New Sampling-Based Estimators for OLAP Queries

Ruoming Jin Leo Glimcher Chris Jermaine Gagan Agrawal
Department of Computer Science and Engineering
Ohio State University, Columbus OH 43210
{jinr,glimcher,agrawal}@cse.ohio-state.edu
Department of Computer and Information Science and Engineering
University of Florida, Gainsville, FL
cjermain@cise.ufl.edu

Abstract

One important way in which sampling for approximate query processing in a database environment differs from traditional applications of sampling is that in a database, it is feasible to collect accurate summary statistics from the data in addition to the sample. This paper describes a set of sampling-based estimators for approximate query processing that make use of simple summary statistics to to greatly increase the accuracy of sampling-based estimators. Our estimators are able to give tight probabilistic guarantees on estimation accuracy. They are suitable for low or high dimensional data, and work with categorical or numerical attributes. Furthermore, the information used by our estimators can easily be gathered in a single pass, making them suitable for use in a streaming environment.

1 Introduction

Interactive-speed evaluation of ad-hoc, OLAP-style queries is quite often an impossible task over the largest databases. Even using modern query processing techniques in conjunction with the latest hardware, OLAP-style queries can still require hours or days to complete execution. A quick perusal of the latest TPC-H benchmark supports this assertion [28].

One obvious way to address this problem is to rely on *approximation*. If an approximate answer can be given at subsecond, interactive speeds, it may encourage the use of large databases for ah-hoc data exploration and analysis. One of the most often-cited possibilities for supporting approximation in a database environment is to rely on *random sampling* [13, 12, 1, 2, 6, 7, 10, 14, 19, 22, 21, 18]. Unlike other estimation methods such as wavelets [29, 8, 9] and histograms [11], sampling has the advantage that most estimates over samples are unaffected by data dimensionality, and sampling is equally applicable to categorical and numerical data. Another primary advantage of random sampling is that sampling techniques are well-understood and sampling as a field of scientific study is very mature. Fundamental results from statistics can be used as a guide when applying sampling to data management tasks [27, 20, 26, 25].

However, there is one important way in which sampling in a database environment differs from sampling as it has been studied in statistics and related fields. Specifically, most work from statistics makes the assumption that it is impossible to gather accurate summary statistics from the population that is to be sampled from. After all, in traditional applications like sociology and biology, sampling is used when it is impossible to directly study the entire population.

In a database, this assumption is far too restrictive. There is no reason that accurate and extensive summary statistics over the underlying data set cannot be maintained. In databases, unlike in traditional statistical inference, the entire data set is available and can be pre-processed or processed in an online fashion as it is loaded. The reason that approximation may be required in a database environment is that there are too many possible queries to pre-compute the answer to every one, and it may be too expensive to evaluate every single query to get an exact answer. This does not imply that it is impractical to maintain a very comprehensive set of summary statistics over the data.

A recent paper [14] proposed a new technique called *Approximate Pre-Aggregation* (APA) that makes use of such summary statistics to greatly increase the accuracy of sampling-based-estimation in a database environment. APA is useful for providing extremely accurate answers to aggregate queries (`SUM`, `COUNT`, `AVG`) over high-dimensional database tables, with a complex relational selection predicate attached to the aggregate query. For example, imagine we have a database table `COMPLAINTS (PROF, SEMESTER, NUM_COMPLAINTS)` and the SQL query:

```
SELECT SUM (NUM_COMPLAINTS)
FROM COMPLAINTS
WHERE PROF = 'Smith'
  AND SEMESTER = 'Fa03';
```

This query asks: "How many complaints did Prof. Smith receive in the Fall 2003 semester?" APA was shown experimentally to produce estimates for the answer to this type of query that have only a small fraction of the error associated with other techniques such as stratified random sampling and wavelets, especially over categorical data.

However, there is one glaring weakness with APA: it is difficult or impossible to formally reason about the accuracy of the approach. In particular, while APA was shown experimentally to produce excellent results, there is no obvious way to associate statistically meaningful confidence bounds with the estimates produced by APA. A confidence bound is an assertion of the form, "With a probability of .95, Prof. Smith received 27 to 29 complaints in the Fall of 2003." The lack of confidence bounds poses a severe limitation on the applicability of APA to real-life estimation problems.

This Paper's Contributions: In this paper, we propose a new method for combining summary statistics with sampling to improve estimation accuracy, called *APA+*. APA+ is in many ways simpler than APA, and yet it has the overwhelming advantage of giving statistically meaningful confidence bounds on its estimates. As we will show experimentally in this paper, APA+ produces confidence bounds that are tighter and more accurate than those that are produced using more traditional, sampling-based estimators. Furthermore, computing an APA+ estimate is computationally very efficient, which renders APA+ an excellent candidate for use in applications like online aggregation [13], where an estimate must be recomputed every second or so in order to inform a user as to the current accuracy of the estimate.

Given the accuracy and wide applicability of the techniques presented in the paper, we assert that APA+ should be used in place of traditional, sampling-based estimation in any environment where it is possible to augment a sample with simple summary statistics over a database.

Paper Organization: Section 2 of the paper gives an overview of APA and our new method, APA+. The simplest version of APA+ is formally described in Section 3. Section 4 generalizes APA+ to work with more complicated and comprehensive summary statistics, and Section 5 discusses the issue of correlation in the individual estimates that make up the final APA+ estimator. Section 6 presents our experimental evaluation, and Section 7 gives an overview of related work. Finally, the paper is concluded in Section 8.

2 Overview of APA and APA+

In this Section, we will describe both APA and APA+ at a high level in the context of the example database table depicted in Table 1 [14].

This table models the situation where a number of students take a course each semester, and the number of student grading complaints is recorded.

2.1 Approximate Pre-Aggregation (APA)

In our example, we are interested in answering a query of the form:

S	Prof.	Sem.	Cmpl.	S	Prof.	Sem.	Cmpl.
	Adams	Fa 02	3	√	Smith	Su 01	7
	Jones	Fa 02	2	√	Smith	Sp 01	8
	Adams	Sp 02	9	√	Adams	Fa 00	4
√	Jones	Sp 02	2		Smith	Su 01	7
√	Smith	Sp 02	21		Smith	Fa 00	33
	Smith	Fa 01	36	√	Adams	Su 00	3
√	Jones	Su 01	1	√	Jones	Su 00	0
	Adams	Su 01	2		Jones	Sp 99	1

Table 1. Number of complains over three years

```
SELECT SUM (NUM_COMPLAINTS)
FROM COMPLAINTS
WHERE PROF = 'Smith'
```

Imagine that we decided to answer this query using a 50% sample of the database, where the sampled tuples are indicated by a check mark in Table 1. Estimating the answer to the query using the sample is very straightforward. The number of complaints in the sample is 36, and since the sample constitutes 50% of the database tuples, we can estimate that Professor Smith received 72 complaints.

However, as we see from Table 1, the actual number of students who came to see Professor Smith was 121 (yielding 40.5% relative error). The problem is the variance in the number of students who complained to Professor Smith each semester. This ranges from a low of 7 to a high of 36, and it happens that our sample missed the two semesters when the greatest number of students complained to Professor Smith.

APA makes use of additional summary statistics to reduce sampling's vulnerability to variance. The APA process begins by using the relational selection predicate of the query to divide the data space into 2^n quadrants, where n is the number of clauses in the predicate. APA then associates a probability density function with each quadrants, and uses the additional summary information that is available to create a set of constraints on the means of these distributions. A Maximum Likelihood Estimation (MLE) is then used to adjust the means in order to correct violations of the constraints. Depending on the dimensionality of the summary information used, different variations of APA are possible. For example, APA0 makes use of *zero-dimensional* facts about the data, or assertions of the form `(SUM (COMPLAINTS)) = 148`. APA1 makes use of *one-dimensional* facts about the data, or assertions of the form `(SUM (COMPLAINTS)) WHERE (SEMESTER = 'Fa02') = 5`, which have one clause in their relational selection predicate.

2.2 APA+

The fundamental weakness of APA is a lack of a formal analysis of the accuracy of the approach, which in turn precludes analytically-derived confidence bounds on APA estimates. Such an analysis is difficult for several reasons:

1. First, it can be difficult in general to reason about the accuracy of maximum likelihood estimators. While MLEs are usually Minimum Variance Unbiased Estimators (MVUEs) and are typically normally distributed

[16], this only provides a lower bound on the variance of the MLE using the Carmer-Rao inequality; it does not say whether or not the MLE actually achieves this lower bound.

2. Second, things are even more difficult in the case of APA, which uses a constrained MLE where the space of possible solutions is restricted by the available summary statistics. It is far from clear how to formally reason about the effect of this on the accuracy of the approach.

3. Finally, the various estimators that are combined in APA are correlated, and this correlation is ignored by APA. If one sample falls in the quadrant associated with Professor Smith, than it cannot fall in the quadrant holding tuples not associated with Professor Smith. Considering this correlation is absolutely necessary in order to develop confidence bounds for APA, and yet it makes the analysis even more difficult.

In this paper, we provide a simple alternative to APA, called *APA+*. APA+ is similar to APA, except that it does not rely on MLE and is far more amenable to a rigorous statistical analysis of its accuracy.

At the highest level, just like APA, APA+ also makes use of additional summary statistics, but in a fundamentally different way. APA+ relies on the idea of a *negative estimator* to make use of summary statistics. In APA+, the original, straightforward estimate of 72 is called a *positive estimator*, in that it is derived by directly considering the tuples that match our relational selection predicate. However, it is also possible to produce a negative estimator by first estimating the number of complaints not attributed to professor Smith (which is $2 \times (20 + 1 + 0 + 3 + 4) = 20$), and subtracting this quantity from 148, which is the total number of complaints. Thus, 148 - 20 = 128 is our *negative estimator* for the number of complaints attributed to Professor Smith.

If we denote the positive and negative estimator by $\hat{t}_p$ and $\hat{t}'_p$ respectively, then

$$\hat{t} = \alpha \hat{t}_p + (1 - \alpha)\hat{t}'_p$$

is itself an unbiased estimator of the number of complaints for Professor Smith for any value of α where $0 \leq \alpha \leq 1$. By choosing α so as to minimize the variance of $\hat{t}$, it is possible to develop an extremely accurate estimator.

This alternative method for incorporating summary statistics into our estimate has several obvious advantages:

1. Intuitively, the accuracy of APA+ should be very good, just like for APA. The reason is that the ensemble approach leverages more information into the estimation. By using the information from the negative domain, we can reduce the variance of our estimation. Considering that our ensemble estimator is unbiased for the answer of the query, the error associated with it must be small, leading to very accurate estimates.

2. Unlike APA, APA+ relies on simple arithmetic in order to make use of summary information, and is thus very amenable to statistical analysis, including the derivation of confidence bounds.

3. Just like for APA, it is possible to extend APA+ to work with more complicated summary information involving specific subsets of the data and more complicated queries. In the general case, APA+ makes use of many negative estimators, all of which are combined to form a single, unbiased estimator for the answer to the query.

An additional advantage of APA+ is its capability to deal with both categorical and numerical attributes in `WHERE` clause. In our technical report [15], we have shown how APA+ can be extended to handle numerical attributes.

Similar to the different versions of APA, i.e., APA0, APA1, etc., we also have different versions of APA+, which are referred to as APA0+, APA1+, and so on. The version that uses i-dimensional aggregates is referred to as APAi+. In the remainder of the paper, we formally describe APA+ and the derivation of its associated confidence bounds with a focus on categorical attributes in the `WHERE` clause. In the next section, we first study APA0+.

3 Negative Estimators and APA0+

To facilitate our discussion, we first introduce some definitions.

Let N be the total size of dataset, and T be the *zero-dimensional* fact that describes the total value of the attribute that is to be aggregated over all tuples in the database (in our example, T is `SUM (COMPLAINTS) = 148`). Let S denote a sample of size n, extracted from the complete data set using *sampling without replacement* [27].

Given a `WHERE` clause (or relational selection predicate) p, the *domain* of p (denoted D_p) is the set of data points in the complete dataset which satisfy the clause. The *domain sample* is the set of samples from S that satisfy the predicate p. The size of the domain sample is denoted as n_d. The i-th unit in the domain sample is represented by y_i. In our example, the selection clause `WHERE PROF = 'Smith'` specifies the domain sample $\{21, 7, 8\}$. The domain sample can be extended to the *domain data space* by assigning 0 to all the samples from S that do not satisfy the predicate. In our example, the *domain data space* is $\{0, 21, 0, 7, 8, 0, 0, 0\}$. We will use u_i to represent the i-th unit in the domain data space, and $\overline{u}$ to represent the mean of the domain data space, computed as $(\sum_1^n u_i)/n$.

Given these definitions, let $\hat{t}_p$ be the *naive* estimator for the value of a `SUM` query over D_p. This can be formally expressed as

$$\hat{t}_p = \frac{N}{n} \sum_{i=1}^{n_d} y_i = \frac{N}{n} \sum_{i=1}^{n} u_i$$

Note that in this paper, we also call this estimator a *positive* estimator, since its estimate is only based on the samples satisfying the predicate p. Also, $\hat{t}_p$ is unbiased:

Lemma 1 *The estimation of $\hat{t}_p$ is* unbiased, *i.e. the expectation of this estimator equals to the true value, $E(\hat{t}_p) = t_p$ [27].*

Furthermore, we also know that the variance of $V(\hat{t}_p)$ can be estimated as:

$$\widehat{V}(\hat{t}_p) = N^2(\frac{N-n}{Nn})s_u^2$$

where, $s_u^2 = 1/(n-1)\sum_{i=1}^{n}(u_i - \overline{u})^2$ is the sample variance in the domain data space [27].

In our example, $\hat{t}_p$ estimates the number of Prof. Smith's complaints to be $2 \times (21 + 7 + 8) = 72$. However, the correct number of Prof. Smith's complaints is 121. Clearly, the naive estimator significantly under-estimates the target variable. For this example, the *standard error (SE)*, which is the square root of the variance, (SE($\hat{t}_p$)=$\sqrt{\widehat{V}(\hat{t}_p)}$), is 68.2. In practice, such large relative standard errors can be common.

In the following subsections, we will describe how to utilize the zero-dimensional fact T to improve the naive estimator of the total of the domain D_p. Section 4 will show how these results can be generalized to make use of facts that provide more detailed information.

3.1 The Negative Estimator

Let $\overline{p}$ be the negation of the clause p. Thus, $\overline{p}$ specifies those tuples from the data set that are not in the domain D_p. Clearly, $\hat{t}_{\overline{p}}$ is also an unbiased estimator for $t_{\overline{p}}$, the total of the domain $D_{\overline{p}}$. In our running example, the negation of the original clause produces the following query:

```
SELECT SUM (NUM_COMPLAINTS)
  FROM COMPLAINTS
  WHERE PROF ≠ 'SMITH'
```

Just as $\hat{t}_p$ is a naive estimator for the number of complaints received by Professor Smith, $\hat{t}_{\overline{p}}$ is a naive estimator for the total number of complaints for all the professors except Smith: $\hat{t}_{\overline{p}} = 16/8 \times (2 + 1 + 0 + 3 + 4) = 20$.

A simple but important fact connects t_p, the total of the domain D_p, and $t_{\overline{p}}$, the total of the domain $D_{\overline{p}}$:

$$T = t_p + t_{\overline{p}}$$

Given this, we introduce a new estimator based on the negative clause:

Definition 1 *The* **negative estimator** *of the domain total t_p is $\hat{t}'_p = T - \hat{t}_{\overline{p}}$.*

Just as the positive estimator is unbiased, so is the negative estimator:

Lemma 2 *The negative estimator $\hat{t}'_p$ is an unbiased estimator for t_p, the total of domain D_p, and its variance is the same as the variance of $\hat{t}_{\overline{p}}$, i.e. $V(\hat{t}'_p) = V(\hat{t}_{\overline{p}})$.*

Proof: From Lemma 1, we can see that $\hat{t}_{\overline{p}}$ is an unbiased estimator of $t_{\overline{p}}$, i.e. $E(\hat{t}_{\overline{p}}) = t_{\overline{p}}$. Therefore, $E(\hat{t}'_p) = E(T - \hat{t}_{\overline{p}}) = E(T) - E(\hat{t}_{\overline{p}}) = T - t_{\overline{p}} = t_p$. Further, by the linear property of variance [3], $V(\hat{t}'_p) = V(\hat{t}_{\overline{p}})$. □

Also, directly from the above discussion, the variance of the *negative* estimator can be estimated as:

$$\widehat{V}(\hat{t}'_p) = \widehat{V}(\hat{t}_{\overline{p}})$$

Applying this new estimator to our example, the negative estimate for the number of complaints for Professor Smith is $148 - 20 = 128$, significantly reducing the standard error to 13.4. The improved accuracy can be attributed to the fact that the sample variance of the domain sample for the negative predicate is much smaller than the one for the original predicate in this example.

3.2 An Optimized Unbiased Estimator

So far, we have introduced two unbiased estimators, $\hat{t}_p$ and $\hat{t}'_p$ for the total of domain D_p. In this subsection, we study how to combine these two estimators together to have an optimized unbiased estimator whose variance is lower than the variance of either of the existing estimators.

Let the new estimator be the linear combination of two existing estimators, $\hat{t}_p$ and $\hat{t}'_p$:

$$\hat{t}_{APA0} = \alpha\hat{t}_p + (1-\alpha)\hat{t}'_p$$

where, $0 \leq \alpha \leq 1$. By the linearity of expectation, this new estimator is clearly unbiased as well:

Lemma 3 *The new estimator $\hat{t}_{APA0}$ is an unbiased estimator for the domain total t_p.*

Now the question is how to find an optimized α to minimize the variance of $\hat{t}_{APA0}$. For simplicity, in this section we will assume the positive estimator and negative estimator are *independent*. In this case, Lemma 4 states the existence of an *optimal* unbiased estimator. Note that the two estimator are actually correlated. In Section 5, we will discuss how to incorporate this correlation into our estimation approach.

Lemma 4 *Assume $V(\hat{t}_p) + V(\hat{t}_{\overline{p}}) \neq 0$, let $\alpha = \frac{V(\hat{t}_{\overline{p}})}{V(\hat{t}_p)+V(\hat{t}_{\overline{p}})}$, the estimator $\hat{t}^{opt}_{APA0} = \alpha\hat{t}_p + (1-\alpha)\hat{t}'_p, 0 \leq \alpha \leq 1$ is an unbiased estimator of the domain total t_p, and minimizes the variance of $\hat{t}_{APA0}$.*

Proof: We know that:

$$V(\hat{t}_{APA0}) = \alpha^2 V(\hat{t}_p) + (1-\alpha)^2 V(\hat{t}_{\overline{p}})$$

To minimize the variance of $\hat{t}_{APA0}$, we first look at the first order derivative of $V(\hat{t}_{APA0})$ with respect to variable α. By forcing the derivative to 0, we have:

$$\alpha = \frac{V(\hat{t}_{\overline{p}})}{V(\hat{t}_p) + V(\hat{t}_{\overline{p}})}$$

Also, the second order derivative of $\hat{t}_{APA0}$ is always larger than 0, when $V(\hat{t}_p) + V(\hat{t}_{\overline{p}}) \neq 0$. Therefore, the variance of $\hat{t}_{APA0}$ is minimized at our specified α. □

We call above estimator the *optimal* estimator. Since the variances of $\hat{t}_p$ and $\hat{t}_{\overline{p}}$ are usually unknown, it is not possible to compute α exactly; rather, the natural estimator $\hat{\alpha} = \frac{\widehat{V}(\hat{t}_{\overline{p}})}{\widehat{V}(\hat{t}_p)+\widehat{V}(\hat{t}_{\overline{p}})}$ is used.

Applying this new estimator to our running example, the optimized $\hat{\alpha}$ is 0.0373, and the estimated total is 125.95, with a reduced standard error 13.1.

3.3 Constructing a Confidence Interval

The standard method for describing the accuracy of an estimate is to associate a *confidence interval* with the estimate. For example, to associate a 95% confidence interval with our estimate, we choose a range of values so that with probability 95%, the true answer to the query will fall within this range. In practice, in order to provide a tight confidence interval the distribution of the estimator must be known.

As a result of the central limit theorem (CLT), the distribution of the positive estimator $\hat{t}_p$ is approximately normal (Gaussian) with mean t_p and variance $\widehat{V}(\hat{t}_p)$ for large n [20]. Thus, a large-sample $100(1-\alpha)\%$ confidence interval (CI) for the domain total t_p is:

$$[\hat{t}_p - z_{\alpha/2}SE(\hat{t}_p), \hat{t}_p + z_{\alpha/2}SE(\hat{t}_p)]$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)^{th}$ percentile of the standard normal distribution. The size (or width) of confidence interval is $2 \times z_{\alpha/2}SE(\hat{t}_p)$. For example, for a 95% CI, α is 5% and $z_{2.5}$ is 1.96. Therefore, an approximate 95% CI for the estimator $\hat{t}_p$ is given by:

$$[\hat{t}_p - 1.96SE(\hat{t}_p), \hat{t}_p + 1.96SE(\hat{t}_p)]$$

The following Lemma establishes an important property of of the distributions of the negative and optimal estimators.

Lemma 5 *If the distribution of the naive estimator on the negative domain is normal with mean $t_{\overline{p}}$, then the distribution of the negative estimator $\hat{t}'_p$ is normal with mean t_p and variance $V(\hat{t}_{\overline{p}})$. Furthermore, if the distribution of the naive estimator on the positive domain is also normal with mean t_p, the distribution of the optimal estimator $\hat{t}^{opt}_{APA0}$ is normal with mean t_p and variance $V(\hat{t}^{opt}_{APA0})$.*

As a result, the confidence intervals for $\hat{t}_p$ and $\hat{t}^{opt}_{APA0}$ can be derived as described above.

4 Generalization

In practice, it is easy to compute and store more complicated pre-aggregated results than those used in APA0+. This information can be used to achieve much greater estimation accuracy. In this Section, we will study how to build optimal estimators when more complicated pre-aggregation results are available. We refer to a fact with i predicates in its

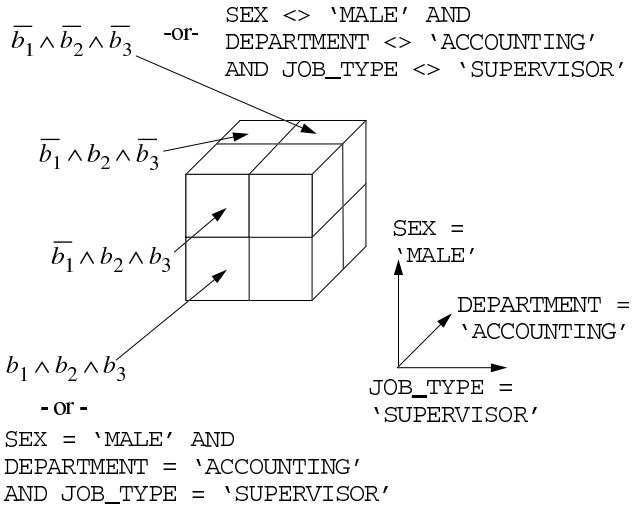


Figure 1. Spatial representation of query predicates

relational selection clause as an *i-dimensional fact*, and we refer to the version of APA+ that makes use of i-dimensional facts as APAi+. We begin this section with a few definitions, specifically, we define the idea of a negative clause with respect to more complicated pre-aggregation results.

We assume a relational selection clause p that can be expressed as a conjunction of m boolean predicates, $b_1, \cdots, b_m$. For example, consider the following query [14]:

```
SELECT SUM (SALARY)
  FROM EMPLOYEE
  WHERE SEX='M'
    DEPARTMENT='ACCOUNT'
    AND JOB_TYPE='SUPERVISOR'
```

In this case, we have b_1 = (SEX = 'M'), b_2 = (DEPARTMENT = 'ACCOUNT'), b_3 = (JOB_TYPE = 'SUPERVISOR'). We will also consider the negation of each of these predicates: $\overline{b_1}, \overline{b_2}, \overline{b_3}$. Note that if we take all possible meaningful combinations of these predicates, the entire dataset can be partitioned into 2^m non-overlapping domains, specified by 2^m clauses. We call such clauses *domain clauses*. In our example query, the $2^3 = 8$ domain clauses are $b_1 \wedge b_2 \wedge b_3, b_1 \wedge b_2 \wedge \overline{b_3}, b_1 \wedge \overline{b_2} \wedge b_3, b_1 \wedge \overline{b_2} \wedge \overline{b_3}, \overline{b_1} \wedge b_2 \wedge b_3, \overline{b_1} \wedge b_2 \wedge \overline{b_3}, \overline{b_1} \wedge \overline{b_2} \wedge b_3, \overline{b_1} \wedge \overline{b_2} \wedge \overline{b_3}$. Note that each of these clauses corresponds to a single cell in the multidimensional data cube defined by b_1, b_2, and b_3 (see Figure 2) [14].

Any conjunction over a subset of $\{b_1, b_2, \cdots, b_m\}$ can be expressed as a disjunction of the various domain clauses. Let the conjunctive clause $b_{j_1} \wedge \cdots \wedge b_{j_i}$ be denoted by $p_{j_1,\cdots,j_i}$. For APAi+, the fact involving such a conjunctive clause with i predicates is referred to as an *i-dimensional fact*, and is assumed to be available. This fact is an aggregate over the domain specified by this clause. In order to utilize i-dimensional facts, we introduce the following definitions.

Definition 2 *The set of 2^{m-i} domain clauses whose disjunction is equivalent to $p_{j_1,\cdots,j_i}$ is called the set of* **domain clauses with respect to** $p_{j1,\cdots,ji}$.

In our example, the domain clauses with respect to the conjunctive clause $p_1 = b_1$ (equivalent to SEX = MALE) are $b_1 \wedge b_2 \wedge b_3, b_1 \wedge b_2 \wedge \overline{b_3}, b_1 \wedge \overline{b_2} \wedge b_3, b_1 \wedge \overline{b_2} \wedge \overline{b_3}$; the domain clauses with respect to the conjunctive clause $p_{2,3} = b_2 \wedge b_3$ are $b_1 \wedge b_2 \wedge b_3$ and $\overline{b_1} \wedge b_2 \wedge b_3$.

This definition is important, because it will allow us to define a *negative clause* that will be used to form the negative estimators in higher-order versions of APA+:

Definition 3 *The set of domain clauses with respect to $p_{j1,\cdots,ji}$ except the predicate p is called the set of* **negative clauses with respect to** $p_{j_1,\cdots,j_i}$, *or $\overline{p}$ w.r.t. $p_{j_1,\cdots,j_i}$ for short.*

In our example, $\overline{p}$ w.r.t. p_1 is $(b_1 \wedge b_2 \wedge \overline{b_3}) \vee (b_1 \wedge \overline{b_2} \wedge b_3) \vee (b_1 \wedge \overline{b_2} \wedge \overline{b_3})$, and $\overline{p}$ w.r.t. $p_{2,3}$ is $\overline{b_1} \wedge b_2 \wedge b_3$.

Finally, we use $t_{j_1,\cdots,j_i}$ to denote the sum of the target attribute over the domain specified by $p_{j_1,\cdots,j_i}$, and use $t_{\overline{p}_{j_1,\cdots,j_i}}$ to denote the total of the domain defined by $\overline{p}$ w.r.t. $p_{j_1,\cdots,j_i}$. For example, $t_{b_1 \wedge b_2 \wedge \overline{b_3}}$ refers to the sum of the target attribute over the domain specified by the domain clause $b_1 \wedge b_2 \wedge \overline{b_3}$, and t_p refers to the answer to our query. Note that $t_{j_1,\cdots,j_i}$ is an i-dimensional fact.

The next four Subsections describe how to extend the techniques of Section 3 to develop APA1+. A generalization to APAi+ for arbitrary values of i is given in Section 4.5.

4.1 Negative Estimators in APA1+

Just like APA0+, APA1+ relies on the idea of a *negative estimator*. However, unlike APA0+, APA1+ will make use of *many* negative estimators. In APA1+, we assume that we have access to all one-dimensional facts over the data set. In our running example, this means that the totals t_1, t_2, t_3 are available. Using the notation given above, the positive estimator for the sum of the target attribute over the domain D_p is denoted by $\hat{t}_p = \hat{t}_{b_1 \wedge b_2 \wedge b_3}$. An estimator for the sum of the target attribute over the domain of a negative clause can be defined similarly, for example:

$$\hat{t}_{\overline{p}_1} = \hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}} + \hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3} + \hat{t}_{b_1 \wedge \overline{b_2} \wedge \overline{b_3}}$$

Since in APA1+ it is assumed that t_1 is available, t_1 and $\hat{t}_{\overline{p}_1}$ can easily be combined to form a new estimator for the answer to our query:

$$\hat{t}_{p_1} = t_1 - \hat{t}_{\overline{p}_1}$$

This is an example of a *negative estimator* in APA1+. In general:

Definition 4 $\hat{t}_{p_i} = t_i - \hat{t}_{\overline{p}_i}$ *is the* **APA1+ negative estimator** *for t_p where $p = b_1 \wedge b_2 \cdots b_m$, and $\overline{p}_i$ is the disjunction of all negative clauses with respect to the clause b_i, $1 \leq i \leq m$.*

Thus, there are a total of m negative estimators available in APA1+. Based on the same argument as in Lemma 2, we have the following lemma:

Lemma 6 *Each APA1+ negative estimator $\hat{t}_{p_i}$ is an unbiased estimator for t_p, the sum of the target attribute over D_p. The variance of $\hat{t}_{p_i}$ is the same as the variance of $\hat{t}_{\overline{p}_i}$, i.e. $V(\hat{t}_{p_i}) = V(\hat{t}_{\overline{p}_i})$.*

4.2 Combining Positive and Negative Estimators in APA1+

Just as in APA0+, the positive estimator and the negative estimators in APA1+ can be combined together to produce a new estimator with smaller variance.

The basic approach is similar to the method described in Subsection 3.2. We first introduce a new family of estimators based on the linear combination of the existing estimators. Formally, for a selection clause $p = b_1 \wedge b_2 \cdots b_m$, the new estimators are of the form:

$$\hat{t}_{APA1} = \alpha_0 \hat{t}_p + \alpha_1 \hat{t}_{p_1} + \cdots + \alpha_m \hat{t}_{p_m}$$

where $\hat{t}_p$ is the positive estimator, $\hat{t}_{p_i}$ is the negative estimator for $i > 0$, and $\alpha_0 + \alpha_1 + \cdots + \alpha_m = 1$. Clearly, $\hat{t}_{APA1}$ is unbiased as well:

Lemma 7 *The new estimator $\hat{t}_{APA1}$ is an unbiased estimator for the domain total t_p.*

4.3 Minimizing the Variance of $\hat{t}_{APA1}$

Because several negative estimators can contain the estimate associated with each cell in the cube illustrated in Figure 1, they will be strongly correlated. In order to derive the appropriate parameters to minimize the variance of $\hat{t}_{APA1}$, we first decompose each negative estimator based on the negative clause so that we can manage this correlation. For example, in our running example, the negative estimator $\hat{t}_{\overline{p}_1}$ can be decomposed as:

$$\hat{t}_{\overline{p}_1} = \hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}} + \hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3} + \hat{t}_{b_1 \wedge \overline{b_2} \wedge \overline{b_3}}$$

By decomposing each APA1+ negative estimator in this fashion, we can group like terms in order to reduce the pairwise correlation. The process is demonstrated in the following steps over the APA1+ estimator for our example query:

$$\begin{aligned}
\hat{t}_{APA1} = \alpha_0 \hat{t}_{b_1 \wedge b_2 \wedge b_3} + \\
\alpha_1(t_1 - (\hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}} + \hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3} + \hat{t}_{b_1 \wedge \overline{b_2} \wedge \overline{b_3}})) + \\
\alpha_2(t_2 - (\hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}} + \hat{t}_{\overline{b_1} \wedge b_2 \wedge b_3} + \hat{t}_{\overline{b_1} \wedge b_2 \wedge \overline{b_3}})) + \\
\alpha_3(t_3 - (\hat{t}_{\overline{b_1} \wedge b_2 \wedge b_3} + \hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3} + \hat{t}_{\overline{b_1} \wedge \overline{b_2} \wedge b_3})) = \\
\alpha_0 \hat{t}_{b_1 \wedge b_2 \wedge b_3} - (\alpha_1 + \alpha_2)\hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}} - (\alpha_1 + \alpha_3)\hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3} \\
-(\alpha_2 + \alpha_3)\hat{t}_{\overline{b_1} \wedge b_2 \wedge b_3} - \alpha_1 \hat{t}_{b_1 \wedge \overline{b_2} \wedge \overline{b_3}} \\
-\alpha_2 \hat{t}_{\overline{b_1} \wedge b_2 \wedge \overline{b_3}} - \alpha_3 \hat{t}_{\overline{b_1} \wedge \overline{b_2} \wedge b_3} + \alpha_1 t_1 + \alpha_2 t_2 + \alpha_3 t_3
\end{aligned}$$

In this way, the APA1+ estimator is transformed into the linear combination of estimators for the totals on a set of non-overlapping domains. Since the domains are non-overlapping, for the moment we will assume that all of the

estimators used in $\hat{t}_{APA1}$ are *pair-wise independent*. In reality, the estimators of the non-overlapping domain totals are correlated, and we will study correlation in Section 5. Using the pair-wise independence assumption, we can derive the variance for the new estimator:

$$V(\hat{t}_{APA1}) = \alpha_0^2 V(\hat{t}_{b_1 \wedge b_2 \wedge b_3}) + (\alpha_1 + \alpha_2)^2 V(\hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}}) + (\alpha_1 + \alpha_3)^2 V(\hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3}) + (\alpha_2 + \alpha_3)^2 V(\hat{t}_{\overline{b_1} \wedge b_2 \wedge b_3}) + \alpha_1^2 V(\hat{t}_{b_1 \wedge \overline{b_2} \wedge \overline{b_3}}) + \alpha_2^2 V(\hat{t}_{\overline{b_1} \wedge b_2 \wedge \overline{b_3}}) + \alpha_3^2 V(\hat{t}_{\overline{b_1} \wedge \overline{b_2} \wedge b_3})$$

To minimize the variance, in most of the cases, we can simply use Lagrange multipliers to optimize the values of the various α parameters. This procedure involves using a linear solver to solve a $m \times m$ linear equation, with cost $O(m^3)$. If the parameters (α_i) are invalid, i.e., they are negative, we can apply *quadratic programming* [14] to find the minimal variance. We refer to the estimator with the minimal variance as the *optimal estimator*, and denote it as $\hat{t}^{opt}_{APA1}$.

4.4 APA1+ Confidence Intervals

As discussed before, in order to provide a confidence interval, we have to know the distribution of the estimator.

Using the same argument as in Lemma 5, we have the following important property of optimal estimator $\hat{t}^{opt}_{APA1}$:

Lemma 8 *The distribution of the optimal estimator $\hat{t}^{opt}_{APA1}$ is normal if the distributions of each positive estimator for the total of each non-overlapped domain (the domain total specified by each conjunctive clause) are normal.*

Thus, just as in APA0+, for sufficiently large n each positive estimator of the domain total is approximately normal. Therefore, the distribution of the new optimal estimator is also approximately normal, with variance $\widehat{V}(\hat{t}^{opt}_{APA1})$.

The confidence interval of the new estimators can be derived the same as the positive estimator, i.e. a $100(1-\alpha)\%$ confidence interval (CI) for the domain total t_p is:

$$[\hat{t}^{opt}_{APA1} - z_{\alpha/2} SE(\hat{t}^{opt}_{APA1}), \hat{t}^{opt}_{APA1} + z_{\alpha/2} SE(\hat{t}^{opt}_{APA1})]$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)^{th}$ percentile of the standard normal distribution.

4.5 Generalizing the Method

The methods developed in this Section can be generalized to utilize higher-dimensional facts. For example, we can define APA2+ to be the version of APA+ that makes use of facts having two clauses in their relational selection predicate: `SUM (SALARY) WHERE (SEX = 'M' AND DEPARTMENT = 'ACCOUNT') = $2.1M`. For APA2+ we have:

$$\hat{t}_{APA2} = \alpha_0 \hat{t}_p + \alpha_{12} \hat{t}_{p_{12}} + \alpha_{13} \hat{t}_{p_{13}} + \alpha_{23} \hat{t}_{p_{23}}$$

where $\alpha_0 + \alpha_{12} + \alpha_{23} + \alpha_{13} = 1$ and $0 \leq \alpha_{12}, \alpha_{13}, \alpha_{23} \leq 1$. In a manner similar to APA1+, the optimal estimator for APA2+ is:

$$\hat{t}^{opt}_{APA2} = \alpha_0 \hat{t}_{b_1 \wedge b_2 \wedge b_3} + \alpha_{12}(t_{12} - \hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}}) + \alpha_{13}(t_{13} - \hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3}) + \alpha_{23}(t_{23} - \hat{t}_{\overline{b_1} \wedge \overline{b_2} \wedge b_3}) = \alpha_0 \hat{t}_{b_1 \wedge b_2 \wedge b_3} - \alpha_{12} \hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}} - \alpha_{13} \hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3} - \alpha_{23} \hat{t}_{\overline{b_1} \wedge \overline{b_2} \wedge b_3} + \alpha_{12} t_{12} + \alpha_{13} t_{13} + \alpha_{23} t_{23}$$

with each α chosen so as to minimize the variance.

It is also possible to extend the method to develop a *hierarchical* version of APA where we make use of APA0+, APA1+, and perhaps higher-order pre-aggregation together. For example, to use APA0+, APA1+, and APA2+ together we make use of the following linear combination:

$$\hat{t} = \alpha \hat{t}_p + \alpha_0 \hat{t}'_p + \alpha_1 \hat{t}_{p_1} + \alpha_2 \hat{t}_{p_2} + \alpha_3 \hat{t}_{p_3} + \alpha_{12} \hat{t}_{p_{12}} + \alpha_{13} \hat{t}_{p_{13}} + \alpha_{23} \hat{t}_{p_{23}}$$

where $\hat{t}_p$ is the positive estimator of the domain total t_p; $\hat{t}'_p$ is the negative estimator using APA0; $\hat{t}_{p_1}$, $\hat{t}_{p_2}$, and $\hat{t}_{p_3}$ are the negative estimators of APA1+; and $\hat{t}_{p_{12}}$, $\hat{t}_{p_{13}}$, and $\hat{t}_{p_{23}}$ are the negative estimators for APA2. As in APA0+ and APA1+, we have the constraint that $\alpha + \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_{12} + \alpha_{13} + \alpha_{23} = 1$,

5 Dealing with Correlation

In Sections 3 and 4, we assumed the estimators on each individual domain are pair-wise independent. However, the correlation does exist among these estimators. In this Section, we will first derive the closed formula of the correlation (covariance) among them (Subsection 5.1). Then we study how to incorporate the correlation factor into our estimators to improve their accuracy (Subsection 5.2).

5.1 Correlations among Estimators

In this Subsection, we first study the correlation for the estimators of the domain totals specified by a clause and its negation. This corresponds to the situation in APA0+, where the combined estimator is based on a pair of these clauses. Then we will study the general case, where we compute the correlation between two estimators on the non-overlapped domain totals. This corresponds to the situation where higher-order pre-aggregation information is used by the APA+ methodology.

We capture the correlation by computing the *covariance*. The covariance (Cov) between two estimators t_1 and t_2 is defined as follows [3]:

$$Cov(t_1, t_2) = E(t_1 t_2) - E(t_1)E(t_2)$$

We begin by considering the case of APA0+, where we have a domain clause and its negation. Let $\hat{t}_p$ and $\hat{t}_{\overline{p}}$ be the estimators of the totals for the domains specified by a selection clause p and its negative clause $\overline{p}$; t_p and $t_{\overline{p}}$ are the true totals of these domains, respectively. Lemma 9 states the covariance of these two estimators:

Lemma 9 *The covariance between the two estimators, $\hat{t}_p$ and $\hat{t}_{\overline{p}}$, is:*

$$Cov(\hat{t}_p, \hat{t}_{\overline{p}}) = -(t_p t_{\overline{p}})(\frac{N-n}{n(N-1)})$$

Proof: Please see our technical report [15]. □

Since N is typically very large, the covariance can be approximated as:

$$Cov(\hat{t}_p, \hat{t}_{\overline{p}}) \approx -\frac{t_p t_{\overline{p}}}{n}$$

Note that the covariance depends on the totals of the domain of p, and its negative $\overline{p}$, which are estimation targets. We will use the naive estimator to replace them, and therefore, we have:

$$Cov(\hat{t}_p, \hat{t}_{\overline{p}}) \approx -\frac{\hat{t}_p \hat{t}_{\overline{p}}}{n}$$

In the general case, we have two estimators $\hat{t}_{p_1}$ and $\hat{t}_{p_2}$ which are estimators of the totals for the domains specified by two clauses p_1 and p_2 where p_1 and p_2 are non-overlapping. Let t_{p_1} and t_{p_2} be the true sum of the target attribute over these domains, respectively. Lemma 10 shows that the covariance of these two estimators is actually the same as the case in which they are negations of one another, i.e. $p_2 = \overline{p_1}$.

Lemma 10 *The covariance between the two estimators, $\hat{t}_{p_1}$ and $\hat{t}_{p_2}$, is:*

$$Cov(\hat{t}_{p_1}, \hat{t}_{p_2}) = -(t_{p_1} t_{p_2})(\frac{N-n}{n(N-1)})$$

Proof: Please see our technical report [15]. □

Using the same techniques as we did in case of APA0+, we can approximate the covariance as follows:

$$Cov(\hat{t}_{p_1}, \hat{t}_{p_2}) \approx -\frac{t_{p_1} t_{p_2}}{n} \approx -\frac{\hat{t}_{p_1} \hat{t}_{p_2}}{n}$$

5.2 Using the Covariance

In this subsection, we study how to improve the new estimators developed in Section 3 and Section 4 after considering the correlation among each individual estimator. Note we will only discuss the variance of these new estimators since the expectation of these estimators is *not* affected by correlation.

APA0+: For the estimator $\hat{t}_{APA0} = \alpha \hat{t}_p + (1-\alpha)\hat{t}'_p, 0 \leq \alpha \leq 1$, considering the correlation between $\hat{t}_p$ and $\hat{t}'_p$, we have

$$\begin{aligned} V(\hat{t}_{APA0}) &= V(\alpha \hat{t}_p + (1-\alpha)(T - \hat{t}_{\overline{p}})) \\ &= V(\alpha \hat{t}_p - (1-\alpha)\hat{t}_{\overline{p}}) \\ = \alpha^2 V(\hat{t}_p) + (1-\alpha)^2 V(\hat{t}_{\overline{p}}) &- 2\alpha(1-\alpha) Cov(\hat{t}_p, \hat{t}_{\overline{p}}) \end{aligned}$$

To minimize the variance, we can use the standard mathematical methods, i.e. looking at the first and second order derivatives with respect to α, to find the desired parameter.

APAi+: Recall that in Section 4, we assumed the estimators on the domain clauses are pair-wise independent. As we showed in Subsection 5.1, they are in fact correlated. The covariance between any two estimators on the non-overlapped domains, $\hat{t}_{p_1}$ and $\hat{t}_{p_2}$, can be approximated as $Cov(\hat{t}_{p_1}, \hat{t}_{p_2}) \approx -\frac{\hat{t}_{p_1} \hat{t}_{p_2}}{n}$.

Given this, we can incorporate the correlation factor into our new estimators as follows. Consider the new estimator based on APA2+ as given in Subsection 4.5. The variance of the new estimator taking correlation into account is as follows:

$$\begin{aligned} V(\hat{t}_{APA2}) = \alpha_0^2 V(\hat{t}_{b_1 \wedge b_2 \wedge b_3}) + \alpha_{12}^2 V(\hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}}) + \\ \alpha_{13}^2 V(\hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3}) + \alpha_{23}^2 V(\hat{t}_{\overline{b_1} \wedge \overline{b_2} \wedge b_3}) - \\ \alpha_0 \alpha_{12} Cov(\hat{t}_{b_1 \wedge b_2 \wedge b_3}, \hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}}) - \\ \alpha_0 \alpha_{13} Cov(\hat{t}_{b_1 \wedge b_2 \wedge b_3}, \hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3}) - \\ \alpha_0 \alpha_{23} Cov(\hat{t}_{b_1 \wedge b_2 \wedge b_3}, \hat{t}_{\overline{b_1} \wedge \overline{b_2} \wedge b_3}) + \\ \alpha_{12} \alpha_{13} Cov(\hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}}, \hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3}) + \\ \alpha_{12} \alpha_{23} Cov(\hat{t}_{b_1 \wedge b_2 \wedge \overline{b_3}}, \hat{t}_{\overline{b_1} \wedge \overline{b_2} \wedge b_3}) + \\ \alpha_{13} \alpha_{23} Cov(\hat{t}_{b_1 \wedge \overline{b_2} \wedge b_3}, \hat{t}_{\overline{b_1} \wedge \overline{b_2} \wedge b_3}) \end{aligned}$$

Given this formula, an approach similar to the one given in Section 4 can be used to find the desired parameters to minimize the variance.

One issue that still needs to be explored is the effect of the covariance on the distribution of the various APA+ estimators. So far, we have relied on the CLTas justification of the normality of the APA+ estimators. However, after incorporating the correlation factor, the distribution of the combined estimator may not necessarily be normal, even if the CLT holds for each of the component estimators. It turns out that such distributions are very close to the class of *spherically symmetric distributions* [17], of which normal distribution is a special case. The analytical formula of the distribution of combined estimator is available in a technical report [15]. Though the exact confidence interval for such distributions is still hard to derive, our analysis and later experimental evaluation (Subsection 6.5) give strong evidence that normal distributions can serve as a good approximation for such distributions. Therefore, in the experimental results presented in Section 6, we simply utilize normal distributions to derive confidence intervals. Our experimental results validate this approach.

6 Experimental Results

This section presents an experimental benchmark of the methods proposed in the paper. Traditionally, benchmarks of approximate query processing techniques have focused on testing the absolute or relative error of the estimates provided by an approximation methodology. While we are guilty of publishing a few benchmarks of this type ourselves [14], we assert that this style of benchmarking can be of limited practical value. In practice, any useful approximation methodology must be able to *differentiate* among the cases when the estimation accuracy is likely to be good and when the accuracy is likely to be poor. In other words, an accurate estimate is of little use unless the methodology can *recognize* that the estimate is very accurate and the user can be notified accordingly.

As a result, this section focuses on studying the ability of APA+ to provide accurate estimates, and to recognize when those estimates are accurate. This is done by studying the *confidence bounds* provided by APA+. Specifically, our study focuses on two important questions:

1. How *accurate* are the confidence bounds provided by APA+? In other words, if APA+ guarantees a certain accuracy on an estimate, what are the real chances that the estimate meets those accuracy guarantees?
2. How *tight* are the bounds of our confidence intervals, as compared to traditional methods that could be used as an alternative to APA+? In other words, how much accuracy is APA+ able to guarantee, compared to the obvious alternatives?

6.1 Approximation Techniques Tested

In our experiments, we compare APA+ with two sampling-based alternatives. We concentrate on sampling for several reasons. Samples are widely used in the data management literature, and there is an extensive statistical literature relevant to associating confidence bounds with samples [27, 20]. Furthermore, sampling is arguably the most widely applicable approximation technique. Unlike other methods such as wavelets [29, 8, 9], samples are unaffected by data dimensionality, and work equally well with categorical and numerical data. Thus, the following approximation techniques were used in our comparison:

Simple Random Sampling: The sample is extracted from the original data set using the *sampling without replacement* method. The estimator is the *naive* unbiased estimator, and confidence bounds are derived using the central limit theorem (CLT).

Stratified Sampling: Stratification is a standard statistical technique that can be used to boost sampling-based estimation accuracy, and has previously been applied to problems in data management [4]. Stratified sampling works by partitioning the data set into a set of non-overlapping strata. In order to increase estimation accuracy, partitions can be chosen so as to minimize subsequent estimation variance. In our experiments, we use a uniform allocation of samples to strata. Experiments were conducted with 2 different sizes for each strata: 10 samples per strata and 100 samples per strata (in general, for a fixed number of total samples, using more strata and fewer samples per strata tends to increase estimation accuracy). Given a uniform allocation, the partitions are chosen so as to minimize the variance of the estimate of an aggregate query over the entire data space.

APA0+: We also implemented and tested *APA0+*, as described in this paper. APA0+ uses *zero-dimensional* facts to produce confidence bounds for each query. Recall that zero-dimensional facts are facts of the form `SUM(COMPLAINTS) = 123`. The additional memory cost of this method is minimal compared to simple random sampling; e.g. in the datasets that we tested with it was only a few bytes. APA0+ takes into account the covariance of its various sub-estimators while performing the estimation, as described in Sections 3 and 5.

exp. %DB match query	1 pred	2 pred	3 pred	4 pred	5 pred	6 pred	7 pred	8 pred
10%	4%	4%	4%	4%	4%	4%		
1%		4%	4%	4%	4%	4%	4%	
0.1%		4%	4%	4%	4%	4%	4%	4%
0.01%			4%	4%	4%	4%	4%	4%

Table 2. Distribution of Testing Queries

APA1+: In this method, *one-dimensional* facts are used. Each one-dimensional fact has exactly one clause in the `WHERE` predicate associated with the fact. For example, a one-dimensional fact is `SUM(SALARY) WHERE (SEX='F') = $1.9M`. The additional memory requirements of APA1+ compared to simple random sampling are still small, in practice only on the order of a few kilobytes. Just like the APA0+, APA1+ accounts for the covariance that possibly exists between the estimator's sub-cubes (Figure 4).

6.2 Experimental Setup

The four datasets used in our benchmark are derived from four real high dimensional data sets, previously used in [14]. The four data sets are: Forest Cover data (from the UCI KDD archive), River Flow data, William Shakespeare data (word proximity information), and Image Feature Vector data. The transformation of these datasets for our experiments were previously described in [14].

Some basic features of our experimental data sets are as follows. The first data set has 8 categorical attributes, and the last three have 30 categorical attributes each. These attributes are the ones that we use for *selection*. Each of the categorical attributes has five different categories. Each dataset also has 2 numerical attributes which are designated as our *measure* attributes. Selection attributes only appear in the relational selection predicates (`WHERE` clauses), whereas measure attributes appear in the aggregation functions of the `SELECT` clauses. Each `SELECT` clause is of the form `SELECT SUM(ATT)`. We create a total of 2000 queries for each data set, 1000 for each measure attribute. The `WHERE` clause in each predicate is a conjunction of boolean equality predicates on various categorical values. The number of predicates in each such conjunction varies from one to eight and are generated to vary the expected selectivity as shown in Table 2. The sampling rate used for our experiments was fixed at 10%.

6.3 Confidence Interval Accuracy

This subsection reports the results of a set of experiments designed to test the accuracy of the guarantees provided by each method. For each method and each query, a confidence interval is computed using a 95% user-specified confidence level. This confidence interval is considered to be *correct* if

Data Set	APA0+	APA1+	Sampling	ss = 10	ss = 100
Rivers Att 1	72.34%	96.80%	79.30%	74.80%	76.10%
Rivers Att 2	72.34%	92.57%	78.47%	74.85%	75.05%
Forest Att 1	71.73%	92.33%	64.75%	59.64%	59.53%
Forest Att 2	66.71%	92.57%	68.05%	57.82%	57.83%
Image Att 1	51.37%	77.37%	57.83%	51.59%	49.37%
Image Att 2	57.77%	82.97%	59.31%	54.17%	51.31%
Shakes Att 1	44.55%	90.14%	60.36%	56.14%	54.02%
Shakes Att 2	39.24%	89.40%	60.46%	55.03%	54.93%

Table 3. Observed Confidence When 95% Confidence Is Specified (*ss* stands for strata size used in conjunction with the stratified sampling approach).

the exact answer to the query falls within the bounds of the interval bounds. Since a user-specified confidence level of 95% is used, if each method works exactly as is theoretically expected, we should find that 95% of the confidence intervals were correct. The *actual* percentage of correct intervals for each of the data sets and approximation methodologies is reported in Table 3.

Discussion

The specified confidence for all of these methods was 95%, but it is clear from the results that only APA1+ comes close to achieving this confidence level experimentally. The confidence level achieved by APA1+ averaged 89.27% over all tests, which was 23.2% higher than simple random sampling (the next best alternative in terms of accuracy). The difference between APA1+ and the stratified-sampling-based techniques is even greater. Not only did APA1+ achieve the highest experimental confidence level, but it also comes very close to achieving the theoretically-expected 95% confidence based on the parameter settings (90+% confidence was observed for 5 out of 8 cases tested).

A few other points worth mentioning are:

- One possible explanation for the poor accuracy of the non-APA+ estimates is that the CLT-based intervals used turned out to be overly aggressive. The CLT is a *limiting theorem*, which is only guaranteed to hold given an infinite number of samples from a distribution. In practice, the CLT often holds after only a few dozen samples. However, under adverse circumstances (particularly with very skewed or bi-modal distributions) many more samples can be required. Given that several of our data sets were rather poorly-behaved (particularly the Image Feature Vector data set), this may be a cause of the problems that are evident in Table 3. On possible solution would be to use less-aggressive bounds (such as Chebychev or Hoeffding bounds). However, while this would improve the accuracy of the bounds, it will also increase the width of the associated confidence intervals, which already do not compare favorably with APA1+ (see Section 6.4).
- A factor that seems to cause problems with all five alternatives is that the variance used in computing the accuracy of any sampling-based estimate is always *itself* an estimate, that may or may not be accurate. This is a difficult problem without an easy solution, and is always present in practical statistical inference. The fact that stratified sampling demonstrates a confidence level that is lower than that of simple random sampling (with an average drop off of 5.93% over all data sets) seems to support this fact. Since the size of each strata is small, the variance estimate for each strata is likely to be more inaccurate than for simple random sampling, which effectively uses only a single strata. However, APA1+ seems to be least-affected by this problem.
- Finally, it is interesting to note that with APA1+ performed by far the best, APA0+ was on par with (or slightly worse than) stratified sampling in terms of the accuracy of the confidence level given. The key reason is that APA0+ uses only one negative estimator, as compared to APA1+, which uses $2^m - 2$. While one negative estimator can prove to be inaccurate, the degree of inaccuracy of multiple (APA1+) negative estimators would be far less. Therefore, we see an improved accuracy when a linear combination of them is taken in $\hat{t}^{opt}_{APA1}$.

6.4 Confidence Interval Width

Not only should the confidence bounds provided by an estimation methodology be *accurate*, they should also be *tight*, in the sense that the actual answer to the query should be constrained to fall in a very small interval. In this subsection we analyze tightness of confidence interval bounds produced by various methods. We perform pair-wise comparisons between APA0+ and APA1+ and each of the other three estimators, in order to determine which one produces smallest confidence intervals. For both APA0+ and APA1+, we compute the median change in confidence interval width that would have been obtained had another estimator been used instead. Specifically, we compute the percentage change for those queries that both estimators being compared answered correctly (that is, both estimators produced a correct confidence interval). The reason we focus on cases where both estimators were correct is that we do not want to reward an estimator for producing a very tight bound on an incorrect estimate.

Table 4 illustrates the shows median change in width of the confidence intervals produced by each of the three traditional methods compared to APA0+. For example, if, for a given query, both estimators produce correct confidence intervals with the size of APA0+ interval denoted by CI_0 and size of the other method's confidence interval denoted by CI, then the change (decrease) in confidence interval width for that pair would be $(CI - CI_0)/CI$.

Similarly, next table illustrates the shows median change in width of the confidence intervals produced by each of the three traditional methods compared to APA1+.

Discussion

For the data sets tested, APA0+ always produces tighter confidence intervals than simple random sampling. Across

Data Set	Sampling	ss = 100	ss = 10
Rivers Att 1	3.72%	-7.43%	-15.15%
Rivers Att 2	18.34%	15.30%	0.67%
Forest Att 1	15.08%	12.12%	4.46%
Forest Att 2	29.10%	9.79%	6.87%
Image Att 1	21.70%	-4.26%	8.21%
Image Att 2	29.88%	-8.30%	-14.14%
Shakes Att 1	16.39%	8.84%	-17.01%
Shakes Att 2	31.18%	12.62%	6.79%

Table 4. Confidence Interval Width Compared to APA0+ (*ss* stands for strata size used in conjunction with the stratified sampling approach).

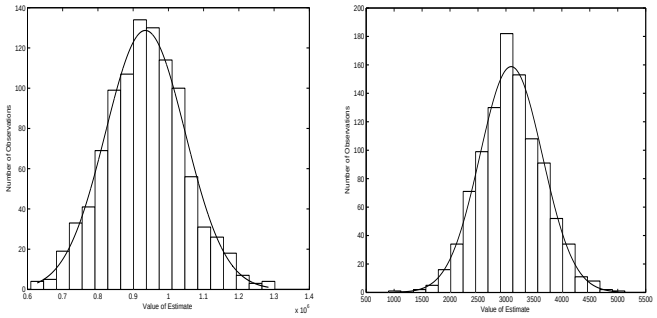
Data Set	Sampling	ss = 100	ss = 10
Rivers Att 1	50.93%	33.60%	21.60%
Rivers Att 2	58.76%	50.25%	30.13%
Forest Att 1	34.82%	30.97%	20.99%
Forest Att 2	43.45%	23.43%	20.50%
Image Att 1	49.26%	23.43%	15.04%
Image Att 2	56.00%	35.81%	6.19%
Shakes Att 1	62.36%	23.53%	13.58%
Shakes Att 2	55.97%	32.00%	23.09%

Table 5. Confidence Interval Width Compared to APA1+ (*ss* stands for strata size used in conjunction with the stratified sampling approach).

all data sets, APA0+ produced confidence intervals that were 14.81% smaller than for simple random sampling. The average improvement with APA1+ is even better. APA1+ provides confidence intervals that are on average 51% smaller than for simple random sampling, which means that when both estimators are predicting correctly, APA1+ will produce a confidence interval $1/2$ the size of that produced by random sampling. APA1+ is also clearly better than stratified sampling. Using 100 samples per strata results in intervals averaging 31.63% wider than those provided by APA1+, which translates into APA1+ intervals being about $2/3$ the size of those produced by this particular stratified sampling approach. Comparing APA1+ to stratified sampling approach using 10 samples per strata, the average of the median decrease (across data sets) is 18.89%, which translates into APA1+ intervals being about $4/5$ the size of their counterparts produced by stratified sampling. Furthermore, as demonstrated in the previous subsection, the APA1+ intervals are also far more accurate than the intervals produced using stratification.

6.5 Empirical Distributions for the APA+ estimators

Our experimental evaluation has shown that utilizing the normal distribution with the variance derived from APA+ estimators provides acceptable approximation of confidence intervals for our estimators. In the following, using two examples, we show that the empirical distributions of the APA+ estimators are indeed very close to the normal distributions. Figure 2 illustrates the empirical distribution (histogram) for two different queries on the Forest Cover dataset. The left one uses the APA0+ estimator and the right one uses the APA1+



(a) (b)

Figure 2. Empirical distributions for APA0+ (a) and APA1+ (b)

estimator. The curves in both the figures are the normal distributions (in the histogram scale) with mean and variance derived from APA0+ and APA1+ estimators, respectively. In other words, the point in the curve is the number of observations that would appear in a bucket-size area for the corresponding normal distributions. These plots show that our estimators have a strong tendency towards the normal distributions.

7 Related Work

Sampling methods have been widely used in the database community for approximate query answering [1, 2, 4, 7, 10, 13, 12, 14] and selectivity estimation [7, 18, 19]. Though confidence intervals have found importance in aggregation estimation [13], this issue was not addressed in these previous research [1, 4, 7, 14]. In particular, most of these methods rely on stratified sampling to improve the estimation accuracy. Our approach focuses on providing accurate confidence intervals associated with the estimation results by utilizing additional summary information. Note that such information has been used in [14] to improve the estimation accuracy. As discussed in previous sections, our work is significantly different from previous approaches with respect to both the goals and the details of the methods.

Improving the accuracy of confidence intervals has been an active research topic in statistics. The typical approaches include stratified sampling [27, 20] and bootstrapping [5]. Our approach has shown better performance than stratified sampling in terms of providing tighter and more accurate confidence intervals. Bootstrapping requires resampling hundreds of times in order to improve the confidence intervals, and therefore is computationally expensive. Recently, Pol and Jermaine have developed a middleware to support bootstrapping in a database system [23]. Compared with bootstrapping, our approach is very simple and computationally efficient. It is very easy implement, and also applicable in a streaming environment (the summary information can be collected in a single pass). Our approach is essentially a composite-based estimation from statistics [24, 17], which targets the accuracy of the estimation. In particular, most of composite estimations in statistics consider only a very small number of estimators (usually two, one unbiased, one biased) [24, 17] and the confi-

dence intervals are very hard to provide for such estimations.

Finally, Garofalakis *et. al.* have studied wavelet synopses with maximal error guarantees [8, 9]. However, such approach is not meant for categorical attributes [14]. In comparison, our approach can handle both categorical attributes and numerical attributes in the WHERE clauses.

8 Conclusions and Future Work

The sampling-based estimators described in this paper are very useful in a database environment when it is possible to collect summary information in addition to the sample. The estimators have the advantage of being unaffected by data dimensionality, and are suitable for use with categorical and numerical data. Furthermore, since the information used by the estimators can be collected in a single pass over a data set, the estimators are suitable for use in a streaming environment.

One important issue that we have not considered in this paper is use of the estimators for queries other than SUM and COUNT queries. In particular, extending our methods to work with AVERAGE queries is an important problem for future work. AVERAGE queries can be treated as a ratio of a SUM and a COUNT query, but since the two estimators will be correlated if they are the result of the same sample, it becomes necessary to use relatively pessimistic confidence bounds in the absence of a rigorous study of the correlation of the two estimators.

References

[1] Swarup Acharya, Phillip B. Gibbons, and Viswanath Poosala. Congressional samples for approximate answering of group-by queries. *SIGMOD Conference Proceedings*, pages 487–498, May 2000.

[2] Swarup Acharya, Phillip B. Gibbons, Viswanath Poosala, and Sridhar Ramaswamy. Join synopses for approximate query processing. *SIGMOD Conference Proceedings*, pages 275–286, June 1999.

[3] George Casella and Roger L. Berger. *Statistical Inference, 2nd. Edition*. DUXBURY Publishers, 2001.

[4] Surajit Chaudhuri, Gautam Das, and Vivek Narasayya. A robust, optimization-based approach for approximate answering of aggregate queries. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, 2001.

[5] B. Efron and R. Tibshirani. *An Introduction to Bootstrap*. Chapman & Hall, New York, 1993.

[6] Sumit Ganguly, Phillip B. Gibbons, Yossi Matias, and Abraham Silberschatz. Bifocal sampling for skew-resistant join size estimation. *SIGMOD Conference Proceedings*, pages 271–281, June 1996.

[7] Venkatesh Ganti, Mong-Li Lee, and Raghu Ramakrishnan. Icicles: Self-tuning samples for approximate query answering. In *VLDB*, pages 176–187, 2000.

[8] Minos Garofalakis and Phillip B. Gibbons. Wavelet synopses with error guarantees. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, 2002.

[9] Minos N. Garofalakis and Amit Kumar. Deterministic wavelet thresholding for maximum-error metrics. In *PODS*, pages 166–176, 2004.

[10] Phillip B. Gibbons and Yossi Matias. New sampling-based summary statistics for improving approximate query answers. *SIGMOD Conference Proceedings*, pages 331–342, June 1998.

[11] Dimitrios Gunopulos, George Kollios, Vassilis J. Tsotras, and Carlotta Domeniconi. Approximating multi-dimensional aggregate range queries over real attributes. *SIGMOD Conference Proceedings*, pages 463–474, May 2000.

[12] Peter J. Haas and Joseph M. Hellerstein. Ripple joins for online aggregation. *SIGMOD Conference Proceedings*, pages 287–298, June 1999.

[13] Joseph M. Hellerstein, Peter J. Haas, and Helen J. Wang. Online aggregation. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, 1997.

[14] Chris Jermaine. Robust estimation with sampling and approximate pre-aggregation. *VLDB Conference Proceedings*, pages 886–897, August 2003.

[15] Ruoming Jin, Leo Glimcher, Chris Jermaine, and Gagan Agrawal. New sampling-based estimators for olap queries. Technical Report OSU-CISRC-6/05-TR44, The Ohio State University, June 2005.

[16] Norman L. Johnson, Samulel Kotz, and Adrienne W. Kemp. *Univariate Discrete Distributions*. Wiley Interscience, 1993.

[17] E.L. Lehmann and George Casella. *Theory of Point Estimation*. Springer-Verlag, 1998, 2nd Edition.

[18] Richard J. Lipton and Jeffrey F. Naughton. Query size estimation by adaptive sampling. *PODS Conference Proceedings*, pages 40–46, April 1990.

[19] Richard J. Lipton, Jeffrey F. Naughton, and Donovan A. Schneider. Practical selectivity estimation through adaptive sampling. *SIGMOD Conference Proceedings*, pages 1–11, May 1990.

[20] Sharon L. Lohr. *Sampling: Design and Analysis*. Duxbury Press, 1999.

[21] James F. Lynch. Analysis and application of adaptive sampling. *PODS Conference Proceedings*, pages 260–267, May 2000.

[22] Frank Olken and Doron Rotem. Simple random sampling from relational databases. *VLDB Conference Proceedings*, pages 160–169, August 1986.

[23] Abhijit Pol and Chris Jermaine. Relational confidence bounds are easy with the bootstrap. In *SIGMOD Conference Proceedings*, June 2005.

[24] Rao03. *Small Area Estimation*. John Wiley & Sons, Inc., New Jersey, 2003.

[25] Carl-Erik Sarndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Spriger-Verlag, 1992.

[26] Jun Shao. *MAthematical Statistics*. Spriger-Verlag, 2003, 2nd Edition.

[27] Steven K. Thompson. *Sampling, Second Edition*. Wiley, 2002.

[28] TPC. Tpc-h benchmark. *http://www.tpc.org/tpch/*, 2004.

[29] Jeffrey Scott Vitter and Min Wang. Approximate computation of multidimensional aggregates of sparse data using wavelets. *SIGMOD Conference Proceedings*, pages 193–204, June 1999.