Sampling-Based Estimation of the Number of Distinct Values of an Attribute

Peter J. Haas* Jeffrey F. Naughton† S. Seshadri‡ Lynne Stokes§

Abstract

We provide several new sampling-based estimators of the number of distinct values of an attribute in a relation. We compare these new estimators to estimators from the database and statistical literature empirically, using a large number of attribute-value distributions drawn from a variety of real-world databases. This appears to be the first extensive comparison of distinct-value estimators in either the database or statistical literature, and is certainly the first to use highly-skewed data of the sort frequently encountered in database applications. Our experiments indicate that a new "hybrid" estimator yields the highest precision on average for a given sampling fraction. This estimator explicitly takes into account the degree of skew in the data and combines a new "smoothed jackknife" estimator with an estimator due to Shlosser. We investigate how the hybrid estimator behaves as we scale up the size of the database.

1 Introduction

Virtually all query optimization methods in relational and object-relational database systems require a means of assessing the number of distinct values of an attribute in a relation. Accurate assessment of the number of distinct values can be crucial for selecting a good query plan. For example, the relative error in the join-selectivity formulas used in the classic System R algorithms (Selinger, et al. [SAC$^+$79]) is directly related to the relative error in the constituent distinct-value estimates, and it is well known that poor join-selectivity estimates can increase the execution time of a join query by orders of magnitude. As another example, consider a query of the form "select * from R, S where $R.A = S.B$ and $f(S.C) > k$" that might be posed to an object-relational database system, and suppose that the predicate f is very expensive to compute; cf Hellerstein and Stonebraker [HS94]. Further suppose that 10% of the tuples in S join with tuples in R, and that 10% of the tuples in S satisfy $f(S.C) > k$. The query optimizer needs to decide whether to do the selection before or after the join. If attribute C has only a few distinct values in S, then by building a cache containing $\big(S.C, f(S.C)\big)$ pairs, the selection can be performed without invoking the function f more than a few times. This approach makes the selection relatively inexpensive, and it is better to do the selection before the join. If, on the other hand, attribute C has many distinct values, then f will be invoked many times when doing the selection, even if a cache is used. Since the selection operation is very expensive in this case, it is better to do the join $R.A = S.B$ first, and then apply the selection predicate to the 10% of the tuples in S that survive the join. It is not hard to see that a poor estimate of the number of distinct values can increase the execution time of the above query by orders of magnitude.

When there is an index on the attribute of interest, the number of distinct values can be computed exactly, in a straightfoward and efficient manner. We focus on the frequently-occurring case in which no such index is available. In the absence of an index, exact computation of the number of distinct values requires at least one full scan of the relation, followed by a sort- or hash-based computation involving each tuple in the relation. For most applications this approach is prohibitively expensive, both in time and space. "Probabilistic counting" methods (Astrahan, Schkolnick, and Whang [ASW87], Flajolet and Martin [FM85], and Whang, Vander-Zanden, and Taylor [WVT90]) estimate the number of distinct values without sorting and require only a small amount of memory. (These hash-based methods are actually deterministic, but

*IBM Research Division, Almaden Research Center, San Jose, CA 95120-6099; peterh@almaden.ibm.com

†Dept. of Computer Sciences, University of Wisconsin, Madison, WI 53706; naughton@cs.wisc.edu

‡Computer Science and Engineering Department, Indian Institute of Technology, Powai, Bombay 400 076; seshadri@cse.iitb.ernet.in

§Dept. of Management Science and Information Systems, University of Texas, Austin, TX 78712; lstokes@mail.utexas.edu

**Proceedings of the 21st VLDB Conference
Zurich, Swizerland, 1995**

use probabilistic arguments to motivate the form of the estimate.) Although much less expensive than exact computation, probabilistic counting methods still require that each tuple in the relation be scanned and processed. As databases continue to grow in size, such exhaustive processing becomes increasingly undesirable; this problem is especially acute in database mining applications. Several *ad hoc* estimation formulas that do not require a scan of the relation have been derived under various "uniformity" assumptions on the data (see, for example, Gelenbe and Gardy [GG82]). These formulas do not provide any indication of the precision of the estimate and can be off by orders of magnitude when the underlying assumptions are violated. In this paper we consider sampling-based methods for estimating the number of distinct values. Such methods process only a small fraction of the tuples in a relation, do not require *a priori* assumptions about the data, and permit assessment and control of estimation error.

Estimating the number of distinct attribute values using sampling corresponds to the statistical problem of estimating the number of classes in a population. A number of estimators have been proposed in the statistical literature (see Bunge and Fitzpatrick [BFi93] for a recent survey), but only three estimators, due to Goodman [Goo49], Chao [Cha84], and Burnham and Overton [BO78, BO79], respectively, have been considered in the database literature; see Hou, Ozsoyoglu, and Taneja [HOT88, HOT89] and Ozsoyoglu, et al. [ODT91]. As discussed in Section 2, each of these three estimators has serious theoretical or practical drawbacks. We therefore turn our attention to estimators from the statistical literature that have not been considered in the database setting and also to new estimators.

Identification of improved estimators is difficult. Ideally, a search for a good estimator would proceed by comparing analytic expressions for the accuracy of the estimators under consideration, and choosing the estimator that works best over a wide range of distributions; cf the comparison of join selectivity estimators in Haas, Naughton, Seshadri, and Swami [HN$^+$93]. Unfortunately, analysis of distinct-value estimators is highly non-trivial, and few analytic results are available. To make matters worse, there has been no extensive empirical testing or comparison of estimators in either the database or statistical literature. As discussed in Section 3, the testing that has been done in the statistical literature has generally involved small sets of data in which the frequencies of the different attribute values are fairly uniform; real data is seldom so well-behaved. In their survey, Bunge and Fitzpatrick provisionally recommend an estimator due to Chao and Lee [CL92], but this recommendation is not based on any systematic study. Moreover, because the Chao and Lee estimator is designed for sampling from an infinite population, that is, for sampling with replacement, it can take on large (and even infinite) values when there are a large number of distinct attribute values in the sample.

In this paper, we develop several new sampling-based estimators of the number of distinct values of an attribute in a relation and compare these new estimators empirically to estimators from the database and statistical literature. Our test data consists of approximately fifty attribute-value distributions drawn from three large real-world databases: an insurance company's billing records, a telecom company's fault-repair records, and the student and enrollment records from a large university. Perhaps the most interesting of the new estimators is a "smoothed jackknife" estimator, derived by adapting the conventional jackknife estimation approach to the distinct-value problem.

Our experimental results indicate that no one estimator is optimal for all attribute-value distributions. In particular, the relative performance of the estimators is sensitive to the degree of skew in the data. (Data with "high skew" has large variations in the frequencies of the attribute values; "uniform data" or data with "low skew" has nonexistent or small variations.) We therefore develop a new estimator that is a hybrid of the new smoothed jackknife estimator and an estimator due to Shlosser [Shl81]. The hybrid estimator explicitly takes into account the observed degree of skew in the data. For our real-world data, this new estimator yields a higher precision on average for a given sampling fraction than previously proposed estimators.

In Sections 2 and 3, we review the various estimators in the database and statistical literature, respectively. We develop several new estimators in Section 4. Our experimental results are in Section 5: after comparing the performance of the various estimators, we investigate how the new hybrid estimator performs as the size of the problem grows. In Section 6 we summarize our results and indicate directions for future work.

2 Estimators from the Database Literature

In this section we review the Goodman, Chao, and jackknife estimators that have been proposed in the database literature. As discussed below, all of these estimators have serious flaws.

Throughout, we consider a fixed relation R consisting of N tuples and a fixed attribute of this relation having D distinct values, numbered $1, 2, \ldots, D$. For $1 \leq j \leq D$, let N_j be the number of tuples in R with attribute value j, so that $N = \sum_{j=1}^{D} N_j$. Both the new and existing estimators described in this section are based on a sample of n tuples selected randomly and uniformly from R, without replacement; we call such a sample a ***simple random sample***. We focus on sampling without replacement because, as indicated in Section 5.2 below, such sampling minimizes estimation errors. (See Olken [Olk93], for a survey of algorithms that can be used to obtain a simple random sample from a relational database.) Denote by n_j the number of tuples in the sample with attribute value j for $1 \leq j \leq D$. Also denote by d the number of distinct attribute values that appear in the sample and, for $1 \leq i \leq n$, let f_i be the number of attribute values that appear exactly i times in the sample. Thus, $\sum_{i=1}^{n} f_i = d$ and $\sum_{i=1}^{n} i f_i = n$.

2.1 Goodman's Estimator

Goodman [Goo49] shows that

$$\widehat{D}_{\text{Good}} = d + \sum_{i=1}^{n} (-1)^{i+1} \frac{(N-n+i-1)!\,(n-i)!}{(N-n-1)!\,n!} f_i$$

is the unique unbiased estimator of D when $n > \max(N_1, N_2, \ldots, N_D)$. He also shows that there exists no unbiased estimator of D when $n \le \max(N_1, N_2, \ldots, N_D)$. Hou, Ozsoyoglu, and Taneja [HOT88] propose $\widehat{D}_{\text{Good}}$ for use in the database setting.

Although $\widehat{D}_{\text{Good}}$ is unbiased, Goodman [Goo49], Hou, Ozsoyoglu, and Taneja [HOT88] and Naughton and Seshadri [NS90] all observe that $\widehat{D}_{\text{Good}}$ can have extremely high variance and numerically unstable behavior at small sample sizes. Our own preliminary experiments confirmed this observation. We found $\widehat{D}_{\text{Good}}$ to be very unstable, with relative estimation errors in excess of 20,000% for some distributions and sample sizes (even when $\widehat{D}_{\text{Good}}$ is truncated so that it lies between d and N). Moreover, $\widehat{D}_{\text{Good}}$ was extremely expensive to compute numerically, requiring the use of multiple precision arithmetic to avoid overflows. These problems were particularly severe for large relations and small sample sizes. We therefore do not consider $\widehat{D}_{\text{Good}}$ further.

2.2 Chao Estimator

Ozsoyoglu, et al. [ODT91] propose the estimator

$$\widehat{D}_{\text{Chao}} = d + \frac{f_1^2}{2f_2},$$

due to Chao [Cha84], for application in the database setting. This estimator, however, estimates only a lower bound on D; cf Section 1.3.3 in [BFi93]. As a result, the Chao estimator usually underestimates the actual number of distinct values (unless $f_2 = 0$, in which case the estimator blows up). For these reasons, $\widehat{D}_{\text{Chao}}$ has been superseded by the estimator $\widehat{D}_{\text{CL}}$ discussed in Section 3.1 below, and we do not consider $\widehat{D}_{\text{Chao}}$ further.

2.3 Jackknife Estimators

Burnham and Overton [BO78, BO79], Heltshe and Forrester [HF83], and Smith and van Bell [SvB84] develop jackknife schemes for estimating the number of species in a population. Ozsoyoglu et al. [ODT91] propose the use of the procedures developed in [BO78, BO79] for estimating the number of distinct values of an attribute in a relation.

The jackknife estimators are defined as follows (see Efron and Tibshirani [ET93] for a general discussion of jackknife estimators). Denote by d_n the number of distinct values in the sample; in this section we write $d = d_n$ to emphasisize the dependence on the sample size n. Number the tuples in the sample from 1 to n and for $1 \le k \le n$ denote by $d_{n-1}(k)$ the number of distinct values in the sample after tuple k has been removed. Note that $d_{n-1}(k) = d_n - 1$ if the attribute value for tuple k is unique; otherwise, $d_{n-1}(k) = d_n$. Set $d_{(n-1)} = (1/n)\sum_{k=1}^{n} d_{n-1}(k)$. Then the conventional "first-order" jackknife estimator is defined by

$$\widehat{D}_{CJ} = d_n - (n-1)\big(d_{(n-1)} - d_n\big).$$

The rationale given in [BO78, BO79] for $\widehat{D}_{CJ}$ is as follows. Suppose that there exists a sequence of nonzero constants $\{\, a_k \colon k \ge 1 \,\}$ such that

$$E\,[d_n] = D + \sum_{k=1}^{\infty} \frac{a_k}{n^k}. \qquad (1)$$

Equation (1) implies that d_n, viewed as an estimator of D, has a bias of $O(n^{-1})$. It can be shown that, under the assumption in (1), the bias of the estimator $\widehat{D}_{CJ}$ is only $O(n^{-2})$, so that $\widehat{D}_{CJ}$ can be viewed as a "corrected" version of the crude estimator d_n.

A second-order jackknife estimator can be based on the n quantities $d_{n-1}(1), d_{n-1}(2), \ldots, d_{n-1}(n)$ together with $n(n-1)/2$ additional quantities of the form $d_{n-2}(i,j)$ $(i < j)$, where $d_{n-2}(i,j)$ is the number of distinct values in the sample after tuples i and j have been removed. Under the assumption in (1), it can be shown that the resulting estimator has a bias of order (n^{-3}). This procedure can be carried out to arbitrary order; the mth order estimator has bias $O(n^{-m+1})$ provided that (1) holds. As the order increases, however, the variance of the estimator increases. General formulas for the mth order estimator are given in [BO78, BO79], along with a procedure for choosing the order of the estimator so as to minimize the overall mean square error (defined as the variance plus the square of the bias).

The difficulty with the above approach is that, unlike the problem considered in [BO78, BO79, HF83, SvB84], $E\,[d_n]$ is *not* of the form (1) in our estimation problem; see (6) below. It can be shown, in fact, that in our setting the bias of $\widehat{D}_{CJ}$ decreases and then *increases* as the sample size increases from 1 to N. This behavior can be seen empirically in Figures 6.1 and 6.2 of [ODT91]. Our own preliminary experiments also indicated that estimators based on the formulas of Burnham and Overton do not work well in our setting, and we do not consider them further. In Section 4 we derive a new first-order jackknife estimator that takes into account the true bias structure of d.

3 Estimators from the Statistical Literature

In this section we review several estimators that have been proposed in the statistical literature but not considered in the database literature. None of the estimators in this section have ever been extensively tested or compared.

3.1 Chao and Lee Estimator

The *coverage* C of a random sample is the fraction of tuples in R having an attribute value that appears in the sample:

$$C = \sum_{\{j\colon\, n_j > 0\}} \frac{N_j}{N}.$$

When the attribute-value distribution is perfectly uniform, we have $C = d/D$. Therefore, given an estimator $\widehat{C}$ of the coverage, a natural estimator of the number of distinct values is $\widehat{D} = d/\widehat{C}$. When sampling is performed with replacement, an estimate of $\widehat{C}$ can be obtained for any attribute-value distribution by observing that

$$\begin{aligned} 1 - E[C] &= \sum_{j=1}^{D} \frac{N_j}{N} P\{\, n_j = 0 \,\} \\ &= \sum_{j=1}^{D} \frac{N_j}{N}\left(1 - \frac{N_j}{N}\right)^n \end{aligned}$$

and (using binomial probabilities)

$$E[f_1] = \sum_{j=1}^{D} P\{\, n_j = 1 \,\} = \sum_{j=1}^{D} n\frac{N_j}{N}\left(1 - \frac{N_j}{N}\right)^{n-1},$$

so that $E[C] \approx 1 - E[f_1]/n$. A natural estimator of C is therefore given by $\widehat{C} = 1 - f_1/n$. Chao and Lee [CL92] combine this coverage estimator with a correction term to handle skew in the data and obtain the estimator

$$\widehat{D}_{\text{CL}} = \frac{d}{\widehat{C}} + \frac{n(1-\widehat{C})}{\widehat{C}}\hat{\gamma}^2,$$

where $\hat{\gamma}^2$ is an estimator of

$$\gamma^2 = \frac{(1/D)\sum_{j=1}^{D}(N_j - \overline{N})^2}{\overline{N}^2}, \tag{2}$$

the squared coefficient of variation of the frequencies $N_1, N_2, \ldots, N_D$. (In the above formula, $\overline{N} = (1/D)\sum_{j=1}^{D} N_j = N/D$.) Note that $\gamma^2 = 0$ when all the attribute-value frequencies are equal (uniform data); the larger the value of γ^2, the greater the skew in the data. In their survey, Bunge and Fitzpatrick [BFi93] recommended $\widehat{D}_{\text{CL}}$ as their "provisional choice" among the available estimators of D.

Chao and Lee derive $\widehat{D}_{\text{CL}}$ under the assumption that samples are taken from an infinite population. Consequently, when sampling from a finite relation $\widehat{D}_{\text{CL}}$ can take on overly-large (and even infinite) values when there are many distinct attribute values in the sample. In [CL92], the performance of $\widehat{D}_{\text{CL}}$ was analyzed using simulations based on synthetic data. In all of the data sets, the skew parameter γ^2 was always less than 1 and the number of distinct values was always relatively small (< 200). In our data, we found many values of γ^2 larger than 10, with one value equal to 81.6. In Section 5.1 we examine the performance of $\widehat{D}_{\text{CL}}$ against both uniform and highly-skewed data when $\widehat{D}_{\text{CL}}$ is truncated at N, the largest possible number of distinct values in the relation.

3.2 Shlosser's Estimator

Shlosser [Shl81] derives the estimator

$$\widehat{D}_{\text{Shloss}} = d + \frac{f_1 \sum_{i=1}^{n}(1-q)^i f_i}{\sum_{i=1}^{n} iq(1-q)^{i-1} f_i}$$

under the assumption that each tuple is included in the sample with probability $q = n/N$, independently of all other tuples. This "Bernoulli sampling" scheme approximates simple random sampling when both n and N are large. Shlosser's (rather complicated) derivation rests on the assumption that

$$\frac{E[f_i]}{E[f_1]} \approx \frac{F_i}{F_1}, \tag{3}$$

where F_i is the number of attribute values that appear exactly i times in the entire relation. Note that when each attribute value appears approximately m times in the relation, where $m > 1$, then the relation in (3) does *not* hold. For this reason we would not expect $\widehat{D}_{\text{Shloss}}$ to perform well when the attribute-value distribution is close to uniform.

The estimator $\widehat{D}_{\text{Shloss}}$ performed well in Shlosser's simulations. He only tested his estimator, however, against two small, moderately skewed data sets consisting of 1,474 and 18,032 elements, respectively.

3.3 Sichel's Parametric Estimator

The idea behind a parametric estimator is to fit a probability distribution to the observed relative frequencies of the different attribute values. The number of distinct attribute values in the relation is then estimated as a function of the fitted values of the parameters of the distribution. According to Bunge and Fitzpatrick [BFi93], the most promising of the parametric estimators in the literature is due to Sichel [Si86a, Si86b, Si92]

Sichel's estimator is based on fitting a "zero-truncated generalized inverse Gaussian-Poisson" (GIGP) distribution to the frequency data. This distribution has three parameters, denoted b, c, and ν. In [Si92], Sichel shows that a wide variety of well-known distributions, including the Zipf distribution, can be closely approximated by the GIGP distribution. The specific estimator we consider is based upon a two-parameter version of the GIGP distribution obtained by fixing the parameter ν at the value $-1/2$; Sichel asserts that this approach suffices for most of the distributions that he has encountered. For such a two-parameter GIGP model, the number of distinct attribute values in the population can be expressed as $2/bc$, and the parameters b and c can be estimated as follows [Si86a]. Set $A = 2n/d - \ln(n/f_1)$ and $B = 2f_1/d + \ln(n/f_1)$, and let g be the solution of the equation

$$(1+g)\ln(g) - Ag + B = 0 \tag{4}$$

such that $f_1/n < g < 1$. Also set

$$\widehat{b} = \frac{g\ln(ng/f_1)}{1-g}.$$

and

$$\widehat{c} = \frac{1-g^2}{ng^2}.$$

Then the final estimate of the number of distinct attribute values is

$$\widehat{D}_{\text{Sichel}} = \frac{2}{\widehat{b}\,\widehat{c}}.$$

Development of practical estimation methods based on the full three-parameter GIGP model is an area of current research; see Burrell and Fenton [BFe93].

In preliminary experiments, we found $\widehat{D}_{\text{Sichel}}$ to be unstable for a number of the attribute-value distributions that we considered. The problem was that for these distributions the equation in (4) did not have a solution in the required range $(f_1/n, 1)$. As a result, the estimates took on values of 0 or ∞. (Even when $\widehat{D}_{\text{Sichel}}$ was truncated at d or N, respectively, the relative estimation errors still exceeded 2000%.) This phenomenon is due to a poor fit of the (two parameter) GIGP distribution to the data. It is possible that use of the more flexible three parameter GIGP would permit a better fit to the data, but practical methods for fitting the three parameter distribution are not yet available. Because of these problems, we do not consider $\widehat{D}_{\text{Sichel}}$ further.

3.4 Method-of-Moments Estimator

When samples are taken from an infinite population and the frequencies of the distinct attribute values are all equal ($N_1 = N_2 = \cdots = N_D$), it can be shown (see Appendix A) that $E[d] \approx D(1 - e^{-n/D})$. A simple estimator $\widehat{D}_{\text{MM0}}$ is then obtained by using the observed number d of distinct attribute values in the sample as an estimate of $E[d]$. That is, $\widehat{D}_{\text{MM0}}$ is defined as the solution D of the equation

$$d = D(1 - e^{-n/D}).$$

The above equation can be solved numerically using, for example, Newton-Raphson iteration. (The technique of replacing $E[d]$ by an estimate of $E[d]$ is called the *method of moments.*) The basic properties of the estimator $\widehat{D}_{\text{MM0}}$ have been extensively studied for the case of sampling from infinite populations with equal attribute-value frequencies; see Section 1.3.1 in [BFi93] for references.

$\widehat{D}_{\text{MM0}}$ is designed for sampling from infinite populations. When sampling from a finite relation, the estimator can take on overly-large (and even infinite) values if there are many distinct attribute values in the sample. $\widehat{D}_{\text{MM0}}$ also can be inaccurate when the data is heavily skewed. In Section 4 below we derive modifications of $\widehat{D}_{\text{MM0}}$ that attempt to address these difficulties and in Section 5.1 we compare the performance of the resulting estimators to $\widehat{D}_{\text{MM0}}$ when the value of $\widehat{D}_{\text{MM0}}$ is truncated at N.

3.5 Bootstrap Estimator

Smith and van Bell [SvB84] propose the bootstrap estimator for a species-estimation problem closely related to the estimation problem considered here. Although our sampling model is slightly different, the resulting estimator

$$\widehat{D}_{\text{Boot}} = d + \sum_{\{j:\, n_j > 0\}} \left(1 - \frac{n_j}{n}\right)^n.$$

is identical to the one in [SvB84]. (Recall that n_j denotes the number of tuples in the sample with attribute value j.) Observe that $\widehat{D}_{\text{Boot}} \leq 2d$, so that $\widehat{D}_{\text{Boot}}$ may perform poorly when D is large and n is small. See [ET93] for a general discussion of bootstrap estimators.

4 New Estimators

In this section, we derive several new estimators of the number of distinct values of an attribute in a relation. After first deriving a "Horvitz-Thompson"-type estimator, we then develop method-of-moments estimators that explicitly take into account both skewness in the data and the fact the we are sampling from a finite relation. Finally, we derive a new "smoothed jackknife" estimator.

4.1 Horvitz-Thompson Estimator

In this section we obtain an estimator of D by specializing an approach due to Horvitz and Thompson; see Sarndal, Swensson, and Wretman [SSW92] for a general discussion of Horvitz-Thompson estimators. Set $Y_j = 1$ if $n_j > 0$ and set $Y_j = 0$ otherwise. Observe that

$$E\left[\sum_{j=1}^{D} \frac{Y_j}{P\{\, n_j > 0 \,\}}\right] = \sum_{j=1}^{D} \frac{E[Y_j]}{P\{\, n_j > 0 \,\}} = \sum_{j=1}^{D} 1 = D.$$

Thus, if $P\{\, n_j > 0 \,\}$ is known for each j, then the estimator

$$\widehat{D} = \sum_{\{j:\, n_j > 0\}} \frac{1}{P\{\, n_j > 0 \,\}}$$

is an unbiased estimator of D. It can be shown (see Appendix A) that $P\{\, n_j > 0 \,\} = 1 - h_n(N_j)$, where

$$h_n(x) = h_n(x; N) = \frac{\Gamma(N - x + 1)\Gamma(N - n + 1)}{\Gamma(N - n - x + 1)\Gamma(N + 1)} \qquad (5)$$

for $x > 0$. Here Γ denotes the standard gamma function; see Section 6 of [AS72]. Of course, N_j, and hence $P\{\, n_j > 0 \,\}$, is unknown in practice. However, we can estimate $P\{\, n_j > 0 \,\}$ by $1 - h_n(\widehat{N}_j)$, where $\widehat{N}_j = (n_j/n)N$. The resulting estimator is

$$\widehat{D}_{\text{HT}} = \sum_{\{j:\, n_j > 0\}} \frac{1}{1 - h_n(\widehat{N}_j)}.$$

4.2 Method-of-Moments Estimators

The estimator $\widehat{D}_{\text{MM0}}$ was derived under the equal-frequency assumption $N_1 = N_2 = \cdots = N_D$ and the assumption that samples are taken from an infinite population. Under the equal-frequency assumption but with sampling from a finite relation, it can be shown (see Appendix A) that $E[d] = D\bigl(1 - h_n(N/D)\bigr)$, where $h_n(x)$ is given by (5). We can thus define a new method-of-moments estimator $\widehat{D}_{\text{MM1}}$ as the solution D of the equation

$$d = D\bigl(1 - h_n(N/D)\bigr).$$

It is reasonable to expect that this estimator would perform well for reasonably uniform attribute-value distributions.

When the frequencies of attribute values are unequal, we have (Appendix A)

$$E[d] = D - \sum_{j=1}^{D} h_n(N_j). \qquad (6)$$

To obtain an estimator that can handle skewed data, we approximate each term $h_n(N_j)$ in (6) by a second-order Taylor expansion about the point $\overline{N} = N/D$. After some straightforward computations, we obtain

$$\frac{E[d]}{D} \approx 1 - h_n(\overline{N}) + \tfrac{1}{2}\overline{N}^2\gamma^2 h_n(\overline{N})\big(g_n'(\overline{N}) - g_n^2(\overline{N})\big), \quad (7)$$

where γ^2 is defined as in (2) and

$$g_n(x) = \sum_{k=1}^{n} \frac{1}{N - x - n + k}. \quad (8)$$

It can be shown that

$$\gamma^2 = \frac{N-1}{\overline{N}n(n-1)} \sum_{i=1}^{n} i(i-1)E[f_i] + \frac{1}{\overline{N}} - 1,$$

so that a natural method-of-moments estimator $\hat{\gamma}^2(D)$ of γ^2 is given by

$$\hat{\gamma}^2(D) = \frac{(N-1)D}{Nn(n-1)} \sum_{i=1}^{n} i(i-1)f_i + \frac{D}{N} - 1. \quad (9)$$

An estimator of the number of distinct attribute values in the relation can be obtained by replacing $E[d]$ by d and γ^2 by $\hat{\gamma}^2(D)$ in (7) and numerically solving for D, but this approach is computationally expensive. Alternatively, the first-order estimate $\widehat{D}_{\mathrm{MM1}}$ can be used to estimate $\overline{N}$ and γ^2, and the resulting approximate version of (7) can be solved to yield an estimator $\widehat{D}_{\mathrm{MM2}}$ defined by

$$\widehat{D}_{\mathrm{MM2}} = d\Big(1 - h_n(\tilde{N}) + \tfrac{1}{2}\tilde{N}^2\hat{\gamma}^2(\widehat{D}_{\mathrm{MM1}})h_n(\tilde{N})\big(g_n'(\tilde{N}) - g_n^2(\tilde{N})\big)\Big)^{-1},$$

where $\tilde{N} = N/\widehat{D}_{\mathrm{MM1}}$. Preliminary numerical experiments indicated that when $\gamma^2 < 1$ and $n/N \geq 0.05$ the estimator $\widehat{D}_{\mathrm{MM2}}$ is essentially identical to the estimator obtained by numerical solution of (7). (As shown in Section 5, neither estimate performs satisfactorily when $\gamma^2 > 1$ or $n/N < 0.05$.) The estimator $\widehat{D}_{\mathrm{MM2}}$ can be viewed as a variant of $\widehat{D}_{\mathrm{MM1}}$ that has been "corrected" to account for the variability of the N_j's.

4.3 A Smoothed Jackknife Estimator

Recall the notation of Section 2.3. In the usual derivation of the first-order jackknife estimator, we seek a constant K such that

$$K(E[d_{n-1}] - E[d_n]) \approx \text{ bias of } d_n = E[d_n] - D.$$

Given K, we then estimate $E[d_{n-1}]$ by $d_{(n-1)}$ and $E[d_n]$ by d_n, and the final bias-corrected estimator is given by

$$\widehat{D} = d_n - K\big(d_{(n-1)} - d_n\big). \quad (10)$$

In the case of the conventional jackknife estimator (as in [BO78]), we have $K = (n-1)$. As discussed in Section 2.3, the key assumption in (1) that underlies the derivation of the conventional jackknife estimator is not satisfied in our setting. We show in Appendix B that the appropriate expression for K is

$$K \approx -\frac{N - \overline{N} - n + 1}{\overline{N}}\left(1 - \frac{\overline{N}\gamma^2 h_{n-1}'(\overline{N})}{h_{n-1}(\overline{N})}\right),$$

where $\overline{N} = N/D$ and h_n is defined as in (5). After substituting this expression for K into (10) and "smoothing" the resulting estimation equation (see Appendix B for details), we obtain the final estimator

$$\widehat{D}_{\mathrm{sjack}} = \left(1 - \frac{(N - \tilde{N} - n + 1)f_1}{nN}\right)^{-1} \Big(d_n + Nh_n(\tilde{N})g_{n-1}(\tilde{N})\hat{\gamma}^2(\widehat{D}_0)\Big), \quad (11)$$

where $\hat{\gamma}^2$ is given by (9), g_{n-1} is given by (8), and $\widehat{D}_0$ is defined by

$$\widehat{D}_0 = \big(d_n - (f_1/n)\big)\left(1 - \frac{(N - n + 1)f_1}{nN}\right)^{-1},$$

and $\tilde{N} = N/\widehat{D}_0$.

5 Experimental Results

In this section, we compare the performance of the most promising of the distinct-value estimators described in Sections 3 and 4 and develop a new hybrid estimator that performs better than any of the individual estimators. We then investigate how the hybrid estimator behaves as we scale up the size of the database. Our performance measure is the mean absolute deviation (MAD) expressed as a percentage of the true number of distinct values; that is, our performance measure for an estimator $\widehat{D}$ of D is

$$100\left(\frac{E[|\widehat{D} - D|]}{D}\right).$$

To motivate this performance measure, suppose that the a distinct-value estimator is used in conjunction with the classical System R formula as in [SAC$^+$79] to estimate the selectivity of a join. Then a MAD of x% in the distinct-value estimator induces an error of approximately ±x% in the selectivity estimate. (In our experiments, we also looked at the root-mean-square (RMS) error $100E^{1/2}[(\widehat{D} - D)^2]/D$; the RMS was consistently about 4% to 6% higher than the MAD, but the relative performance of the estimators based on the RMS error was the same as the relative performance based on MAD.)

In our experiments we always apply "sanity bounds" to each estimator. That is, we increase an estimator $\widehat{D}$ to d if $\widehat{D} < d$ and decrease $\widehat{D}$ to N if $\widehat{D} > N$.

5.1 Empirical Performance Comparison

Our comparison is based on 47 attribute-value distributions obtained from three large (> 1.5 GB) databases, one containing student and enrollment records, one containing fault-repair records from a telecom company, and

Dist.	tuples	d.v.'s	γ^2	Dist.	tuples	d.v.'s	γ^2	Dist.	tuples	d.v.'s	γ^2
1	624473	624473	0.00	17	1547606	3	0.38	33	173805	61	31.71
2	1288928	1288928	0.00	18	633756	202462	1.19	34	597382	17	14.27
3	15469	15469	0.00	19	597382	437654	1.53	35	1547606	21	6.30
4	113600	110074	0.04	20	1463974	624472	0.94	36	633756	221480	15.68
5	597382	591564	0.01	21	931174	110076	1.63	37	1547606	49	6.55
6	621498	591564	0.05	22	178525	52	9.07	38	633756	213	16.16
7	15469	131	3.76	23	1654700	624473	1.13	39	1463974	535328	7.60
8	1341544	1288927	0.05	24	624473	168	3.90	40	1463974	10	8.12
9	100655	29014	0.67	25	113600	6155	24.17	41	931174	73	12.96
10	147811	110076	0.47	26	173805	72	16.98	42	1547606	909	7.99
11	162467	83314	0.47	27	931174	29	3.22	43	931174	398	19.70
12	113600	3	0.70	28	178531	23	19.30	44	1341544	37	33.03
13	173805	109688	0.93	29	73561	287	55.77	45	624473	14047	81.63
14	73950	278	3.86	30	147811	62	34.68	46	1654700	235	30.85
15	1547606	51168	0.23	31	1547606	33	3.33	47	1463974	233	37.75
16	73950	8	6.81	32	1547606	194	3.35				

Table 1: Characteristics of 47 experimental attribute-value distributions.

one containing billing records from a large insurance company. Table 1 shows for each attribute-value distribution the total number of tuples, total number of distinct attribute values, and squared coefficient of variation of the attribute-value frequencies (that is, the parameter γ^2 given by (2)). The attributes corresponding to distributions 1–3 are primary keys, so that all attribute values are distinct and $\gamma^2 = 0$. For a given estimator and attribute-value distribution, we estimate the MAD by repeatedly drawing a sample from the distribution, evaluating the estimator, and then computing the absolute deviation. The final estimate is obtained by averaging over all of the experimental replications. We use 100 repetitions, which is sufficient to estimate the MAD with a standard error of $\leq 5\%$ in virtually all cases; typically, the standard error is much less.

Unlike the MAD of a join-selectivity estimator or an estimator of a population mean, the MAD of a distinct-value estimator is not independent of the population size; see, for example, p. 215 in Lewontin and Prout [LP56]. It follows that the MAD cannot be viewed as a simple function of the sample size. Initial experiments indicated that the MAD can more reliably be viewed as a function of the sampling fraction (see also Section 5.3), and so we vary the sampling fraction, rather than the sample size, in our experiments.

All of the estimators except $\widehat{D}_{\mathrm{HT}}$ and $\widehat{D}_{\mathrm{Boot}}$ were perfectly accurate for attribute-value distributions 1–3. The reason is that all of the estimators except $\widehat{D}_{\mathrm{HT}}$ and $\widehat{D}_{\mathrm{Boot}}$ assume that if all the attribute values in a sample are distinct (as they must be when sampling from distributions 1–3), then all the attribute values in the relation are distinct.

Tables 2 and 3 display the average and maximum MAD for the remaining eight estimators when applied to distributions with low skew and high skew, respectively. (We exclude the three attribute-value distributions in which all values are distinct.) As can be seen from these results, the relative performance of the estimators for distributions with low skew is quite different from the relative performance for distributions with high skew. In particular, estimators $\widehat{D}_{\mathrm{MM2}}$, $\widehat{D}_{\mathrm{CL}}$, and $\widehat{D}_{\mathrm{sjack}}$ perform well for distributions with low skew but perform poorly for distributions with high skew. To understand this effect, recall that these three estimators are derived essentially using Taylor-series expansions in γ^2 about the point $\gamma^2 = 0$. When the skew is high, γ^2 tends to be large and the underlying Taylor-series expansions are no longer valid; when the skew is low, the Taylor-series expansions are accurate. As discussed in Section 3, estimator $\widehat{D}_{\mathrm{Shloss}}$ has the opposite behavior: due to the assumption in (3) the estimator does not work well for distributions with low skew. Because the derivation of $\widehat{D}_{\mathrm{Shloss}}$ does not depend on Taylor-series expansions in γ^2, the estimator can achieve reasonable accuracy even when γ^2 is large.

The estimators $\widehat{D}_{\mathrm{Boot}}$ and $\widehat{D}_{\mathrm{HT}}$ do not perform particularly well. As discussed earlier, $\widehat{D}_{\mathrm{Boot}}$ is bounded above by $2d$, and thus yields poor estimates when D is large and n is small. The poor performance of $\widehat{D}_{\mathrm{HT}}$ may be due to the fact that the least frequent attribute values in the sample have the greatest effect on the value of the estimator, but for each infrequent value j it is difficult to accurately estimate the frequency N_j of the value in the relation.

It is interesting to note that $\widehat{D}_{\mathrm{MM0}}$ performs better than $\widehat{D}_{\mathrm{MM1}}$. "Correcting" $\widehat{D}_{\mathrm{MM0}}$ to account for sampling without replacement (as opposed to just truncating the value of $\widehat{D}_{\mathrm{MM0}}$ at N) appears to result in underestimation problems for this type of estimator. The reason for this is that the degradation in accuracy due to errors in estimating γ^2 outweighs the advantages of using γ^2.

As can be seen from Table 2, estimator $\widehat{D}_{\mathrm{sjack}}$ gives the lowest MAD for the distributions with low skew. The superior performance of $\widehat{D}_{\mathrm{sjack}}$ is possibly due to the stabilizing effect of smoothing the estimator. On the other hand, the results in Table 3 show that $\widehat{D}_{\mathrm{Shloss}}$ gives the lowest MAD for the distributions with high skew. These observations suggest that a hybrid estimator that explicitly takes data skew into account might perform better overall. We develop and test such an estimator in the next section.

samp.	Estimator							
frac.	$\widehat{D}_{\text{Shloss}}$	$\widehat{D}_{\text{MM0}}$	$\widehat{D}_{\text{MM1}}$	$\widehat{D}_{\text{MM2}}$	$\widehat{D}_{\text{HT}}$	$\widehat{D}_{\text{Boot}}$	$\widehat{D}_{\text{CL}}$	$\widehat{D}_{\text{sjack}}$
5%	64.58	19.87	37.82	39.42	60.64	62.39	21.18	20.88
	(252.85)	(46.40)	(209.55)	(246.92)	(92.14)	(93.10)	(75.43)	(41.28)
10%	33.95	17.61	32.68	35.44	51.71	53.59	21.97	16.52
	(107.29)	(36.60)	(178.20)	(245.29)	(84.52)	(86.21)	(123.04)	(32.77)
20%	17.61	14.10	20.46	22.80	38.81	40.77	30.80	11.31
	(58.67)	(36.50)	(64.07)	(106.72)	(70.04)	(72.45)	(196.13)	(33.87)

Table 2: Estimated mean absolute deviation (%) for 8 distinct-value estimators– low skew case. Average value and (maximum value) over 20 "low skew" attribute-value distributions for sampling fractions of 5%, 10%, and 20%.

samp.	Estimator							
frac.	$\widehat{D}_{\text{Shloss}}$	$\widehat{D}_{\text{MM0}}$	$\widehat{D}_{\text{MM1}}$	$\widehat{D}_{\text{MM2}}$	$\widehat{D}_{\text{HT}}$	$\widehat{D}_{\text{Boot}}$	$\widehat{D}_{\text{CL}}$	$\widehat{D}_{\text{sjack}}$
5%	32.45	34.82	36.37	58.45	35.26	36.59	262.28	45.99
	(132.06)	(70.27)	(70.27)	(515.05)	(80.88)	(83.09)	(4194.38)	(186.15)
10%	23.56	27.81	29.44	36.77	27.59	28.26	158.60	39.61
	(95.07)	(60.08)	(60.08)	(166.51)	(65.70)	(69.04)	(1235.51)	(186.15)
20%	13.60	19.86	21.51	22.69	18.36	18.48	156.08	33.68
	(49.23)	(45.81)	(45.81)	(59.37)	(45.81)	(47.04)	(1072.38)	(186.15)

Table 3: Estimated mean absolute deviation (%) for 8 distinct-value estimators– high skew case. Average value and (maximum value) over 24 "high skew" attribute-value distributions for sampling fractions of 5%, 10%, and 20%.

5.2 Performance of a Hybrid Estimator

To obtain an estimator that is accurate over a wide range of attribute-value distributions, it is natural to try a hybrid approach in which the data is tested to see whether there is a large amount of skew. If the data appears to be skewed, then $\widehat{D}_{\text{Shloss}}$ is used; otherwise, $\widehat{D}_{\text{sjack}}$ is used. One straightforward way to detect skew is to perform an approximate χ^2 test for uniformity. Specifically, we set $\bar{n} = n/d$ and compute the statistic

$$u = \sum_{\{j:n_j>0\}} \frac{(n_j - \bar{n})^2}{\bar{n}}.$$

For $k > 1$ and $0 < \alpha < 1$, let $x_{k-1,\alpha}$ be the unique real number such that if χ^2_{k-1} is a random variable having $k-1$ degrees of freedom then $P\left\{ \chi^2_{k-1} < x_{k-1,\alpha} \right\} = \alpha$. Then the estimator $\widehat{D}_{\text{hybrid}}$ (with parameter α) is defined by

$$\widehat{D}_{\text{hybrid}} = \begin{cases} \widehat{D}_{\text{sjack}} & \text{if } u \le x_{n-1,\alpha} \\ \widehat{D}_{\text{Shloss}} & \text{if } u > x_{n-1,\alpha}. \end{cases}$$

In our experiments we take $\alpha = 0.975$.

Table 4 shows the average and maximum MAD over all 47 attribute distributions for $\widehat{D}_{\text{hybrid}}$ and for the eight estimators considered in the previous section. As can be seen from the table, the estimator $\widehat{D}_{\text{hybrid}}$ is able to exploit the relative strengths of the estimators $\widehat{D}_{\text{sjack}}$ and $\widehat{D}_{\text{Shloss}}$ to achieve the lowest overall average MAD. For a sampling fraction of between 10% and 20%, $\widehat{D}_{\text{hybrid}}$ estimates the number of distinct values to within an average error of ±10% to ±16%. Astrahan, et al. [ASW87] found this degree of precision adequate in the setting of query optimization.

Though details are not given here, we also computed the average and maximum MAD of the various estimators over all 47 attribute distributions using sampling *with* replacement. The relative performance of the estimators remained essentially the same as indicated above, except that $\widehat{D}_{\text{MM0}}$ occasionally had a lower MAD than $\widehat{D}_{\text{hybrid}}$. The MAD for the best-performing estimator under sampling with replacement, however, was always higher than the MAD for $\widehat{D}_{\text{hybrid}}$ under sampling without replacement. Thus, our results indicate that, as might be expected, sampling without replacement minimizes estimation errors.

5.3 Scaleup Performance of the Hybrid Estimator

Up to this point in this paper we have not addressed an important but difficult question: when is sampling-based estimation of the number of distinct values an attractive alternative to exact computation of the number of distinct values? Unfortunately, the error behavior of $\widehat{D}_{\text{hybrid}}$ is sufficiently complex that it is difficult to make general statements about the cost of sampling to a specified accuracy. For this reason, our goal in this section is not to provide an exact answer to the question "when should one use sampling for distinct-value estimation." Rather, we seek to identify trends in the performance of $\widehat{D}_{\text{hybrid}}$ that indicate how it performs as the size of the problem grows. It turns out that these trends in performance depend upon how the problem is scaled.

One way to scale up the problem is to keep the number of distinct attribute values fixed and multiply the frequency of each distinct value by the scaleup factor (thereby leaving the relative frequencies unchanged). This sort of scaleup appears, among other places, in the enrollment table of our university database. Each record of this table represents a student taking a course, with attributes ***student id***, ***credits***, ***grade***, and so forth. Consider, for example, the ***credits*** attribute of this table. The values of ***credits*** vary

samp. frac.	Estimator								
	$\widehat{D}_{\mathrm{Shloss}}$	$\widehat{D}_{\mathrm{MM0}}$	$\widehat{D}_{\mathrm{MM1}}$	$\widehat{D}_{\mathrm{MM2}}$	$\widehat{D}_{\mathrm{HT}}$	$\widehat{D}_{\mathrm{Boot}}$	$\widehat{D}_{\mathrm{CL}}$	$\widehat{D}_{\mathrm{sjack}}$	$\widehat{D}_{\mathrm{hybrid}}$
5%	44.05	26.24	34.66	46.62	49.70	51.18	142.95	32.37	23.85
	(252.85)	(70.27)	(209.55)	(515.05)	(92.21)	(93.17)	(4194.38)	(186.15)	(135.22)
10%	26.48	21.69	28.94	33.86	41.50	42.75	90.34	27.26	15.65
	(107.29)	(60.08)	(178.20)	(245.29)	(84.66)	(86.33)	(1235.51)	(186.15)	(44.63)
20%	14.44	16.14	19.69	21.29	30.37	31.42	92.81	22.01	10.33
	(58.67)	(45.81)	(64.07)	(106.72)	(70.26)	(72.65)	(1072.38)	(186.15)	(42.22)

Table 4: Estimated mean absolute deviation (%) for 9 distinct-value estimators– combined results. Average value and (maximum value) over 47 attribute-value distributions for sampling fractions of 5%, 10%, and 20%.

Tuples	MAD (%)	Tuples	MAD (%)
100K	7.01	600K	8.58
200K	8.06	700K	8.81
300K	8.56	800K	8.38
400K	8.80	900K	8.83
500K	8.60	1000K	8.57

Table 5: Performance of $\widehat{D}_{\mathrm{hybrid}}$ for bounded-domain scaleup, 10K samples in each case.

from 0 to 9, with 3, 4, and 5 being very popular values. The relative frequencies of the specific *credits* values remain largely unchanged whether we look at a database of 10,000 or 1,000,000 enrollment records. We call this kind of scaleup ***bounded-domain scaleup***, since here the size of the domain of the attribute does not vary.

For bounded-domain scaleup, $\widehat{D}_{\mathrm{hybrid}}$ performs very well. Table 5 gives one example of a bounded-domain scaleup experiment. We generate the data sets for this experiment by adding tuples to the relation according to the distribution of values in a highly-skewed generalized Zipf distribution (Zipf(2.00) with 33 distinct values.) In more detail, we begin with a 1000 tuple relation drawn from this distribution. It turns out that in this relation the most frequent attribute value appears 609 times, the next most frequent value appears 153 times, and so forth. We scale this to a 100,000 tuple relation by making the most frequent value appear 60,900 times, the next most frequent value appear 15300 times, and so forth. For the 200,000 tuple relation, the most frequent value appears 121,800 times, the next most frequent value 30,600 times, etc. Table 5 shows that for a constant sample size the MAD remains approximately constant as the relation grows. That is, the sampling fraction required to achieve a given precision decreases as the relation grows.

Another way to scale up the problem is to add new distinct attribute values as the relation grows such that for each $1 \leq i \leq N$ the fraction of distinct values that appears exactly i times in the relation remains unchanged. We call this kind of scaleup ***unbounded-domain*** scaleup. Unbounded-domain scaleup also appears in the university database. For example, consider the ***student id*** attribute of the enrollment table. Here, if we consider a database with 10,000 or 1M enrollment records the number of distinct values of ***student id*** grows proportionally, while the number of occurrences of each value does not vary significantly.

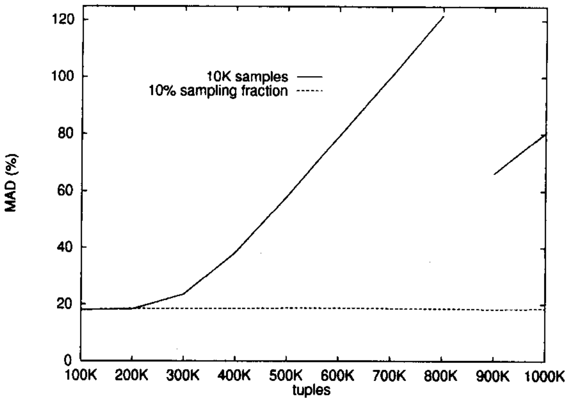

Figure 1: Performance of $\widehat{D}_{\mathrm{hybrid}}$ for unbounded-domain scaleup.

Figure 1 gives an example of an unbounded-domain scaleup experiment. Here we begin with the same generalized Zipf distribution used in Table 5, but scale up the relation by adding new distinct values. Specifically, we start with the same 1000 tuple relation as before. Recall that in this relation the most frequent attribute value appears 609 times, the second most frequent value appears 153 times, and so forth. To generate the 100,000 tuple relation, we add 99 new distinct values, each of which appears 609 times; another 99 new distinct values, each of which appears 153 times; and so forth.

Figure 1 shows that, unlike bounded-domain scaleup, for a constant sample size the MAD increases as the relation grows. The discontinuity between 800K and 900K tuples is a result of the hybrid estimator switching from the Shlosser estimator (at 800K tuples and below) to the new smoothed jackknife estimator (at 900K and 1M) due to the decreasing skew in the distribution as the relation scales.

Figure 1 also shows that if we keep the sample size a fixed percentage of the input, the MAD remains roughly constant as the relation grows; this suggests that while sampling does not get more attractive for larger inputs under unbounded-domain scaleup, it does not get less attractive either, so the effectiveness of sampling-based distinct-value estimation depends on the statistical properties of the data but not on the relation size.

6 Conclusions

Sampling-based estimation of the number of distinct values of an attribute in a relation is a very challenging problem. It is much more difficult, for example, than estimation of the selectivity of a join (cf [HN+93]). Perhaps for this reason, distinct-value estimation has received less attention than other database sampling problems, in spite of its importance in query optimization. Our results indicate, however, that in certain situations sampling-based methods can be used to estimate the number of distinct values at potentially a fraction of the cost of other methods.

In this paper we have provided a new distinct-value estimator that differs from previous estimators in that it explicitly adapts to differing levels of data skew. This new estimator, $\widehat{D}_{\text{hybrid}}$, is free from the flaws found in all of the previous distinct-value estimators in the database literature. Moreover, in empirical tests based on attribute-value distributions actually encountered in practice, $\widehat{D}_{\text{hybrid}}$ outperformed all previous known estimators. The performance of the new hybrid estimator $\widehat{D}_{\text{hybrid}}$ as the size of the problem grows depends on the precise nature of the scaleup. If the number of distinct attribute values remains fixed as the relation grows, then the cost of sampling (relative to processing all the tuples in the relation) decreases. If the number of distinct attribute values increases as the relation grows, then the relative cost of sampling remains roughly constant for a fixed sampling fraction.

There is ample scope for future work. A key research question is how to extend the applicability of the $\widehat{D}_{\text{hybrid}}$ estimator. Since $\widehat{D}_{\text{hybrid}}$ is based on the sampling of individual tuples rather than pages of tuples and can require a 10-20% sampling fraction, this estimator is best suited for situations in which reduction of CPU costs is a key concern. For example, the work described here was partially motivated by a situation in which a relation needed to be scanned for a variety of purposes, distinct-value estimation was desired, and the scan was CPU bound. $\widehat{D}_{\text{hybrid}}$ is also well-suited to distinct-value estimation for main-memory databases, where CPU costs dominate I/O costs. $\widehat{D}_{\text{hybrid}}$ can also be used effectively when I/O costs dominate and tuples are assigned to pages independently of the attribute value (that is, no "clustering" of attribute values on pages). In this case, the tuples required for the $\widehat{D}_{\text{hybrid}}$ estimator can be sampled a page at a time without compromising estimation accuracy. The estimator $\widehat{D}_{\text{hybrid}}$ needs to be extended, however, to permit sampling of tuples a page at a time when the attribute values are clustered on pages.

It is probable that more sophisticated hybrid estimators can be developed, resulting in further improvements in performance. It is also possible that if a practical parametric estimator based on the full three-parameter GIGP distribution could be constructed, it could fruitfully be used by itself or incorporated into a hybrid estimator. We are investigating techniques for estimating the variance of $\widehat{D}_{\text{hybrid}}$ and other estimators. This error information could potentially be used to develop fixed-precision estimation procedures for the number of distinct values. We are also considering various techniques for incorporating information from the system catalog and from previous monitoring of the database system into the estimator to improve estimation accuracy. Finally, we are starting to investigate how to incorporate sampling-based estimates into query optimizers.

Acknowledgements

Hongmin Lu made substantial contributions to the development of the smoothed jackknife estimator. This work was partially supported by NSF Grant IRI-9113736.

References

[AS72] Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions.* Ninth printing. Dover. New York.

[ASW87] Astrahan, M. M., Schkolnick, and Whang, K. (1987). Approximating the number of unique values of an attribute without sorting. *Inform. Systems* **12**, 11–15.

[BFi93] Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *J. Amer. Statist. Assoc.* **88**, 364–373.

[BO78] Burnham, K. P. and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**, 625–633.

[BO79] Burnham, K. P. and Overton, W. S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology* **60**, 927–936.

[BFe93] Burrell, Q. L. and Fenton, M. R. (1993). Yes, the GIGP really does work–and is workable! *J. Amer. Soc. Information Sci.* **44**, 61–69.

[Cha84] Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian J. Statist., Theory and Applications*, **11**, 265–270.

[CL92] Chao, A. and Lee, S. (1992). Estimating the number of classes via sample coverage. *J. Amer. Statist. Assoc.* **87**, 210–217.

[ET93] Efron, B. and Tibshirani, R. F. (1993). *An Introduction to the Bootstrap.* Chapman and Hall. New York.

[FM85] Flajolet, P. and Martin, G. N. (1985). Probabilistic counting algorithms for data base applications. *J. Computer Sys. Sci.* **31**, 182–209.

[GG82] Gelenbe, E, and Gardy, D. (1982). On the sizes of projections: I. *Information Processing Letters* **14**, 18–21.

[Goo49] Goodman, L. A. (1949). On the estimation of the number of classes in a population. *Ann. Math. Stat.* **20**, 572–579.

[HN+93] Haas, P. J., Naughton, J. F., Seshadri, S., and Swami, A. N. (1993). Selectivity and cost estimation for joins based on random sampling. Technical Report RJ 9577. IBM Almaden Research Center. San Jose, CA.

[HS94] Hellerstein, J. M. and Stonebraker, M. (1994). Predicate migration: optimizing queries with expensive predicates. *Proc. ACM-SIGMOD International Conference on Management of Data*, 267–276. Association for Computing Machinery. New York.

[HF83] Heltshe, J. F. and Forrester, N. E. (1983). Estimating species richness using the jackknife procedure. *Biometrics* **39**, 1–11.

[HOT88] Hou, W., Ozsoyoglu, G., and Taneja, B. (1988). Statistical estimators for relational algebra expressions. *Proc. 7th ACM Symposium on Principles of Database Systems*, 276–287. Association for Computing Machinery. New York.

[HOT89] Hou, W., Ozsoyoglu, G., and Taneja, B. (1989). Processing aggregate relational queries with hard time constraints. *Proc. ACM-SIGMOD International Conference on Management of Data*, 68–77. Association for Computing Machinery. New York.

[LP56] Lewontin, R.C. and Prout, T. (1956). Estimation of the number of different classes in a population. *Biometrics* **12**, 211–223.

[NS90] Naughton, J. F. and Seshadri, S. (1990). On estimating the size of projections. *Proc. Third Intl. Conf. Database Theory*, 499–513. Springer-Verlag. Berlin.

[Olk93] Olken, F. (1993). Random Sampling from Databases. Ph.D. Dissertation. Department of Computer Science. University of California at Berkeley. Berkeley, California.

[ODT91] Ozsoyoglu, G. Du, K., Tjahjana, A., Hou, W., and Rowland, D. Y. (1991). On estimating COUNT, SUM, and AVERAGE relational algebra queries. *Proc. Database and Expert Systems Applications (DEXA 91)*, 406–412. Springer-Verlag. Vienna.

[SSW92] Sarndal, Swensson, and Wretman (1992). *Model Assisted Survey Sampling*. Springer-Verlag. New York.

[SAC$^+$79] Selinger, P. G., Astrahan, D. D., Chamberlain, R. A., Lorie, R. A., and Price, T. G. (1979). Access path selection in a relational database management system. *Proc. ACM-SIGMOD International Conference on Management of Data*, 23–34. Association for Computing Machinery. New York.

[Shl81] Shlosser, A. (1981) On estimation of the size of the dictionary of a long text on the basis of a sample. *Engrg. Cybernetics* **19**, 97–102.

[Si86a] Sichel, H. S. (1986). Parameter estimation for a word frequency distribution based on occupancy theory. *Commun. Statist.- Theor. Meth.* **15**, 935–949.

[Si86b] Sichel, H. S. (1986). Word frequency distributions and type-token characteristics. *Math. Scientist* **11**, 45–72.

[Si92] Sichel, H. S. (1992). Anatomy of the generalized inverse Gaussian-Poisson distribution with special applications to bibliometric studies. *Information Processing and Management* **28**, 5–17.

[SvB84] Smith, E. P. and van Belle, G. (1984). Nonparametric estimation of species richness. *Biometrics* **40**, 119–129.

[WVT90] Whang, K., Vander-Zanden, B. T., and Taylor, H. M. (1990). A linear-time probabilistic counting algorithm for database applications.

A Expected Number of Distinct Values in a Sample

Consider a simple random sample of size n drawn from a relation with N tuples and suppose that the attribute of interest has D distinct values in the relation. The probability that the attribute value j does not appear in the sample (that is, the probability that $n_j = 0$) is equal to the hypergeometric probability

$$\begin{aligned} h_n(N_j) &= \binom{N-N_j}{n} \Big/ \binom{N}{n} \\ &= \frac{\Gamma(N-N_j+1)\Gamma(N-n+1)}{\Gamma(N-n-N_j+1)\Gamma(N+1)}, \end{aligned}$$

where N_j is the frequency of attribute value j in the relation. We have used the fact that $\Gamma(x+1) = x!$ whenever x is a nonnegative integer. Let $Y_j = 1$ if $n_j > 0$ and $Y_j = 0$ otherwise. Observe that

$$E[Y_j] = P\{Y_j = 1\} = P\{n_j > 0\} = 1 - h_n(N_j).$$

Letting d denote the number of distinct attribute values in the sample, we find that

$$E[d] = E\left[\sum_{j=1}^{D} Y_j\right] = \sum_{j=1}^{D} E[Y_j] = D - \sum_{j=1}^{D} h_n(N_j).$$

In particular, if $N_1 = N_2 = \cdots = N_D = N/D$, we have $E[d] = D\big(1 - h_n(N/D)\big)$. If, in addition, N is very large relative to n, so that we are effectively sampling from an infinite population, we have $h_n(x) \approx \left(1 - \frac{x}{N}\right)^n$, so that

$$\frac{E[d]}{D} \approx 1 - \exp\big(n \ln(1 - D^{-1})\big) \approx 1 - e^{-n/D},$$

where we have used the additional approximation $\ln(1 - D^{-1}) \approx -1/D$.

B Derivation of the Smoothed Jackknife Estimator

As discussed in Section 4.3, we seek a constant K such that

$$K \approx \frac{E[d_n] - D}{E[d_{n-1}] - E[d_n]}.$$

The jackknife estimator is then given by $\widehat{D} = d - K\big(d_{(n-1)} - d_n\big)$. Observe that $d_{(n-1)} = d_n - f_1/n$, so that $\widehat{D}$ can be written as

$$\widehat{D} = d_n + K\frac{f_1}{n}. \tag{12}$$

Using (6), we have

$$K = \frac{\sum_{j=1}^{D} h_n(N_j)}{\sum_{j=1}^{D} \left(\frac{N_j}{N-n+1}\right) h_{n-1}(N_j)}. \tag{13}$$

Set $\overline{N} = N/D$. Writing $h_n(N_j) \approx h_n(\overline{N}) + (N_j - \overline{N})h'_n(\overline{N})$ and

$$\begin{aligned} \left(\frac{N_j}{N-n+1}\right) h_{n-1}(N_j) &\approx \left(\frac{\overline{N}}{N-n+1}\right) h_{n-1}(\overline{N}) \\ &+ (N_j - \overline{N}) \left(\frac{N_j h'_{n-1}(\overline{N})}{N-n+1} + \frac{h_{n-1}(\overline{N})}{N-n+1}\right) \end{aligned}$$

for $1 \leq j \leq D$, substituting into (13), and using the approximation

$$\frac{a}{b+x} \approx \frac{a}{b}\left(1 - \frac{x}{b}\right)$$

for small x, we obtain

$$K \approx -\frac{N - \overline{N} - n + 1}{\overline{N}}\left(1 - \frac{\overline{N}\gamma^2 h'_{n-1}(\overline{N})}{h_{n-1}(\overline{N})}\right). \tag{14}$$

As before, γ^2 is the squared coefficient of deviation of the numbers $N_1, N_2, \ldots, N_D$; see (2). Substituting (14) into (12) and using the easily-established fact that $h'_k(x) = -h_k(x)g_k(x)$ for $k \geq 1$, we obtain

$$\begin{aligned} \widehat{D}&\left(1 - \frac{(N - \overline{N} - n + 1)f_1}{nN}\right) \\ &= d_n + \frac{(N - \overline{N} - n + 1)f_1 g_{n-1}(\overline{N})\gamma^2}{n}. \end{aligned} \tag{15}$$

We then "smooth" the jackknife by replacing the right side of the estimation equation (15) by its expected value. It can be shown that

$$\begin{aligned} \frac{E[f_1]}{n} &= \sum_{j=1}^{D} \frac{N_j}{N - N_j + n - 1} h_n(N_j) \\ &\approx \frac{\overline{N}}{N - \overline{N} + n - 1} h_n(\overline{N}). \end{aligned} \tag{16}$$

Replacing f_1/n by $E[f_1]/n$ on the right side of (15) and using (16) yields

$$\begin{aligned} \widehat{D}&\left(1 - \frac{(N - \overline{N} - n + 1)f_1}{nN}\right) \\ &= d_n + N h_n(\overline{N}) g_{n-1}(\overline{N})\gamma^2. \end{aligned} \tag{17}$$

A distinct-value estimator can be obtained by replacing $\overline{N}$ by $N/\widehat{D}$ and γ^2 by $\hat{\gamma}^2(\widehat{D})$ in (17) and solving (17) iteratively for $\widehat{D}$. As in the case of the method-of-moments estimator, however, we can obtain an estimate that is almost as accurate and much cheaper to compute by starting with a crude estimate of D and then correcting this estimate using (17). To do this, we replace each N_j in (13) with $\overline{N}$ and substitute the resulting expression for K into (12) to obtain the relation

$$\widehat{D} = d_n + \frac{(N - \overline{N} - n + 1)f_1}{nN}. \tag{18}$$

We then approximate $\overline{N}$ by $N/\widehat{D}$ in (18) and solve for $\widehat{D}$. The resulting solution, denoted by $\widehat{D}_0$, is given by

$$\widehat{D}_0 = \left(d_n - (f_1/n)\right)\left(1 - \frac{(N - n + 1)f_1}{nN}\right)^{-1}$$

and serves as our initial crude estimate. To obtain the final estimator in (11), we approximate $\overline{N}$ by $\tilde{N} = N/\widehat{D}_0$ and γ^2 by $\hat{\gamma}^2(\widehat{D}_0)$ in (17) and solve.

Modification of Chao and Lee's derivation of $\widehat{D}_{\mathrm{CL}}$ to account for random sampling from a finite relation yields an estimator essentially identical to $\widehat{D}_{\mathrm{sjack}}$. Thus, the new estimator $\widehat{D}_{\mathrm{sjack}}$ can be viewed as a modification of $\widehat{D}_{\mathrm{CL}}$ for sampling from a finite relation. Conversely, our derivation shows that $\widehat{D}_{\mathrm{Chao}}$ can be viewed as essentially a jackknife estimator. Unlike $\widehat{D}_{\mathrm{CL}}$, the estimator $\widehat{D}_{\mathrm{sjack}}$ does not equal ∞ when all attribute values in the sample are distinct.