW, and then analytically derive an improved version of GEE.

5.1 The Modified Hybrid Estimator HYBGEE

From the definition of GEE and the intuition behind the way the definition was derived, we see that GEE only errs on the low frequency elements; the number of high frequency elements are estimated fairly accurately. This would imply that GEE does well in cases when f_1 is small relative to d , i.e., when the number of low frequency elements is small relative to the total number of distinct values. Thus, GEE should perform well for data with high skew or with relatively few low frequency elements, and this is borne out by the results of our experiments. By the same reasoning, we expect that GEE will not perform as well (in fact be a severe underestimate) for data which has *both* low skew and a large number of distinct values; again, this is borne out by our experimental results. In the case of high-skew synthetic data, and in fact on all real-world data, we found that

GEE outperforms the Shlosser Estimator. This suggests a modified version, called HYBGEE (for Hybrid with GEE), of the HYBSKEW estimator which substitutes GEE for the Shlosser estimator in the case of high-skew data. As expected, our experimental results show that HYBGEE gives a significant reduction in error compared to HYBSKEW. The experimental results are described in Section 6.

5.2 The Adaptive Estimator AE

One undesirable aspect of both HYBSKEW and the modified version HYBGEE is that it combines the results of two very different estimators; one of the two is selected depending on a test designed to measure the skew of the data. As a result, for data that is neither clearly high-skew or clearly low-skew, the value returned by the estimator is the value of one of the two estimators where the choice is somewhat arbitrary depending on the specific test used to measure skew. Usually, the values of the two estimators are very different. This results in instability in the final value of the estimator, resulting in high variance. In fact, it is possible that for the same data, some random samples result in the choice of one estimator while others cause the other to be chosen. In order to avoid this, it is desirable to have an estimator which has a uniform behavior.

Also, while the new estimator HYBGEE outperforms HYBSKEW, it does not have the desired property of having a guaranteed error bound as in the case of GEE. The question arises: Can we devise a new estimator which has guaranteed error and does not suffer from the same problem as GEE for low-skew data? The reason behind this problem with GEE is that it estimates the number of low frequency elements incorrectly. It fixes the coefficient of f_1 to be $\sqrt{n/r}$ which turns out to be too low for low-skew data with a large number of distinct values. This suggests devising an estimator which would select the coefficient for f_1 based on the information gathered about the data distribution in the random sample. Of course, since the only estimator for which an error guarantee can be provided is GEE, which is a consequence of the functional form of GEE, we choose to consider only estimators of the form $\hat{D} = d + K f_1$. In fact this is the family of generalized jackknife estimators considered by Haas and Stokes [16]. The insight derived from the preceding discussion suggests selecting the value of K adaptively based on an analysis of the random sample. One constraint on K derives from the desired feature that \hat{D} be an unbiased estimator of D , or that $E[\hat{D}] = D$. This implies that:

$$K = \frac{D - E[d]}{E[f_1]}. \quad (1)$$

Now, $E[d]$ and $E[f_1]$ can be expressed in terms of the p_i 's. (Recall that p_i is the fraction of the column containing value i .) This would give us an expression for K in terms of the p_i 's; unfortunately, we do not know the p_i values. Instead, we use the information about p_i 's that can be derived from the random sample. Since we have f_i values with frequency i in the sample, these f_i values are expected to have frequency approximately in/r in the column. An important contribution to the numerator of the expression comes from the low-frequency values, which are (say) m in number. We obtain an expression for K in terms of the f_i 's and m , the number of low-frequency values. We substitute this expression in (1). Note that D can also be expressed

in terms of f_i 's and m . This implies an equation in terms of m . Solving this equation, we obtain an estimate for m , the number of low frequency values which in turn gives us an estimate for the number of distinct values. We call the resulting estimator AE (for Adaptive Estimator); a detailed, formal derivation of AE is presented in the next section. A development of estimators along these lines is also present in [16]. A distinguishing aspect of our approach to AE compared to [16] is that we separate the contributions of the low frequency and high frequency elements, accounting for them differently. We conjecture that an analytical error guarantee can be established for AE, and leave it as an interesting open problem for the future. We note that a confidence interval can be provided for AE in exactly the same manner as for GEE.

5.3 Analytical Derivation of AE

Our goal is to derive a distinct value estimator of the form $\hat{D} = d + Kf_1$, where we know that

$$E[d] = \sum_{i=1}^D 1 - (1 - p_i)^r = D - \sum_{i=1}^D (1 - p_i)^r,$$

$$E[f_1] = \sum_{i=1}^D r \cdot p_i (1 - p_i)^{r-1},$$

$$E[\hat{D}] = D - \sum_{i=1}^D (1 - p_i)^r + K \sum_{i=1}^D r \cdot p_i (1 - p_i)^{r-1}.$$

Since we require $E[\hat{D}] = D$, it follows that

$$K = \frac{\sum_{i=1}^D (1 - p_i)^r}{\sum_{i=1}^D r \cdot p_i (1 - p_i)^{r-1}}.$$

Of course, we need to know the p_i values in order to evaluate the numerator and denominator of the above expression. We settle for an approximation, which is obtained as follows. Since we observed f_i values occurring i times in the random sample, we will take this to mean that there are f_i distinct values, each of which occupy a fraction $p = i/r$ of the set. This approximation should be fairly accurate for large values of i , but inaccurate for smaller i , especially $i = 1, 2$. Under this approximation, we conclude that

$$K = \frac{\sum_{i=1}^r \left(1 - \frac{i}{r}\right)^r f_i}{\sum_{i=1}^r i \left(1 - \frac{i}{r}\right)^{r-1} f_i}.$$

Now, let us attempt to improve the approximation. The earlier approximation was inaccurate for small i , particularly $i = 1, 2$. We work with the intuition that the values that contribute to the f_i 's for $i > 2$ are the *high-frequency* values in the set. The elements that contribute to f_1 and f_2 are representatives of the *low-frequency* values in the set, whose number could be much larger than $f_1 + f_2$. Let us estimate the number m of low-frequency values represented by each value that contributes to f_1 and f_2 . The fraction of the set occupied by these values is (approximately) $(f_1 + 2f_2)/r$. Thus the total number of distinct values is $D = d - f_1 - f_2 + m$. We further assume that each of them occupies an equal fraction of the set; thus, for each of these m values,

$p_i = \frac{f_1 + 2f_2}{rm}$. This results in the following estimate for K :

$$K = \frac{\sum_{i=3}^r \left(1 - \frac{i}{r}\right)^r f_i + m \left(1 - \frac{f_1 + 2f_2}{rm}\right)^r}{\sum_{i=3}^r i \left(1 - \frac{i}{r}\right)^{r-1} f_i + (f_1 + 2f_2) \left(1 - \frac{f_1 + 2f_2}{rm}\right)^{r-1}}.$$

Recall that we want $\hat{D} = d + Kf_1$ and that $D = d - f_1 - f_2 + m$; hence, $d - f_1 - f_2 + m = d + Kf_1$ or $m - f_1 - f_2 = Kf_1$. Substituting our estimate for K , we obtain the following expression for m :

$$m - f_1 - f_2 = f_1 \frac{\sum_{i=3}^r \left(1 - \frac{i}{r}\right)^r f_i + m \left(1 - \frac{f_1 + 2f_2}{rm}\right)^r}{\sum_{i=3}^r i \cdot \left(1 - \frac{i}{r}\right)^{r-1} f_i + (f_1 + 2f_2) \left(1 - \frac{f_1 + 2f_2}{rm}\right)^{r-1}}.$$

Employing standard approximations for the terms in the above equation, we obtain:

$$m - f_1 - f_2 = f_1 \frac{\sum_{i=3}^r e^{-i} f_i + m e^{-(f_1 + 2f_2)/m}}{\sum_{i=3}^r i e^{-i} f_i + (f_1 + 2f_2) e^{-(f_1 + 2f_2)/m}}.$$

Solving either of these equations to estimate m , using standard numerical methods, we obtain that the AE estimator is $\hat{D} = d + m - f_1 - f_2$. Of course, we apply sanity bounds to ensure that $d \leq \hat{D} \leq n$.

6. EXPERIMENTAL RESULTS

In this section, we report full results of our empirical analysis of the estimators GEE, HYBSKEW [15], AE, HYBGEE, DUJ2A [16] and HYBVAR [16]. We studied the performance of these estimators on three real-world data sets ‘‘Census,’’ ‘‘CoverType,’’ and ‘‘MSSales’’ with 32561, 581012 and 1996290 records respectively, as well as on synthetic data sets. The first two real-world data sets were obtained from the public domain site [13]. MSSales is an internal database in Microsoft which tracks sales of products by the company over a fiscal year. The synthetic data allowed us to vary the following data characteristics in a controlled manner:

- Data Skew.** We generated the data sets according to the generalized Zipfian distribution. We generated columns with Zipfian value $Z \in \{0, 1, 2, 3, 4\}$, where $Z = 0$ gives a uniform distribution (low skew), and $Z = 4$ is a highly-skewed distribution.
- Number of records.** We generated tables where the number of rows, n , ranged over $\{100K, 200K, \dots, 1000K\}$.
- Number of duplicates for each distinct value.** For each skew value, we generated columns with a varying number duplication factor. The number of copies ranged over $\{1, 10, 100, 1000\}$. For example, to generate a column with $n = 1,000,000$, $Z = 2$ and 100 duplicates, we generate Zipfian data for $n = 10,000$, and made 100 copies of each value.

The layout of data for each column was random. We achieved this by clustering the data on tuple-ids that were generated at random. In all our experiments, we varied the sampling fraction over 6 values: 0.2%, 0.4%, 0.8%, 1.6%, 3.2%, 6.4%. For each sampling fraction, we collect ten independent samples, and report the average error of each estimator over these ten samples. The experiments were conducted on an Intel Pentium III Xeon 550 Mhz processor with 256 MB RAM. The databases were stored in Microsoft SQL Server 7.0. We used existing functionality in SQL Server for obtaining a random sample without replacement of a specified sample size. We had to modify the server so that once the random sample was gathered, we obtained the following information: (a) number of distinct values in the sample; (b) f_i , for all $i \geq 1$; and, (c) sample skew [15].

Varying the Sampling Rate. In our first experiment we study the effect of varying the sampling fraction on each estimator for high-skew data ($Z=2$) and low-skew ($Z=0$). We report the results for $n = 1,000,000$ rows with duplication factor 100. Figures 1 and 2 show that on low-skew data HYBGEE performs as well as HYBSKEW, whereas on high-skew data HYBGEE significantly outperforms HYBSKEW. This is because for low skew, both HYBSKEW and HYBGEE use the smoothed jackknife estimator, whereas for high skew, HYBGEE uses GEE (and hence they overlap in the figure) while HYBSKEW uses Shlosser. As noted earlier, GEE does better than Shlosser on high-skew data. Finally, AE does consistently well on low-skew as well as high-skew data and outperforms all other estimators including DUJ2A and HYBVAR.

We also report the standard deviation of each estimator (as a fraction of the actual number of distinct values D) as the sampling fraction is varied, for both low-skew and high-skew data. Figures 3 and 4 show that the variance of all estimators decreases with increasing sample size. For certain sampling fractions in the low-skew case, GEE seems to exhibit slightly higher variance than the other estimators. However, note that the absolute value of the standard deviation for all estimators is already small in the low skew case. The HYBSKEW estimator has the highest variance amongst all estimators in the high-skew case. The variance of AE and DUJ2A are fairly small throughout whereas the variance of HYBVAR exhibits erratic behavior in the high skew case. Finally, for the same data, we show the confidence interval for GEE in Tables 1 and 2. As expected, we see that as the sampling fraction increases, the interval [LOWER,UPPER] rapidly converges to D .

Varying the Skew. Next, we study the effect of varying the skew on accuracy of the estimators for low (0.8%) and high (6.4%) sampling fractions. We present the results for $n = 1,000,000$ rows with 100 duplicates. Figures 5 and 6 show that HYBGEE consistently outperforms HYBSKEW for both low and high sampling fractions. Also, AE does better than all other estimators for the low sampling fraction with a ratio error extremely close to 1. For the high sampling fraction, AE has higher error than HYBGEE and HYBVAR but lower error than HYBSKEW and DUJ2A. We note that in this case the ratio error of all estimators is extremely close to 1, and it is hard to see any clear trend. Figure 6 also shows that GEE and HYBGEE have extremely small errors for high sampling fraction.

Varying the Duplication Factor. In our next experiment, we study the impact of duplication factor on the accuracy of the estimators. We present results for $n = 1,000,000$, $Z = 1$, and two sampling fractions (0.8% and 6.4%). Figures 7 and 8 show that in both cases, HYBGEE significantly outperforms HYBSKEW over the entire range of duplication factors. Also, except in the case of no duplicates for the low sampling fraction, AE is better than both HYBGEE and HYBSKEW throughout. DUJ2A and HYBVAR appear to perform well in the low sampling case whereas in the high sampling case HYBVAR has high error for the no duplicates data set. In general, we expect the error of the estimators to decrease as the duplication factor increases. This is because for large duplication factors, the random sample will almost surely contain all the distinct values in the column. This is roughly what we observe. A notable exception is HYBSKEW for low sampling fraction: its error goes up significantly as the duplication factor is increased from 1 to 10. This is because of the particular form of Shlosser’s estimator and the (invalid) assumptions made in its derivation.

Scaleup. The goal of this experiment is to evaluate the accuracy of estimators as data size increases. We tried two kinds of scaleup: *bounded domain scaleup* where we fixed the sample size and the number of distinct values but increased n , and *unbounded domain scaleup* where we fixed the sampling fraction and increased the number of distinct values with n . For bounded domain scaleup, we varied n over $\{100K, 200K, \dots, 1000K\}$. We generated data with $Z = 2$ which gives 49 distinct values for $n = 1000$. To generate the 100K table, we made 100 copies of each distinct value; to generate the 200K table, we made 200 copies of each distinct value, and so on. The number of rows sampled was fixed at 10K. For the unbounded domain scaleup, we used $Z = 2$ with duplication factor 100, fixing the sampling fraction at 1.6%. Figure 9 shows that for bounded domain scaleup, the error for all estimators except HYBVAR remains approximately constant as n is increased whereas the error of HYBVAR increases approximately linearly with n . The reason for this is that HYBVAR uses the modified Shlosser’s estimator in this case. The modified Shlosser’s estimator is unable to detect situations where data is duplicated, and therefore overestimates by a factor proportional to the number of copies of each distinct value. We also note that HYBGEE uses GEE for high skew data ($Z = 2$) and hence these estimators behave identically in both scaleup experiments. In the unbounded scaleup experiment (Figure 10), we see that once again the errors of all estimators except HYBVAR remains approximately constant as n is increased. The error for HYBVAR however, abruptly increases at $n=400K$ since it switches from using DUJ2A to using the modified Shlosser’s estimator (which has a significantly higher error) for $n \geq 400K$. The two scaleup experiments show that all estimators except HYBVAR scale well with increasing data size.

Comparison of estimators on real-world data. In our last experiment we compare the accuracy and variance of estimators on three real-life data sets. The Census database has 15 columns (Age, Marital-Status, etc.), the CoverType database has 11 columns (Elevation, Aspect, Slope, etc.)

and the table considered in the MSSales database has 20 columns (Product, Division, LicenseNumber, Revenue etc.). Figure 11 shows the average (over all columns) error of each estimator on the Census database as the sampling fraction is varied. We see that GEE, AE, and HYBGEE consistently outperform the other three estimators on this data set. Figure 13 similarly shows that these estimators yield more accurate estimates than HYBSKEW for the CoverType data as well. We observe that HYBGEE performs better than both GEE and HYBSKEW, and the performance of AE is comparable to that of HYBSKEW. On MSSales (Figure 15), all estimators perform reasonably well although both HYBSKEW and HYBGEE appear to give lowest errors. Figures 12 and 14 and 16 show the standard deviation of each estimator as a fraction of the actual number of distinct values in the data. We see that the variance of all estimators is small on real-world data sets; the two exceptions being HYBSKEW and DUJ2A on MSSales. Further, the variance decreases with increase in the sampling fraction.

7. CONCLUDING REMARKS

We presented a powerful negative result establishing the inherent difficulty of distinct-values estimation. At the same time, we devised a new estimator GEE which is provably optimal with respect to this negative result and performs reasonably well in practice, except in the case of low-skew data with a large number of distinct values. Our extensive empirical evaluation points to the fact that to ensure accuracy, the estimation procedure needs to take into account characteristics of the input distribution. We analytically derived the estimator AE, a new version of GEE that adapts to the input distribution so as to avoid the problems faced by GEE. Deriving a formal analytical error guarantee for AE is a key part of our future work.

8. REFERENCES

- [1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the ACM Symposium on the Theory of Computing*, 1996, pp. 20–29.
- [2] M.M. Astrahan, M. Schkolnick, and K.-Y. Whang. Approximating the number of unique values of an attribute without sorting. *Information Systems* 12(1987): 11–15.
- [3] J. Bunge and M. Fitzpatrick. Estimating the Number of Species: A Review. *Journal of the American Statistical Association* 88(1993): 364–373.
- [4] K.P. Burnham and W.S. Overton. Estimation of the size of a closed population when capture possibilities vary among animals. *Biometrika* 65(1978): 625–633.
- [5] K.P. Burnham and W.S. Overton. Robust estimation of population size when capture possibilities vary among animals. *Ecology* 60(1979): 927–936.
- [6] A. Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistical Theory and Applications* 11(1984): 265–270.
- [7] A. Chao and S. Lee. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* 87(1992): 210–217.
- [8] S. Chaudhuri, R. Motwani, and V. Narasayya. Using Random Sampling for Histogram Construction. In *Proceedings of the 1998 ACM SIGMOD Conference on Management of Data*, 1998, pp. 436–447.
- [9] S. Chaudhuri, R. Motwani, and V. Narasayya. Random Sampling over Joins: Efficient Implementations and Limitations. In *Proceedings of the 1999 ACM SIGMOD Conference on Management of Data*, 1999.
- [10] S. Chaudhuri and V. Narasayya. An Efficient, Cost-Driven Index Selection Tool for Microsoft SQL Server. In *Proceedings of the 23rd International Conference on Very Large Databases*, 1997.
- [11] S. Finkelstein, M. Schkolnick, and P. Tiberio. Physical Database Design for Relational Databases. *ACM TODS*, 13(1988): 91–128.
- [12] P. Flajolet and G.N. Martin. Probabilistic counting. In *Proceedings of the IEEE Symposium on the Foundations of Computer Science*, 1983, pp. 76–82.
- [13] Information and Computer Science, University of California, Irvine. <ftp://ftp.ics.uci.edu/pub/machine-learning-databases>
- [14] L.A. Goodman. On the estimation of the number of classes in a population. *Annals of Mathematical Statistics* 20(1949): 572–579.
- [15] P.J. Haas, J.F. Naughton, S. Seshadri, and L. Stokes. Sampling-based estimation of the number of distinct values of an attribute. In *Proceedings of the 21st International Conference on Very Large Databases*, pages 311–322, 1995.
- [16] P.J. Haas and L. Stokes. Estimating the number of classes in a finite population. In *Journal of the American Statistical Association*, 93(1998): 1475–1487.
- [17] J.F. Heltshe and N.E. Forrester. Estimating species richness using the jackknife procedure. *Biometrics* 39(1983): 1–11.
- [18] W. Hou, G. Ozsoyoglu, and B. Taneja. Statistical estimators for relational algebra expressions. In *Proceedings of the 7th ACM Symposium on Principles of Database Systems*, pages 276–287, 1988.
- [19] W. Hou, G. Ozsoyoglu, and B. Taneja. Processing aggregate relational queries with hard time constraints. In *Proceedings of the ACM-SIGMOD International Conference on Management of Data*, pages 68–77, 1989.
- [20] D.E. Knuth. **Sorting and Searching**, Volume 3 of *The Art of Computer Programming*. Addison-Wesley, Reading, MA, 1971.
- [21] R. Motwani and P. Raghavan. **Randomized Algorithms**. Cambridge University Press, 1995.
- [22] J.F. Naughton and S. Seshadri. On Estimating the Size of Projections. In *Proceedings of the Third International Conference on Database Theory*, pages 499–513, 1990.
- [23] F. Olken. *Random Sampling from Databases*. PhD Thesis, Computer Science, U.C. Berkeley, 1993.

- [24] F. Olken and D. Rotem. Random Sampling from Databases – A Survey. Manuscript, 1995.
- [25] G. Ozsoyoglu, K. Du, A. Tjahjana, W. Hou, and D.Y. Rowland. On estimating COUNT, SUM, and AVERAGE relational algebra queries. In *Proceedings of the Conference on Database and Expert Systems Applications*, pages 406–412, 1991.
- [26] V. Poosala, Y. Ioannidis, P. Haas, and E. Shekita. Improved Histograms for Selectivity Estimation of Range Predicates. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 294–305, 1996.
- [27] A. Shlosser. On estimation of the size of the dictionary of a long text on the basis of a sample. *Engineering Cybernetics* 19(1981): 97–102.
- [28] H.S. Sichel. Anatomy of the generalized inverse Gaussian-Poisson distribution with special application to bibliometric studies. *Information Processing and Management* 28(1992): 5–17.
- [29] E.P. Smith and G. Belle. Nonparametric estimation of species richness. *Biometrics* 40(1984):119–129.
- [30] K. Whang, B.T. Vander-Zanden, and H.M. Taylor. A linear-time probabilistic counting algorithm for database applications. *ACM Transactions on Database Systems* 15(1990): 208–229.
- [31] G.E. Zipf. **Human Behavior and the Principle of Least Effort**. Addison-Wesley Press, Inc, 1949.

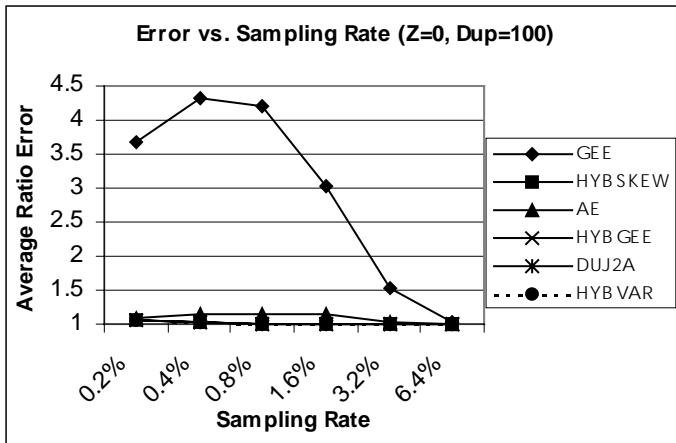


Figure 1. Variation of error with sampling rate (Z=0, Dup=100)

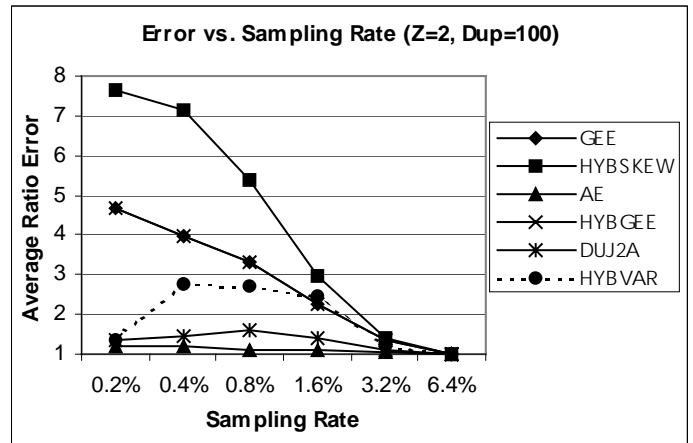


Figure 2. Variation of error with sampling rate (Z=2, Dup=100)

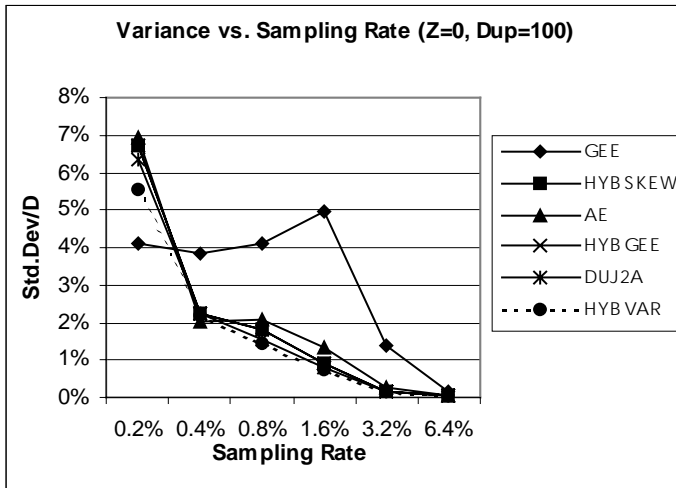


Figure 3. Variance of estimators vs. sampling rate (Z=0, Dup=100)

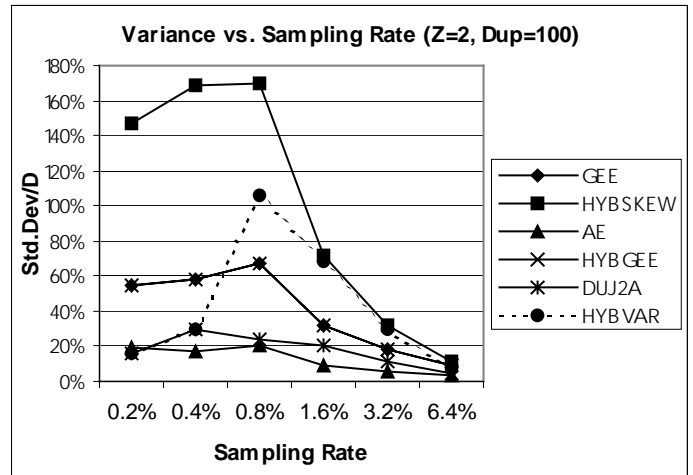


Figure 4. Variance of estimators vs. sampling rate (Z=2, Dup=100)

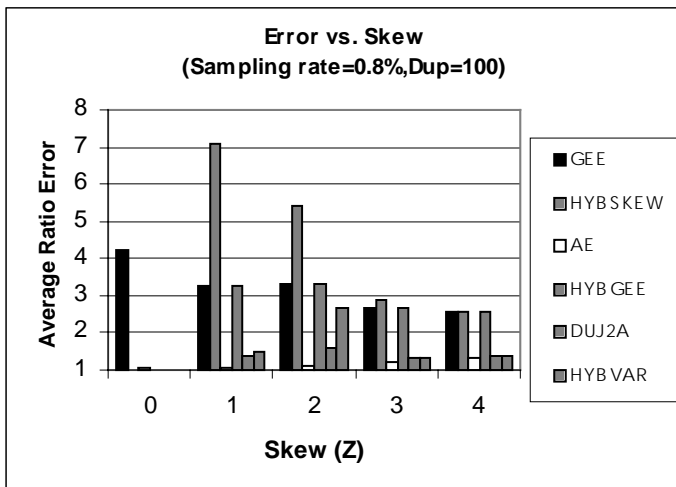


Figure 5. Variation of error with skew (Sampling Rate=0.8%, Dup=100)

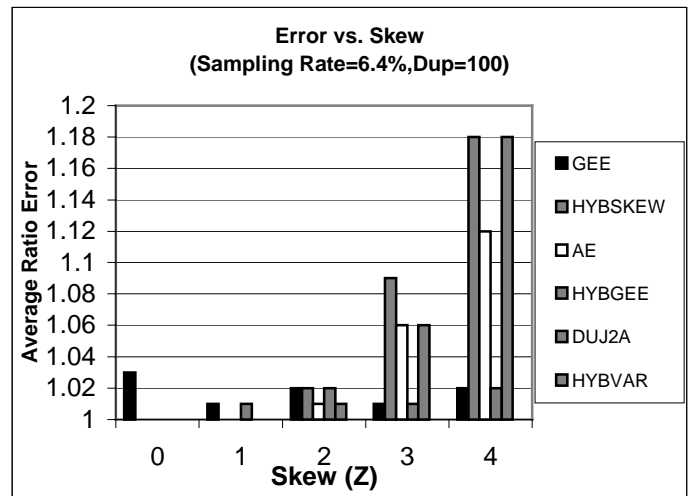


Figure 6. Variation of error with sampling rate (Z=0, Dup=100)

Sampling Rate	ACTUAL	LOWER	UPPPER
0.2%	10000	1814	817300
0.4%	10000	3345	671118
0.8%	10000	5511	452502
1.6%	10000	7999	207963
3.2%	10000	9611	47960
6.4%	10000	9987	11306

Table 1. Error Guarantee for GEE (Z=0, Dup=100, N = 1million)

Sampling Rate	ACTUAL	LOWER	UPPPER
0.2%	156	63	15712
0.4%	156	86	9044
0.8%	156	111	5067
1.6%	156	137	2083
3.2%	156	152	531
6.4%	156	156	169

Table 2. Error Guarantee for GEE (Z=2, Dup=100, N = 1million)

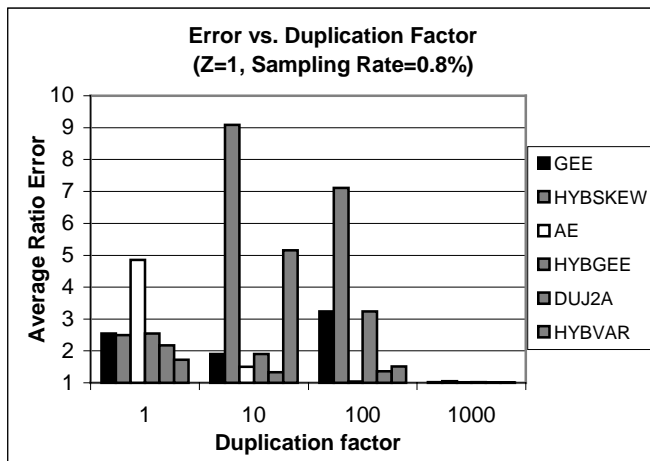


Figure 7. Variation of error with duplication factor (Z=1, Sampling rate=0.8%)

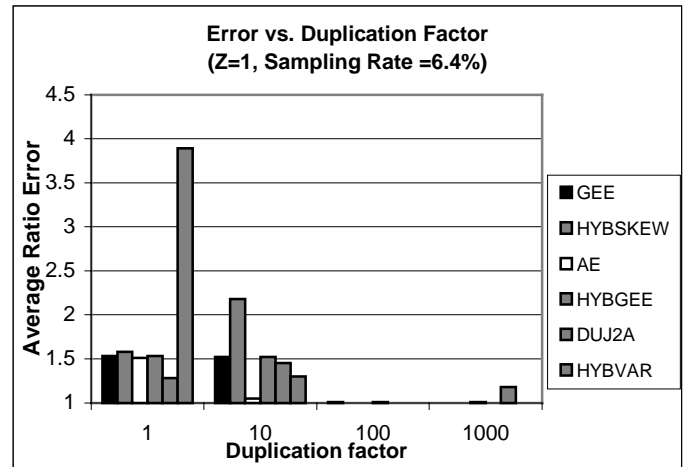


Figure 8. Variation of error with duplication factor (Z=1, Sampling rate=6.4%)

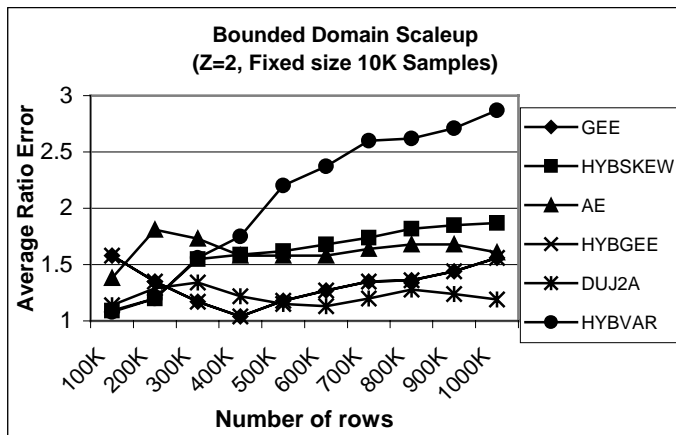


Figure 9. Scaleup when number of distinct values is kept constant.

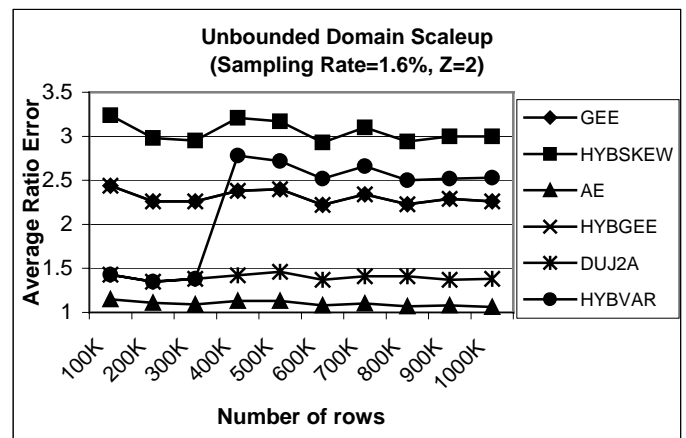


Figure 10. Scaleup when number of distinct values is increased with number of rows.

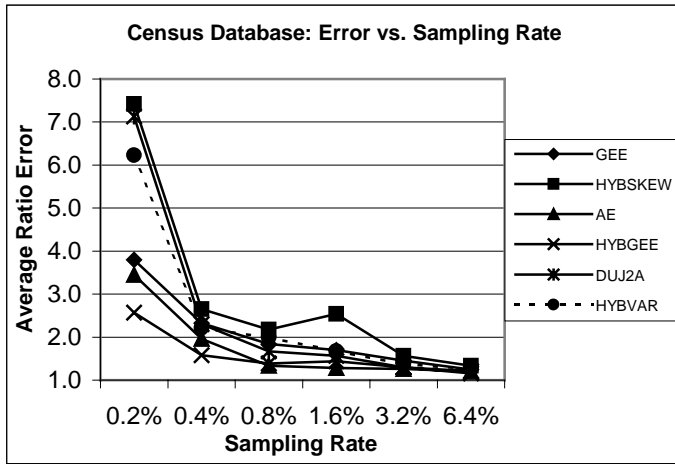


Figure 11. Average error of estimators over all columns of Census database.

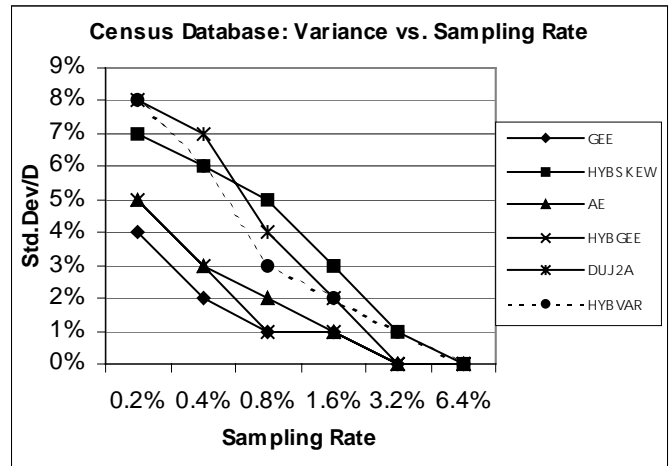


Figure 12. Variance of estimators over all columns of Census database.

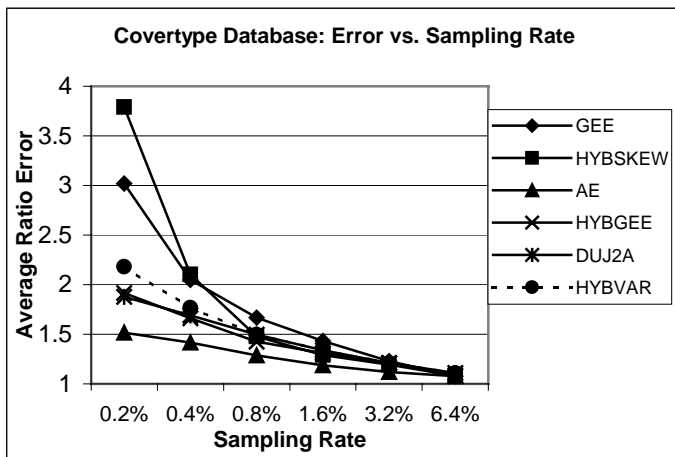


Figure 13. Average error of estimators over all columns of Covertypes database.

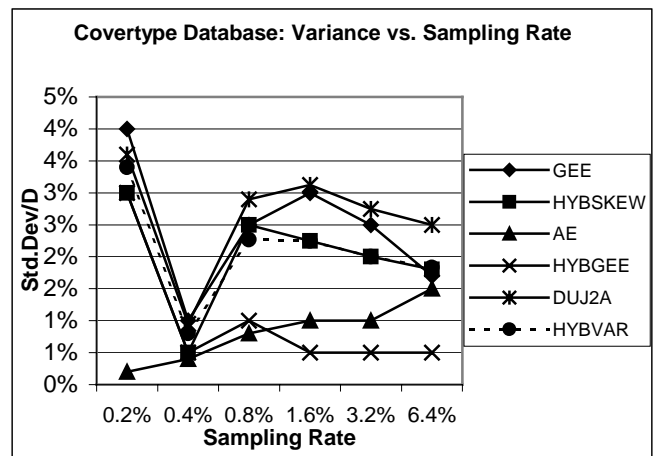


Figure 14. Variance of estimators over all columns of Covertypes database.

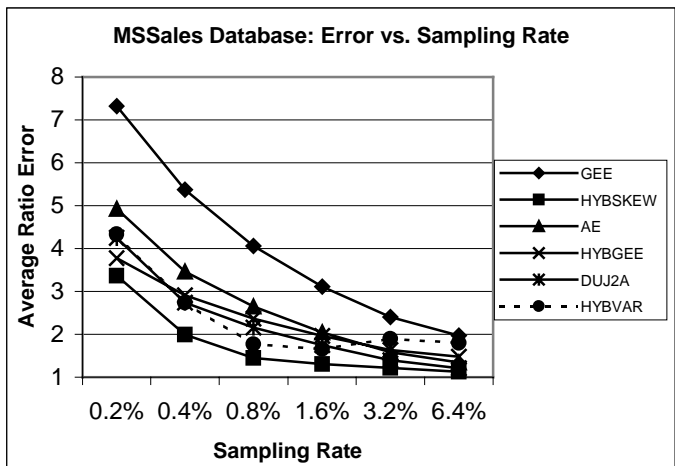


Figure 15. Average error of estimators over all columns of a table in MSSales database.

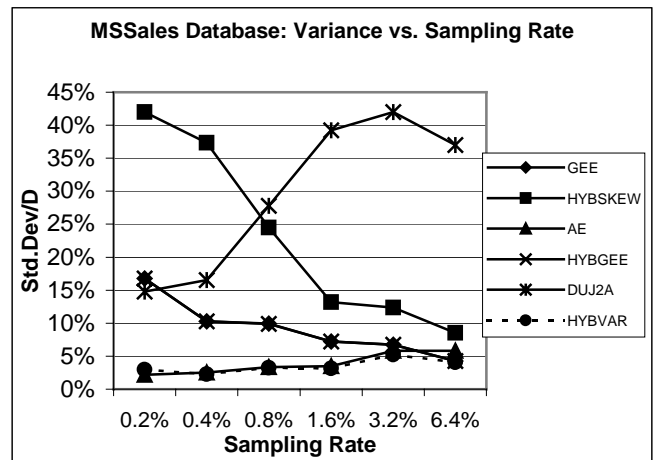


Figure 16. Variance of estimators over all columns of a table in MSSales database.