TRAPEZOIDAL ATTRIBUTE CARDINALITY MAP: A NEW HISTOGRAM-LIKE TECHNIQUE FOR QUERY OPTIMIZATION

B. John Oommen and Murali Thiyagarajah

TR-99-06 February 1999

School of Computer Science, Carleton University
Ottawa, Canada, KIS 5B6

Trapezoidal Attribute Cardinality Map: A New Histogram-like Technique for Query Optimization

B. John Oommen* and **Murali Thiyagarajah†**
School of Computer Science
Carleton University
Ottawa, Canada K1S 5B6
{oommen, murali}@scs.carleton.ca

Key Words: Query Optimization, Query Result Size Estimation

Abstract

Histogram techniques are used to efficiently estimate query result sizes in most of the modern-day database systems. This is done by approximating the frequencies of the attribute values of the underlying data distributions. Even though they have been in use for nearly two decades, there has been no significant mathematical techniques (other than those used in statistics for traditional histogram approximations) to study them. In a recent work [12], we introduced a new histogram-like approximation strategy, called the Rectangular Attribute Cardinality Map (R-ACM), which approximates the density function within a given sector by a rectangular cell. In this paper, we introduce another histogram-like approximation strategy, called the Trapezoidal Attribute Cardinality Map (T-ACM), that aims to approximate the density of the underlying attribute values using the philosophies of numerical integration.

In this new histogram-like approximation method, the density function within a given sector is approximated by a trapezoidal cell, where the slope of the trapezoid is obtained so as to guarantee that the actual probability mass within the cell equals the true probability mass. Analytically, we show that for the T-ACM, the distribution of an attribute value within the sector is Binomially distributed. This permits us to derive worst-case and average-case results for the estimation errors of the probability mass itself. Our theoretical results, which include a rigorous maximum likelihood and expected-case analyses, and an extensive set of experiments demonstrate that the T-ACM scheme (which is essentially histogram-like) is much more accurate than the traditional equi-width and equi-depth histograms for query result size estimation. Due to its high accuracy and low construction costs, we hope that, along with the R-ACM introduced in [12], the T-ACM could become an invaluable tool for query optimization in the future database systems.

*Senior Member, IEEE. Partially supported by the Natural Sciences and Engineering Research Council of Canada.

†Supported by the Natural Sciences and Engineering Research Council of Canada

Vendor	Product	Histogram Type
IBM	DB2-6000 (Client-Server)	Compressed(V,F) Type
IBM	DB2-MVS	Equidepth, Subclass of End-Biased(F,F)
Oracle	Oracle7	Equidepth
Sybase	System 11	Equidepth
Tandem	NonStop SQL/MP	Equidepth
NCR	Teradata	Equidepth
Informix	Online Data Server	Equidepth

Table 1: Histograms used in commercial DBMSs.

1 Introduction

Query optimization for relational database systems is a combinatorial optimization problem, which requires estimation of query result sizes to select the most efficient access plan for a query based on the estimated costs of various query plans.

Query result sizes are usually estimated using a variety of statistics that are maintained in the database catalogue for relations in the database. Since these statistics approximate the distribution of data values in the attributes of the relations, they represent an inaccurate picture of the actual contents of the database. It has been shown in [5] that errors in query result size estimates may increase exponentially with the number of joins. This result, in light of the complexity of present-day queries, shows the the critical importance of accurate result size estimation.

Several techniques have been proposed in the literature to estimate query result sizes, including histograms, sampling, and parametric techniques [2, 3, 7, 9, 11, 13, 14]. Of these, histograms are the most commonly used form of statistics, which are also used in commercial database systems such as Microsoft SQL Server, Sybase and DB2. A more comprehensive list is shown in Table 1.

In a recent work [12], we introduced a new histogram-like approximation strategy called the Rectangular Attribute Cardinality Map (R-ACM), where the density function within a given sector was approximated by a rectangular cell. In the R-ACM, the height of the rectangular cell was obtained so as to guarantee that the actual probability density differs from the approximated one by a maximum user-specified tolerance, τ. In this paper we introduce another new catalogue-based non-parametric statistical model called the *Trapezoidal Attribute Cardinality Map* (T-ACM) that can be used to obtain more accurate estimation results than the traditional estimation techniques. We argue that the T-ACM can be used as a fundamental tool for query result-size estimation, and provide the mathematical foundation for its use. These arguments are fully supported by a formal maximum likelihood analysis, an expected-case analysis of the variance, and the resulting worst-case and average-case errors. A brief summary of some of the experimental results we obtained using a real-world database (U.S. CENSUS) is also included here, which clearly demonstrates the

superiority of the T-ACM over the currently used strategies.

2 Previous Work

In the interest of brevity, it is impossible to give a good review of the field here. Such a review is found in [15]. However, to present our results in the right perspective a brief survey is given.

Equi-width histograms for single-attributes were considered by Christodoulakis [1] and Kooi [7]. Since these histograms traditionally have the same width, they produce highly erroneous estimates if the attribute values are not uniformly distributed. The problem of building equi-depth histograms on a single attribute was first proposed by Piatetsky-Shapiro and Connell [13]. This was later extended as multi-dimensional equi-depth histograms to represent multiple attribute values by Muralikrishna and Dewitt [11].

Ioannidis and Christodoulakis took a different approach by grouping attribute values based on their frequencies [5, 4]. In these serial histograms, the frequencies of attribute values associated with each bucket are either all greater or all less than the frequencies of the attribute values associated with any other bucket. They also considered optimal serial histograms that minimize worst case error propagation in the size of join results [5]. The serial histograms provide optimal results for equality join estimations, but less than optimal results for range queries. Faloutsos *et al* [3] proposed using a multi-fractal assumption for real-data distribution as opposed to the uniformity assumptions made within current histograms.

Ioannidis and Poosala [6] discussed the design issues of various classes of histograms and of strategies for balancing their practicality and optimality in query optimization. They investigated various classes of histograms using different constraints (V-Optimal, MaxDiff, Compressed, and Spline-based) and sort and source parameters (Frequency, Spread, and Area). They also provided various sampling techniques for constructing the above histograms and concluded that the V-optimal histogram is the most optimal one for estimating the result sizes of equality-joins and selection predicates.

3 Trapezoidal Attribute Cardinality Map

The Trapezoidal Attribute Cardinality Map (T-ACM) of a given attribute, in its simplest form, is a one-dimensional integer array that stores the count of the tuples of a relation corresponding to that attribute, and for some equal subdivisions for the range of values assumed by that attribute. The T-ACM is, in fact, a modified form of the histogram. But unlike the Rectangular Attribute Cardinality Map (R-ACM) [12] and the two major forms of histograms, namely equi-width histogram [7] and the equi-depth histogram [13], where the histogram buckets are rectangular cells, the sectors in the T-ACM are trapezoidal cells. The Trapezoidal Attribute Cardinality Map (T-ACM) is a non-parametric histogram-like estimation technique based on the trapezoidal-rule of numerical integration, which generalizes the R-ACM from a "step" representation to a "linear" representation.

Symbol	Explanation
x_i	Number of tuples in attribute X for the i^{th} value of X.
$E(X_i)$	Expected number of tuples in attribute X for the i^{th} value of X.
n_j	No of tuples in the j^{th} sector of an ACM.
l_j	No of distinct values in the j^{th} sector. (Also called sector width).
s	Number of sectors in the ACM.
τ	Allowable tolerance for an R-ACM
ξ	Size of a relation.
N	Number of tuples in the relation.

Table 2: Notations Used in the Paper

In this paper we introduce the T-ACM as one of the two fundamental tools (other being the R-ACM introduced in [12]) for query result-size estimation and provide a mathematical foundation for its use. We also investigate the use of the T-ACM in the result-size estimation of various relational operations. Also, as in the case of the R-ACM, we propose to store and maintain the T-ACM in the DBMS catalogue.

Before we proceed, we present the notations used in this paper in Table 2.

The T-ACM can be either a one-dimensional or a multi-dimensional depending on the number of attributes being mapped. To introduce the concepts formally, we shall deal with the one-dimensional case in this paper.

A trapezoidal ACM is a modified form of the equi-width histogram where each histogram partition is a trapezoid instead of a rectangle. In fact, the trapezoidal ACM is obtained by replacing each of the rectangular sectors of the equi-width histogram by a trapezoid. The beginning and ending frequency values of each trapezoid sector is chosen so that the area of the resulting trapezoid will be equal to the area of the "rectangle" of the histogram it is replacing.

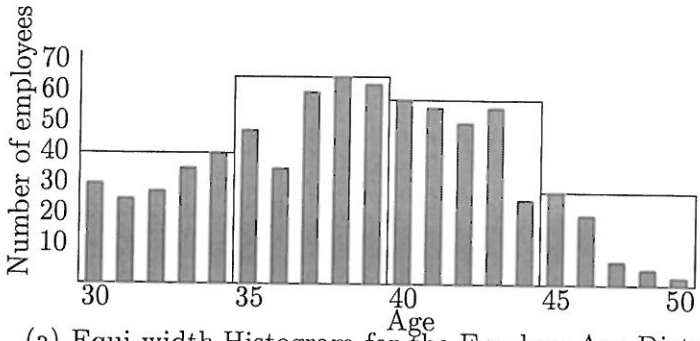
Definition 1 A One dimensional Trapezoidal ACM: *Let* $\mathcal{V} = \{v_i : 1 \leq i \leq |\mathcal{V}|\}$, *where* $v_i < v_j$ *when* $i < j$, *be the set of values of an attribute* X *in relation* R. *Let the value set* $\mathcal{V}$ *be subdivided into* s *equi-width sectors, each having sector width,* l. *We approximate each equi-width sector by a trapezoid in which the* j^{th} *trapezoid is obtained by connecting the starting value,* a_j, *to the terminal value,* b_j, *where the quantities* a_j *and* b_j *satisfy:*

(a) The starting value a_1 *is a user-defined parameter.*

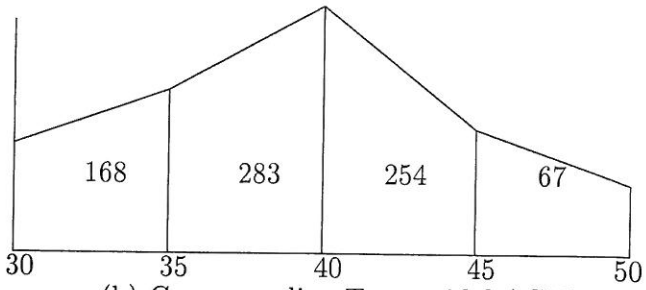
(b) For all $j > 1$, *the starting value of the* j^{th} *trapezoid,* a_j, *is the terminal value of the* $(j-1)^{st}$ *trapezoid,* b_{j-1}.

(c) The area of the j^{th} *trapezoid exactly equals the area of the* j^{th} *equi-width sector from which the exact computation of the quantity,* b_j, *is possible.*

Then the Trapezoidal Attribute Cardinality Map of attribute X *with initial attribute value* X_1 *and width* l *is the set* $\{(a_i, b_i) | 1 \leq i \leq s\}$.



(a) Equi-width Histogram for the Employe Age Distribution



(b) Corresponding Trapezoidal ACM

Figure 1: An Example for Constructing the Trapezoidal Attribute Cardinality Map

Example 1 *Figure 1 shows the equi-width histogram and the trapezoidal ACM of the* `Age` *attribute of a relation* `Emp(SIN, Age, Salary)` *between* `Age = 30` *and* `Age = 49`. *Note that the actual frequency for every age value is shown in the histogram as shaded rectangles. As can be noticed, the starting and ending frequencies of each trapezoidal sector is chosen so that the area under the trapezoid is equivalent to the area of the corresponding rectangular sector of the histogram. From the trapezoidal ACM, the number of tuples in the relation with ages in the range of* $35 \leq$ `Age` ≤ 40 *is 283 and the estimate for the number of employees having* `Age = 48` *is 6.*

Our motivation for proposing the trapezoidal ACM for density approximation (and the query result size estimation) originates from considering the various techniques used in numerical integration. Finding the result size of a selection query on a range-predicate can be considered as a discrete case of finding the area under a curve. Thus any numerical integration technique used to find the area under a curve will fit our purpose well. Though more accurate and sophisticated methods such as Simpson's Rule exist, since the trapezoidal method is relatively easy to use in a DBMS setting and is much superior to the traditional equi-width and equi-depth histograms currently in use, we have opted to use the trapezoidal method. In addition to providing more accurate result estimation on selection queries on range predicates, it also gives better results on equality-match predicates.

3.1 Generating Trapezoidal ACM

Unlike the R-ACM, where the sector widths are variable, the sector widths of a T-ACM are all equal. Each sector or cell of a T-ACM stores the frequency values of the first and last attribute values in that sector, in addition to the number of tuples in the sector. Algorithm `Generate_T-ACM` partitions the value range of the attribute X into s equal width sectors of the T-ACM. The input to the algorithm is the number of partitions, s. The frequency distribution is assumed to be available in an integer array A, which has a total of L entries for each of the L distinct values of X. For simplicity, we assume that the attribute values are ordered integers from 0 to $L-1$. The output of the algorithm is the T-ACM for the given attribute value set. Since choosing the starting frequency value of the first trapezoidal sector is important for obtaining the subsequent a_j's and b_j's, we briefly discuss it below.

3.1.1 Determining the First Frequency Value, a_1

As we shall see later from Lemmas 1 and 2, if the frequency of the first attribute value of the first sector of a T-ACM is known, the subsequent a_j's and b_j's can be easily obtained. The problem of obtaining an optimal starting frequency for building a T-ACM is still open and is currently being investigated. Below we have listed some of the methods that can be used to obtain this quantity:

(1) a_1 is a user-defined frequency value.

(2) a_1 is obtained using the average of all the frequencies in the given attribute value domain.

(3) Use the frequency value from (2) above as the starting frequency of the first sector and compute all the a_j's and b_j's in a *left-to-right* manner. Again use the frequency value from (2) above as the terminal frequency of the last sector and compute all the a_j's and b_j's in a *right-to-left* manner. One possibility is to assign a_1 to be the average of the first frequency values resulting from these two methods.

Before presenting the `Generate_T-ACM` algorithm that generates a T-ACM, we shall first present two lemmas that are used in this algorithm.

Lemma 1 *For each sector in the T-ACM, the number of tuples, n_j, is equal to,*

$$n_j = \left(\frac{a+b}{2}\right).l,$$

where a, b are the frequencies of the first and last attribute value in the sector and l is the number of distinct values in the sector.

Proof: This can be easily shown using the geometry of the trapezoidal sector. □

This lemma is important because ensuring that n_j is close to $(a+b)l/2$ would provide us the desired accuracy using trapezoidal approximation.

Let a_j be the frequency of the first attribute value in the j^{th} sector. The first frequency value of the first sector, a_1 can be chosen to be either the actual frequency of the attribute value (i.e: $a_1 = x_1$) or the average frequency of the entire attribute value range (i.e: $a_1 = \frac{N}{sl}$). The subsequent values for $a_j, 2 \leq j \leq s$, do not need to be stored explicitly and can be obtained from Lemma 2.

Lemma 2 *If the frequency of the first attribute value of the first sector of a T-ACM is a_1, then the frequency of the first attribute value of the j^{th} T-ACM sector, $a_j, 2 \leq j \leq s$, is given by,*

$$a_j = (-1)^{j-1}\frac{2}{l}\left\{a_1 + \sum_{k=1}^{j-1}(-1)^k n_k\right\}$$

where n_k is the number of tuples in the k^{th} sector.

Proof: Given the frequency of the first attribute value in a T-ACM sector, the frequency of the last attribute value in that sector can be obtained by using Lemma 1. Hence we have the following first and last frequency values a_j's and b_j's for a T-ACM.

$$\begin{array}{ll} a_1 = a & b_1 = \frac{2n_1}{l} - a \\ a_2 = b_1 = \frac{2n_1}{l} - a & b_2 = \frac{2n_2}{l} - a_2 \\ \vdots & \vdots \\ a_j = b_{j-1} = (-1)^{j-1}\frac{2}{l}\left\{a_1 + \sum_{k=1}^{j-1}(-1)^k n_k\right\} & b_j = \frac{2n_j}{l} - a_j \end{array}$$

Hence the lemma. □

Algorithm 1 Generate_T-ACM

```
Input:  No of sectors, s, frequency distrib.of X as A[0...L-1]
Output:  T-ACM
begin
   Initialize_ACM;      /* set all entries in ACM to zero */
   ACM[1].a := (sum_{i=0}^{L-1} A[i]) / L ; /* set a_1 to average frequency */
   for j := 1 to s do /* for every sector */
      for i := 1 to l do /* for every attrib.value */
         ACM[j].n := ACM[j].n + A[(j-1)*l + i];
      end; { for }
      if (j > 1) then ACM[j].a := ACM[j-1].b;
      ACM[j].b := 2 * ACM[j].n/l - ACM[j].a;
   end; { for }
end;
EndAlgorithm Generate_T-ACM
```

In practice, we can obtain a_j easily from a_{j-1} as shown in the Algorithm 1. We also obtain a_1 by averaging the frequency values of the entire attribute range. Note that each entry of the ACM array is a record with three fields, namely n, a, b, which store the number of tuples, the frequency of the first value and the frequency of the last value in the sector respectively.

It is obvious that the algorithm, `Generate_T-ACM` generates the T-ACM corresponding to the given frequency value set. Assuming the frequency distribution of X is already available in array A, the running time of the algorithm `Generate_T-ACM` is $O(L)$ where L is the number of distinct attribute values.

3.1.2 Implementation Details

Since the T-ACM is based on the trapezoidal rule of numerical integration, we expect it to be more accurate than the corresponding equi-width histogram of the same sector width. We shall now describe a process by which we can obtain a T-ACM that is even much more accurate than the T-ACM generated by the above algorithm `Generate_T-ACM`.

Let us assume that T_A is the T-ACM derived from the equi-width histogram, H_A. Also assume that the frequencies of the starting attribute value of *every second* histogram sector are available. Observe that this can be computed in $O(s)$ time (as opposed to $O(L)$), where s is the number of sectors. Then we can generate a T-ACM which has a sector width that is *half* of the sector width of H_A and is much more superior than the initial T-ACM, T_A. The strategy to achieve this is given in the algorithm, `Implement_T-ACM`, below.

Algorithm 2 Implement_T-ACM

```
Input:  (i) Equi-width histogram H_A with sector width l.
        (ii) Starting frequencies of every 2^nd sector.
Output: T-ACM with sector width l/2.
begin
   Merge every two adjacent sectors of H_A to get H_B;
   Generate T-ACM, T_A from H_B.
   Estimate frequencies of T_A's middle attribute values using Lemma 1.
   Generate T_B from T_A using frequencies obtained from the last step.
   Estimate frequencies of T_B's middle attribute values using Lemma 1.
   Generate T_C from T_B using frequencies obtained from the last step.
end;
```

EndAlgorithm Implement_T-ACM

Since the trapezoidal rule of numerical integration is more accurate than the left-end or right-end rectangular rule of numerical integration, it is obvious that by virtue of the construction of the T-ACM, T_A is more accurate than the equi-width histogram H_B. We note that the area of a sector in T_A may not be exactly equal to the actual number of tuples

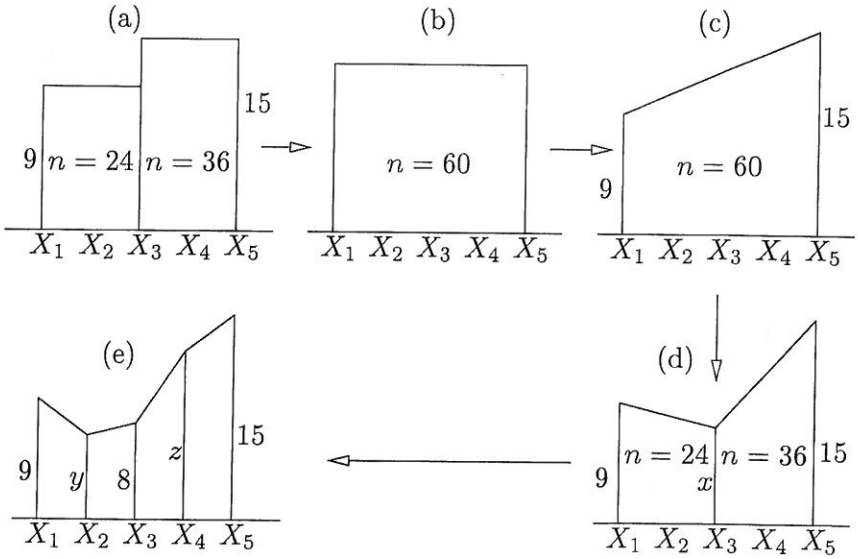
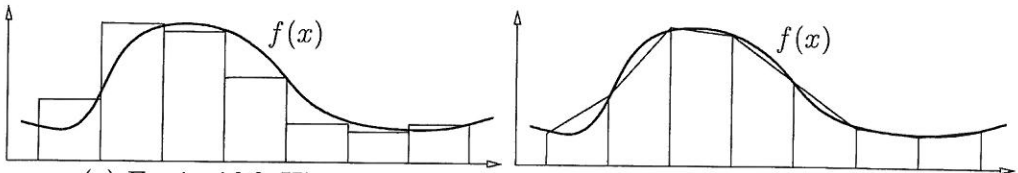


Figure 2: Construction of a T-ACM

falling in that sector. Hence, using the actual number of tuples within the sectors and applying Lemma 1, we can partition the sectors to obtain the T-ACM, T_B, where the sector areas represent the *actual* number of tuples more accurately. Using the same arguments, we obtain the T-ACM, T_C, from T_B. The T-ACM, T_C, can be expected to be more accurate than the T-ACM, T_B, as its boundary frequencies are more accurate estimates based on the *actual number* of tuples falling within the sectors.

Thus, by invoking a small preprocessing step, we have generated a T-ACM that is much more accurate than the original T-ACM derived directly from the corresponding histogram. An example highlighting the steps taken by the above algorithm is shown in Figure 2. In this example, the input to the algorithm is an equi-width histogram with two sectors as shown in Figure 2(a). The number of tuples in the sectors are $n = 24$ and $n = 36$ respectively. Also the starting frequency value of the first sector is 9 and the terminal frequency of the second sector is 15. The first step of the algorithm merges these two sectors to create a single histogram sector shown in Figure 2(b). The next step generates a trapezoidal sector equivalent to this larger histogram sector. Since the trapezoidal sector is only an approximation of the number of tuples represented by the rectangular sector, its area may not reflect the actual number of tuples ($n = 60$) falling within that sector. Hence in the next step of the algorithm, we *estimate* the middle frequency (i.e: $x_3 = 8$) of this sector by applying Lemma 1. The resulting T-ACM sectors are shown in Figure 2(d). Since the actual number of tuples contained in these two T-ACM sectors are already known (they are $n = 24$ and $n = 36$), the next step of the algorithm applies Lemma 1, once again, to estimate the middle frequencies of these two sectors. The result is the T-ACM shown in Figure 2(e).

Before we derive any of the analytical properties of the T-ACM, it would be fitting to



(a) Equi-width Histogram (b) Trapezoidal ACM

Figure 3: Equi-width Histogram and Trapezoidal ACM.

compare the three methodologies from a common conceptual perspective.

3.2 Rationale for the Trapezoidal ACM

Without loss of generality, let us consider an arbitrary continuous frequency function $f(x)$. Figure 3 shows the histogram partitioning of $f(x)$ under the traditional equi-width method and the T-ACM method.

We note that in the equi-width case, regardless of how steep the frequency changes are in a given sector, the sector widths remain the same across the attribute value range. This means even widely different frequency values of all the different attribute values are assumed to be equal to that of the average sector frequency. Thus there is an obvious loss of accuracy with this method. On the other hand, in the equi-depth case, the area of each histogram sector is the same. This method still results in sectors with widely different frequency values and thus suffers from the same problem as the equi-width case. In the T-ACM method, we note that due to the trapezoidal nature of the sectors, the frequency approximation is assured to be closer to the actual frequencies of the attribute values. Thus the trapezoidal partitioning strategy obviously increases the estimation accuracy.

The rationale for partitioning the attribute value range using a trapezoidal approach is to minimize the variance of values in each ACM sector, and this, as we shall see, has the effect of minimizing the estimation errors. Since the variance of an arbitrary attribute value X_k is given as $Var(X_k) = E[(x_k - \mu_k)^2]$, a trapezoidal partitioning guarantees that the approximated frequencies are closer to the actual frequency values, thus ensuring that the variance of the values falls within the acceptable range. To get a flavor for the ultimate goal of our endeavor, we allude to the expression for the ACM variance which we shall derive in a subsequent lemma (Lemma 8). It will later become clear that minimizing the variance of the individual sectors will result in a lower value for the variance of the ACM. The advantages of using the T-ACM are obvious. Indeed, more detailed expressions and experimental results will later strengthen this *initial* claim.

4 Density Estimation Using Trapezoidal ACM

Consider a trapezoidal ACM sector of sector width l with n_j tuples. We assume that the tuples in this sector occur according to a trapezoidal probability distribution. In other words, the number of occurrences of the attribute values is not uniform, but it increases

(decreases) from the left most value a to the right most value b in the sector in a linear fashion as shown in Figure 4 (a).

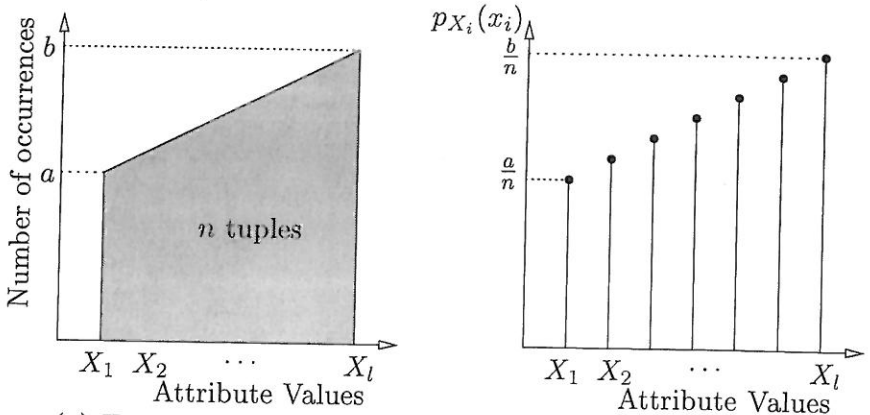


Figure 4: Trapezoidal ACM sector and its corresponding probability mass function

Since the probability of a given attribute value X_i occurring is the number of occurrences of X_i divided by the total number of tuples n_j, the probability mass function (pmf) $p_{X_i}(x_i)$ can be sketched as shown in Figure 4 (b).

Lemma 3 *The probability of a given value X_i occurring in the trapezoidal sector is,*

$$p_{X_i}(x_i) = \frac{a_j}{n_j} + \frac{2(n_j - a_j l)}{n_j l(l-1)}.i \qquad 1 \leq i \leq l-1 \tag{1}$$

where a_j is the frequency for the first attribute value in the j^{th} sector.

Proof: From the geometry of Figure 4 (b), we know that

$$\begin{aligned} p_{X_i}(x_i) &= \frac{a_j}{n_j} + \frac{b_j - a_j}{n_j(l-1)}.i \qquad 1 \leq i \leq l-1 \\ \text{But, } b_j &= \left(\frac{2n_j}{l} - a_j\right) \text{ from Lemma 1.} \\ \text{So, } p_{X_i}(x_i) &= \frac{a_j}{n_j} + \frac{2(n_j - a_j l)}{n_j l(l-1)}.i \qquad 1 \leq i \leq l-1. \end{aligned}$$

This proves the lemma. □

Lemma 4 *The probability mass distribution for the frequencies of the attribute values in a T-ACM is a Binomial distribution with parameters $(n, p_{X_i}(x_i))$.*

Proof: Consider an arbitrary permutation (or arrangement) of the n tuples in the sector. Suppose the value X_i occurs exactly x_i number of times, then all other $(l-1)$ values must occur a combined total of $(n-x_i)$ times. Since the probability of X_i occurring once is $p_{X_i}(x_i)$, where $p_{X_i}(x_i)$ is given by Lemma 3, the probability of this value not occurring is $(1-p_{X_i}(x_i)$. From hereafter, let us denote $p_{X_i}(x_i)$ simply as p_i for convenience. Hence the probability of an arbitrary permutation of the n tuples, where the value X_i occurs exactly x_i number of times is,

$$p_i^{x_i}(1-p_i)^{n-x_i}. \tag{2}$$

There are $\binom{n}{x_i}$ different permutations of the n tuples in the sector where X_i occurs exactly x_i number of times and all other values occur a combined total of $(n-x_i)$ times. Hence we find that the probability that an arbitrary value X_i occurs exactly x_i number of times is,

$$\binom{n}{x_i} p_i^{x_i}(1-p_i)^{n-x_i}. \tag{3}$$

In other words, we note that each of the attribute values $X_1, X_2, \ldots, X_l$ forms a binomial distribution *Binomial* (n, p_i) with a parameter determined by its location in the trapezoidal sector.

□

5 Maximum Likelihood Estimate Analysis for the T-ACM

In the previous section we showed that the frequency distribution of an attribute value in a T-ACM sector is a Binomial distribution whose parameters change with the location of the attribute value. With this knowledge, we shall now derive a maximum likelihood estimate for the frequency of an arbitrary attribute value in a T-ACM sector. Contrary to the classical estimation theory, where we are interested in estimating the parameters, such as the mean, of a distribution of one or more random variables, in our problem, we are interested in estimating the value of the occurrence of the random variable (the frequency x_i) which we assume is "inaccessible". We do this, however, in terms of an observation of one or more accessible random variables (i.e: the total number of tuples, n, and the slope and width of the T-ACM sector). In order to do this, we shall derive the maximum likelihood estimate, which maximizes the **corresponding** likelihood function. Indeed the result which we get is both intuitively appealing and quite easy to comprehend.

Theorem 1 *For a one-dimensional trapezoidal ACM, the maximum likelihood estimate of the number of tuples for a given value X_α of attribute X in the k^{th} T-ACM sector is given by,*

$$\hat{x}_{ML} = a_k + \frac{2(n_k - a_k l)}{l(l-1)}.z_\alpha$$

where n_k is the number of tuples in the k^{th} T-ACM sector, a_k is the frequency of the first attribute value in the k^{th} sector, l is the number of distinct attribute values (or width) of the T-ACM sectors and X_α is the z_α^{th} value in the T-ACM sector.

Proof: We know from Lemma 4 that the frequency distribution of a given attribute value in an T-ACM sector is a Binomial distribution. So the probability mass function of the frequency distribution of an attribute value $X = X_\alpha$ in an T-ACM sector can be written as,

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where x is the number of occurrences of X_α. Let

$$\mathcal{L}(x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

$\mathcal{L}(x)$ is the traditional *likelihood function* of the random variable X on the parameter x which we intend to maximize. We are interested in finding out the maximum likelihood estimate for this parameter x. Taking natural logarithm on both sides of the likelihood function, we have,

$$\begin{aligned} \ln \mathcal{L}(x) &= \ln n! - \ln x! - \ln(n-x)! + x \ln p + (n-x)\ln(1-p) \\ &= \ln \Gamma(n+1) - \ln \Gamma(x+1) - \ln \Gamma(n-x+1) + \\ &\quad x \ln p + (n-x)\ln(1-p) \end{aligned} \tag{4}$$

where $\Gamma(x)$ is the Gamma function given by, $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$. Now using the well known identity,

$$\Gamma(\alpha) = \frac{\Gamma(\alpha+k+1)}{\alpha(\alpha+1)\ldots(\alpha+k)}$$

we find that,

$$\Gamma(n-x+1) = \frac{\Gamma(n+1)}{(n-x+1)(n-x+2)\cdots n} \text{ and } \Gamma(x+1) = \frac{\Gamma(n+1)}{(x+1)(x+2)\cdots n}.$$

Thus substituting the above expressions for $\Gamma(n-x+1)$ and $\Gamma(x+1)$ in Equation 4, we find,

$$\begin{aligned} \ln \mathcal{L}(x) &= -\ln \Gamma(n+1) + x \ln p + (n-x)\ln(1-p) + \\ &\quad \ln(x+1) + \ln(x+2) + \ldots + \ln n + \\ &\quad \ln(n-x+1) + \ln(n-x+2) + \ldots + \ln n \end{aligned}$$

Now differentiating $\ln \mathcal{L}(x)$ with respect to x, we obtain,

$$\frac{d}{dx} \ln \mathcal{L}(x) = \ln p - \ln(1-p) + \sum_{r=x+1}^{n-x} \frac{1}{r}.$$

Setting $\frac{d\{\mathcal{L}(x)\}}{dx} = 0$, and noting that $\sum_{r=x+1}^{n-x} \frac{1}{r} \leq \ln\left(\frac{n-x}{x}\right)$, $\hat{x}_{ML}$ of x is obtained as,

$$\frac{p(n-x)}{(1-p)x} \geq 1.$$

This inequality is solved for $x \leq np$. But, by virtue of the underlying distribution, since we know that the likelihood function monotonically increases till its maximum, we conclude that,

$$\hat{x}_{ML} = np.$$

But we have already seen in Lemma 3 that, the probability of the z_α^{th} attribute value, X_α, occurring in a T-ACM sector is given by,

$$p_{X_\alpha}(x_\alpha) = \frac{a_k}{n_k} + \frac{2(n_k - a_k l)}{n_k l(l-1)}.z_\alpha$$

where a_k is the frequency of the first attribute value in the k^{th} sector. So we have,

$$\hat{x}_{ML} = a_k + \frac{2(n_k - a_k l)}{l(l-1)}.z_\alpha.$$

Hence the theorem. □

In most of the cases, the maximum likelihood estimate, $\hat{x}_{ML} = np$, which we derived using the Gamma function above is not an integer. In fact, the maximum likelihood estimate reaches its upper limit of np at integer values only in very special cases. If we are interested in the integer maximum likelihood value which is related to the above maximum likelihood estimate, we have to discretize the space. Thus, considering the analogous discrete case, we have the following theorem.

Theorem 2 *For a one-dimensional trapezoidal ACM, the maximum likelihood estimate of the number of tuples for a given value X_i of attribute X falls within the range of,*

$$\frac{a_k}{n_k} + \frac{2(n_k - a_k l)}{n_k l(l-1)}.z_\alpha(n_k+1) - 1 \leq \hat{x}_{ML} \leq \frac{a_k}{n_k} + \frac{2(n_k - a_k l)}{n_k l(l-1)}.z_\alpha(n_k+1),$$

where a_k is the frequency of the first attribute value in the k^{th} sector, n_k is the number of tuples in the k^{th} sector containing the value X_i and l is the width of that sector.

Proof: The probability mass function $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$ is a steadily increasing function until it reaches the maximum likelihood value, $x = \hat{x}_{ML}$. For any $x > \hat{x}_{ML}$, $f(x)$ is a steadily decreasing function. Hence we can obtain an integer value for the maximum likelihood estimate by solving the following two discrete inequalities simultaneously.

$$f(x) - f(x+1) > 0 \quad (5)$$
$$f(x) - f(x-1) > 0 \quad (6)$$

From Equation (5), we have,

$$\begin{aligned} f(x) &- f(x+1) > 0 \\ \binom{n}{x} p^x (1-p)^{n-x} &- \binom{n}{x+1} p^{x+1} (1-p)^{n-x-1} > 0 \\ \frac{n!}{x!(n-x)!}(1-p) &- \frac{n!}{(x+1)!(n-x-1)!} p > 0 \\ \frac{1-p}{n-x} &- \frac{p}{x+1} > 0 \text{ or} \\ x &> p(n+1) - 1. \end{aligned}$$

Similarly considering Equation (6), by using similar algebraic manipulation, we get,

$$\begin{aligned} f(x) &- f(x-1) > 0 \\ \binom{n}{x} p^x (1-p)^{n-x} &- \binom{n}{x-1} p^{x-1} (1-p)^{n-x-1} > 0 \text{ or} \\ x &< p(n+1). \end{aligned}$$

Hence,

$$p(n+1) - 1 < x < p(n+1).$$

Since $p = \frac{a_k}{n_k} + \frac{2(n_k - a_k l)}{n_k l(l-1)}.z_\alpha$, the theorem follows. □

6 Expected and Worst-Case Error Analysis for the T-ACM

The maximum likelihood estimation of the frequency of an attribute value tells us that the attribute value would have a frequency of $\hat{x}_{ML}$ with high degree of certainty when compared to the other possible frequency values. But even though the attribute value occurs with the maximum likelihood frequency with high probability, it can also occur with other frequencies with smaller probabilities. Hence, as we did in the case of the R-ACM, when we need to find the worst-case and average-case errors for the result size estimations, we need to obtain the expected value of the frequency of a given attribute value. We use our Binomial model for the T-ACM sector to find the expected value of the frequency of an attribute value as given in the following lemma and develop a series of results regarding the corresponding query result-size estimates.

Lemma 5 *Using a trapezoidal approximation, the expected number of tuples for a given value X_i of attribute X is,*

$$E(X_i) = a_j + \frac{2(n_j - a_j l)}{l(l-1)}.i,$$

where n_j is the number of tuples in the sector which contains value X_i and l is the sector width. The quantity a_j is the number of occurrences of the first attribute value in the j^{th} sector.

Proof: The frequency distribution of attribute values in a T-ACM sector is a binomial distribution with parameters (n, p_i) where p_i is given by Lemma 3. Hence the expected value $E(X_i)$ is its mean,

$$E(X_i) = n_j p_i = a_j + \frac{2(n_j - a_j l)}{l(l-1)}.i$$

and the lemma follows. □

6.1 Estimation Error with the Trapezoidal ACM

It has been shown that even a small error in the estimation results, when propagated through several intermediate relational operations, can become exponential and be devastating to the performance of a DBMS [5]. In this section we provide some definitions for estimating the errors based on the variance, and provide a technique to measure the estimation errors obtained from the T-ACM.

The variance of a random variable X measures the spread or dispersion that the values of X can assume and is defined by $Var(X) = E\{[X - E(X)]^2\}$. It is well known that $Var(X) = E(X^2) - [E(X)]^2$. Thus the variance of the frequency of the k^{th} value of the attribute X in the j^{th} sector is given as,

$$Var(X_k) = E\left[\left(x_k - \frac{n_j}{l_j}\right)^2\right]$$

Expanding the right hand side, we obtain,

$$Var(X_k) = \sum_{i=0}^{n_j} x_k^2 \left(\frac{1}{l_j}\right)^i \left(1 - \frac{1}{l_j}\right)^{n_j - i} - \left(\frac{n_j}{l_j}\right)^2 \tag{7}$$

Lemma 6 *The variance of the frequency of an attribute value X_i in sector j of a trapezoidal ACM is,*

$$Var(X_i) = n_j p_i (1 - p_i), \text{ where } p_i = \frac{a_j}{n_j} + \frac{2(n_j - a_j l)}{n_j l(l-1)}.i$$

Proof: The frequency distribution in a T-ACM sector is a binomial distribution with parameters (n_j, p_i), where p_i is given by Lemma 3. Hence the variance is $n_j p_i (1 - p_i)$. □

Lemma 7 *The sector variance of the j^{th} trapezoidal ACM sector is,*

$$Var_j = n_j - \frac{a_j(l+1)(a_j l - 2n_j)}{3n_j(l-1)} - \frac{2n_j(2l-1)}{3l(l-1)}$$

where a_j is the frequency of the first attribute value in the sector, n_j is the number of tuples in the sector and l is the sector width.

Proof: Since the frequency values in the sector are assumed independent, summing up the variances of all frequency values in the sector will give us the expression for the variance of the entire sector. So we have,

$$\begin{aligned} Var_j &= \sum_{i=0}^{l-1} n_j p_i (1 - p_i) \\ &= \sum_{i=0}^{l-1} n_j \left(\frac{a_j}{n_j} + \frac{2(n_j - a_j l)}{n_j l(l-1)} . i \right) \left(1 - \frac{a_j}{n_j} - \frac{2(n_j - a_j l)}{n_j l(l-1)} . i \right). \end{aligned}$$

Simplifying the above expression gives us,

$$Var_j = n_j - \frac{a_j(l+1)(a_j l - 2n_j)}{3n_j(l-1)} - \frac{2n_j(2l-1)}{3l(l-1)}$$

and the lemma follows. □

Lemma 8 *The variance of a T-ACM is given by,*

$$Var(ACM) = \sum_{j=1}^{s} Var_j$$

where s is the number of sectors in the T-ACM, and Var_j is the sector variance given in Lemma 7.

Proof: The lemma follows directly from the fact that the frequency values in each sector are independent of each other and thus summing up the variances of all the sectors will give the overall variance which is also an estimate for the estimation error. □

6.2 Self-join Error with the Trapezoidal ACM

It is interesting to study the join estimation when a relation is joined with itself. These self-joins frequently occur with 2-way join queries. It is well known [10] that the self-join is a case where the query result size is maximized because the highest occurrences (frequencies) in the joining attributes correspond to the same attribute values. Assuming that the duplicate tuples after the join are not eliminated, we have the following lemma.

Lemma 9 *The error, ϵ, resulting from a self-join of relation R on attribute X using a trapezoidal ACM is given by,*

$$\epsilon = \sum_{j=1}^{s} \left(\sum_{k=1}^{l} x_k^2 - n_j^2 + n_j Var_j \right)$$

where s is the number of sectors in the T-ACM, and n_j is the number of tuples in the j^{th} sector and Var_j is the variance of the j^{th} sector given in Lemma 7.

Proof: Since there are $L = sl$ distinct values for attribute X, the actual value, ξ and expected value κ of the join size can be estimated as follows.

$$\xi = \sum_{i=1}^{L} x_i^2 = \sum_{j=1}^{s} \sum_{k=1}^{l} x_k^2.$$

The frequency of an arbitrary attribute value is computed from the T-ACM as the expected value $E(x_i)$, which is the average frequency of the T-ACM sector. Hence the result of self-joining this attribute value would be $[E(x_i)]^2$. Hence the size of the join computed by the T-ACM, κ, is,

$$\begin{aligned} \kappa &= \sum_{i=1}^{L} [E(x_i)]^2 = \sum_{j=1}^{s} \sum_{i=0}^{l-1} [E(x_i)]^2 \\ &= \sum_{j=1}^{s} n_j^2 \sum_{i=0}^{l-1} p_i^2. \end{aligned} \tag{8}$$

But the variance of the j^{th} sector Var_j is,

$$\begin{aligned} Var_j &= \sum_{i=0}^{l-1} n_j p_i (1 - p_i) = n_j - n_j \sum_{i=0}^{l-1} p_i^2 \\ \text{So, } \sum_{i=0}^{l-1} p_i^2 &= 1 - \frac{Var_j}{n_j}. \end{aligned}$$

Substituting the above expression in Equation (8), we obtain,

$$\kappa = \sum_{j=1}^{s} (n_j^2 - n_j Var_j).$$

Hence the error in estimation of self-join is,

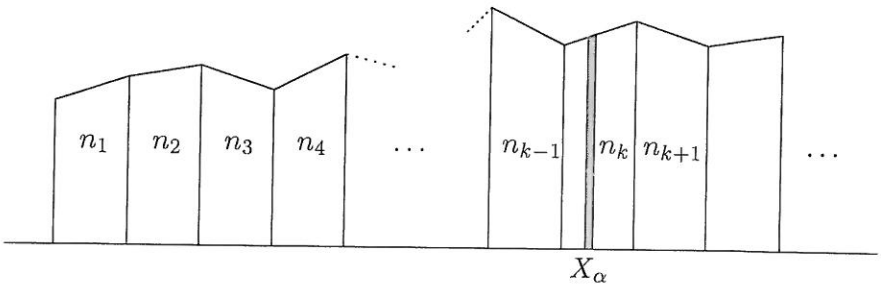
$$\begin{aligned} \xi - \kappa &= \sum_{j=1}^{s} \sum_{k=1}^{l} x_k^2 - \sum_{j=1}^{s} (n_j^2 - n_j Var_j) \\ &= \sum_{j=1}^{s} \left(\sum_{k=1}^{l} x_k^2 - n_j^2 + n_j Var_j \right) \end{aligned}$$

and the lemma is proved. □

Corollary 1 *The error, ϵ, resulting from a self-join of relation R on attribute X using a trapezoidal ACM is given by,*

$$\epsilon = \sum_{j=1}^{s} \left(\sum_{k=1}^{l} x_k^2 - \frac{a_j(l+1)(a_j l - 2n_j)}{3(l-1)} - \frac{2n_j^2(2l-1)}{3l(l-1)} \right)$$

where a_j is the frequency of the first attribute value in the j^{th} sector.



Equality Select with T-ACM

Figure 5: Equality Select Using the T-ACM

Proof: The proof follows from substituting Var_j in the previous lemma with the expression obtained in Lemma 7. □

6.3 Worst-Case Error with the Trapezoidal ACM

Having derived the maximum likelihood estimate for the frequency of an attribute value within a T-ACM sector, we shall now estimate the worst-case error with the T-ACM for both an equality-select query and a range-select query.

Let us consider a relational query with an equality select predicate, $X = X_\alpha$, where X_α is a constant value in the domain of attribute X. We are interested in finding the worst-case error in estimating the result size of the query, $\sigma_{X=X_\alpha}(R)$, where the attribute value X_α is in position α of the k^{th} T-ACM sector. The following lemma gives us this estimate.

Theorem 3 *The worst-case error, ϵ, in estimating the equality select operation, $\sigma_{X=X_\alpha}(R)$ in a T-ACM using the maximum likelihood estimate is given by,*

$$\epsilon = \begin{cases} a_k + \frac{2(n_k - a_k l)}{l(l-1)} z_\alpha & \text{if } z_\alpha < \frac{l(l-1)(n_k - 2a_k)}{4(n_k - a_k l)}, \\ n_k - a_k - \frac{2(n_k - a_k l)}{l(l-1)} z_\alpha & \text{if } z_\alpha \geq \frac{l(l-1)(n_k - 2a_k)}{4(n_k - a_k l)}. \end{cases}$$

where the attribute value X_α is in the z_α^{th} position within the k^{th} T-ACM sector.

Proof: The expected frequency of the attribute value X_α reported as a result of the T-ACM is,

$$\hat{\xi} = n_k \left(\frac{a_k}{n_k} + \frac{2(n_k - a_k l)}{n_k l(l-1)} z_\alpha \right)$$

where $\left(\frac{a_k}{n_k} + \frac{2(n_k - a_k l)}{n_k l(l-1)} z_\alpha \right)$ is the probability that attribute value X_α occurs in the T-ACM sector. But the actual frequency ξ of attribute value X_α can be anywhere in the range of,

$$0 \leq \xi \leq n_k.$$

Hence the maximum worst case error is,

$$\epsilon = \max(\hat{\xi}, n_k - \hat{\xi}).$$

It is easy to observe that whenever $\hat{\xi} < \frac{n_k}{2}$, the maximum worst case error occurs when the actual frequency ξ is equal to n_k. The maximum worst case error in this case is $n_k - \hat{\xi}$. Whereas whenever $\hat{\xi} \geq \frac{n_k}{2}$, the maximum worst case error occurs when the actual frequency $\xi = 0$. The maximum worst case error in this case is of course $\hat{\xi}$. We note that whether the expected frequency value $\hat{\xi}$ is smaller or larger than $\frac{n_k}{2}$ depends on the location of the attribute value X_α within the T-ACM sector. The location of the attribute value when the expected frequency $\hat{\xi}$ is equal to $\frac{n_k}{2}$ can be obtained by solving,

$$a_k + \frac{2(n_k - a_k l)}{l(l-1)} z_\alpha = \frac{n_k}{2}$$

and is equal to,

$$z_\alpha = \frac{l(l-1)(n_k - 2a_k)}{4(n_k - a_k l)}.$$

The theorem follows from the above. □

Unlike the previous case, when estimating the result size of a range select query, we have to consider three distinct cases. These are namely the cases when,

1. The attribute value range spans across one T-ACM sector.
2. The attribute value range falls completely within one T-ACM sector.
3. The attribute value range spans across more than one T-ACM sector.

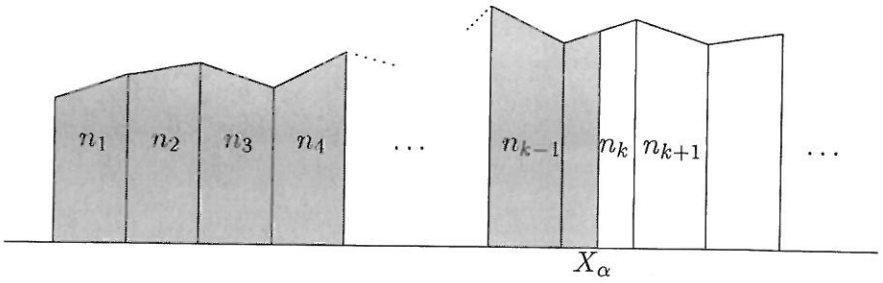
In the first case, estimation using the T-ACM gives the accurate result (n_j) and there is no estimation error. The estimation error in the second case is given by the theorem below. The estimation error in the third case can be obtained by noting that it is in fact the combination of the first and second cases.

Theorem 4 *The worst-case error in estimating the result size of the range-selection query, $\sigma_{X_\alpha \leq X \leq X_\beta}(R)$, where the attribute values X_α and X_β fall completely within the k^{th} T-ACM sector is given by,*

$$\epsilon = \begin{cases} n_k - \mathcal{A} & \text{if } \mathcal{A} < \frac{n_k}{2}, \\ \mathcal{A} & \text{if } \mathcal{A} \geq \frac{n_k}{2} \end{cases}$$

where $\mathcal{A}$ is the expected number of tuples between the attribute values X_α and X_β and is equal to,

$$\mathcal{A} = a_k(\beta - \alpha + 1) + \frac{(n_k - a_k l)(\beta - \alpha + 1)(\beta - \alpha + 2)}{l(l-1)}.$$



Range Select with T-ACM (Shaded region is the result of select)

Figure 6: Result Estimation of Range Select Using the T-ACM

Proof: The sum of the expected frequencies between the attribute values X_α and X_β within a T-ACM sector is,

$$\begin{aligned}\mathcal{A} &= \sum_{i=\alpha}^{\beta} \hat{\xi} = \sum_{i=\alpha}^{\beta} \left(a_k + \frac{2(n_k - a_k l)}{l(l-1)} i \right) \\ &= a_k(\beta - \alpha + 1) + \frac{(n_k - a_k l)(\beta - \alpha + 1)(\beta - \alpha + 2)}{l(l-1)}.\end{aligned}$$

But the actual sum of frequencies ξ between the attribute values X_α and X_β can be anywhere in the range of,

$$0 \leq \xi \leq n_k.$$

Hence the maximum worst case error is,

$$\epsilon = \max(\mathcal{A}, n_k - \mathcal{A}).$$

It is easy to observe that whenever $\mathcal{A} < \frac{n_k}{2}$, the maximum worst case error occurs when the actual sum of frequencies, ξ, is equal to n_k. The maximum worst case error in this case is $n_k - \mathcal{A}$. Whereas whenever $\mathcal{A} \geq \frac{n_k}{2}$, the maximum worst case error occurs when the actual sum of frequencies, $\xi = 0$. The maximum worst case error in this case is of course $\mathcal{A}$. Hence the theorem. □

6.4 Average-Case Error with Trapezoidal ACM

In this section we give an estimate for the average-case error with a trapezoidal ACM. As we shall see, the average case error is much smaller than the worst-case error that we derived in the previous section. The average-case is synonymous with a truly random sector in which all attribute values have the same (or approximately same) frequency value equal to the mean sector frequency, $\frac{n_j}{l}$. The average case error in estimating the frequency of an arbitrary value X_i can be obtained by two different methods, depending on how the

frequencies of the attribute values are averaged. In the first case, the expected frequency of all attribute values in the sector is obtained by averaging over the entire sector. In the second case, we obtain the average frequency of *an attribute* value by averaging all the possible frequencies that this particular attribute value can assume. The average case errors in these two situations are given below in Theorems 5 and 6.

Theorem 5 *Assuming that the T-ACM sector has been obtained by processing a histogram bucket of size l with n_j tuples, the average error in estimating the result size of the equality selection query, $\sigma_{X=X_i}(R)$, obtained by averaging over* all *attribute values in this sector of the trapezoidal ACM is exactly zero.*

Proof: The expected frequency of the attribute values in the sector computed by the T-ACM can be obtained by averaging **over the entire sector** as,

$$\begin{aligned} E(\hat{\xi}_i) &= \frac{1}{l}\sum_{i=0}^{l-1}\hat{\xi}_i \\ &= \frac{1}{l}\sum_{i=0}^{l-1} n_j\left(\frac{a_j}{n_j} + \frac{2(n_j - a_j l)}{n_j l(l-1)}i\right) = \frac{n_j}{l} \end{aligned}$$

where $\left(\frac{a_j}{n_j} + \frac{2(n_j - a_j l)}{n_j l(l-1)}i\right)$ is the probability that attribute value X_i occurs in the T-ACM sector. But, if we assume that the T-ACM sector has been obtained by processing an equivalent histogram bucket of size l with n_j tuples, then the actual frequency ξ_i of attribute value X_i in the average case is equal to, $\xi_i = \frac{n_j}{l}$. Hence the average case error obtained by averaging **over the entire range** is equal to,

$$\epsilon = \hat{\xi}_i - \xi_i = \frac{n_j}{l} - \frac{n_j}{l} = 0.$$

The theorem follows. □

Note that in the case of an R-ACM, the actual frequencies of the attribute values are controlled by the tolerance, τ. That is why in the R-ACM (unlike the T-ACM), the average case error is not equal to zero. In a T-ACM, due to the geometry of the T-ACM sector, each of the negative estimation errors in the right half of the sector cancels out with the corresponding positive estimation error on the left half of the sector, thus resulting in an overall zero average case error **when the expectation operation is carried out by averaging over the entire sector**. Note that this will not be the case if we perform the expectation at any one particular attribute value, and this is discussed in the theorem below.

Theorem 6 *The upper bound of the average-case error, ϵ, in estimating the result size of an equality select query, $\sigma_{X=X_i}(R)$, using a trapezoidal ACM is,*

$$\epsilon = a_k + \frac{2(n_k - a_k l)}{l(l-1)}.i - \frac{n_k}{l},$$

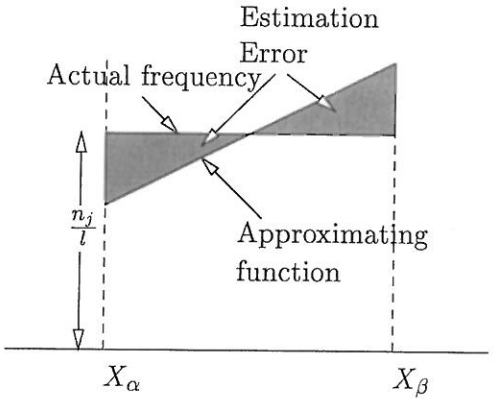


Figure 7: Average Case Error in T-ACM

where a_k is the frequency of the first attribute value in the k^{th} sector and X_i is in the i^{th} position of the T-ACM sector.

Proof: The expected frequency of the attribute value computed by the T-ACM is,

$$\hat{\xi} = n_k \left(\frac{a_k}{n_k} + \frac{2(n_k - a_k l)}{n_k l(l-1)} i \right)$$

where $\left(\frac{a_k}{n_k} + \frac{2(n_k - a_k l)}{n_k l(l-1)} i \right)$ is the probability that attribute value X_i occurs in the T-ACM sector. But assuming that the T-ACM sector has been obtained from an equivalent histogram bucket of size l with n_k tuples, we note that, due to the uniformity assumption within the histogram bucket, the frequency ξ_i of attribute value X_i in this histogram bucket is equal to, $\xi_i = \frac{n_k}{l}$. Hence the average-case error is equal to,

$$\epsilon = \hat{\xi} - \xi = a_k + \frac{2(n_k - a_k l)}{l(l-1)} . i - \frac{n_k}{l}.$$

The theorem follows. □

As before in the worst-case error analysis, when estimating the sum of frequencies in an attribute value range, we have three distinct cases. The theorem given below deals with the case of the average-case error when attribute value range falls completely within one T-ACM sector. The case when the attribute value range spans across one entire T-ACM sector is trivial and does not result in any estimation error. The case when the attribute value range spans across more than one T-ACM sector can be solved by decomposing it into the first two cases.

Theorem 7 *The average-case error in estimating the result size of the selection query, $\sigma_{X_\alpha \leq X \leq X_\beta}(R)$, where the attribute values X_α and X_β fall completely within the k^{th} T-ACM*

sector is given by,

$$\epsilon = \frac{(\beta - \alpha)(\alpha + \beta - 3)(n_k - a_k l)}{l - 1}$$

where $\beta > \alpha$.

Proof: In a random T-ACM sector, all the frequency values are equal (or close) to the mean frequency value. This is shown in Figure 7 along with the T-ACM frequency distribution that is used to approximate the actual frequency distribution. We note that the shaded area between the actual and approximate frequency distribution represents the cumulative estimation error. Also we see that both lines intersect at the center of the sector or at $i = \frac{l-1}{2}$. Hence an estimate for the estimation error is,

$$\begin{aligned}
\epsilon &= \frac{n_k}{l}\left(\frac{l-1}{2} - \alpha + 1\right) - \sum_{i=\alpha}^{\frac{l-1}{2}} \left(a_k + \frac{2(n_k - a_k l)}{l(l-1)}\right) i \\
&\quad + \sum_{i=\frac{l-1}{2}}^{\beta} \left(a_k + \frac{2(n_k - a_k l)}{l(l-1)}.i\right) - \frac{n_k}{l}\left(\frac{l-1}{2} - \beta + 1\right) \\
&= \frac{(\beta - \alpha)(\alpha + \beta - 3)(n_k - a_k l)}{l - 1}
\end{aligned}$$

and the theorem follows. □

6.5 Estimation of Join Error

The estimation error resulting from an equality join of two attributes is usually much higher than the estimation errors resulting from the equality select and range select operations.

Lemma 10 *Considering the equality join of two domain compatible attributes* X *and* Y *with* $X_i = Y_j$, *if the expected result size of the equality selection query,* $\sigma_{X=X_i}$, *using an ACM is* $\hat{x}_i$ *and that of* $\sigma_{Y=Y_j}$ *is* $\hat{y}_j$, *then the maximum error resulting from joining the attributes* X *and* Y *on the values* X_i *and* Y_j *is given by,*

$$\epsilon = |(\hat{x}_i \epsilon_y + \hat{y}_j \epsilon_x + \epsilon_x \epsilon_y)|$$

where ϵ_x *and* ϵ_y *are the estimated errors resulting from the equality selection queries* $\sigma_{X=X_i}$ *and* $\sigma_{X=Y_j}$ *respectively.*

Proof: Assume that the actual frequency values of X_i and Y_j are x_i and y_j respectively. Hence the actual size of the join contribution from these values is,

$$\xi = x_i y_j.$$

But the expected size of the join contribution from the above values is,

$$\hat{\xi} = \hat{x}_i \hat{y}_j.$$

Thus the maximum error resulting from joining the values $X = X_i$ and $Y = Y_j$ is,

$$\begin{aligned} \epsilon &= |\xi - \hat{\xi}| \\ &= |x_i y_j - \hat{x}_i \hat{y}_j| \end{aligned}$$

The possible values for x_i can be either $(\hat{x}_i - \epsilon_x)$ or $(\hat{x}_i + \epsilon_x)$. Similarly the possible values for y_j can be either $(\hat{y}_j - \epsilon_y)$ or $(\hat{y}_j + \epsilon_y)$. We note that out of the 4 possible value combinations of these expected values, only $(\hat{x}_i + \epsilon_x)(\hat{y}_j + \epsilon_y)$ gives the largest error. Hence the maximum error becomes,

$$\begin{aligned} \epsilon &= |\hat{x}_i \hat{y}_j - (\hat{x}_i + \epsilon_x)(\hat{y}_j + \epsilon_y)| \\ &= |\hat{x}_i \epsilon_y + \hat{y}_j \epsilon_x + \epsilon_x \epsilon_y|. \end{aligned}$$

The lemma follows. □

Considering all the values of attributes X and Y, it is possible to find the cumulative error in the estimation of a join. Hence using the results on estimation errors we obtained earlier, we can find the join errors for both the worst-case and average-case situations in the T-ACM.

Corollary 2 *The error resulting from an equality join of two domain compatible attributes X and Y, is given by,*

$$\epsilon = \sum_{j=1}^{s_X} \sum_{i=1}^{l_j} (\hat{x}_i \epsilon_{y_k} + \hat{y}_k \epsilon_{x_i} + \epsilon_{x_i} \epsilon_{y_k})$$

where k is an index on the T-ACM of Y such that $X_i = Y_k$ and $\epsilon_{x_i}, \epsilon_{y_k}$ are the errors resulting from the equality selection queries $\sigma_{X=X_i}$ and $\sigma_{Y=Y_k}$ respectively.

Proof: The proof follows from the previous lemma. □

7 Comparison of Trapezoidal ACM and Equi-Width Histogram

The rationale for the trapezoidal ACM is that the trapezoidal rule for numerical integration provides a more accurate estimation of the area under a curve than the left-end or right-end histogram (or rectangular rule) based methods. This is formally given by the following lemma.

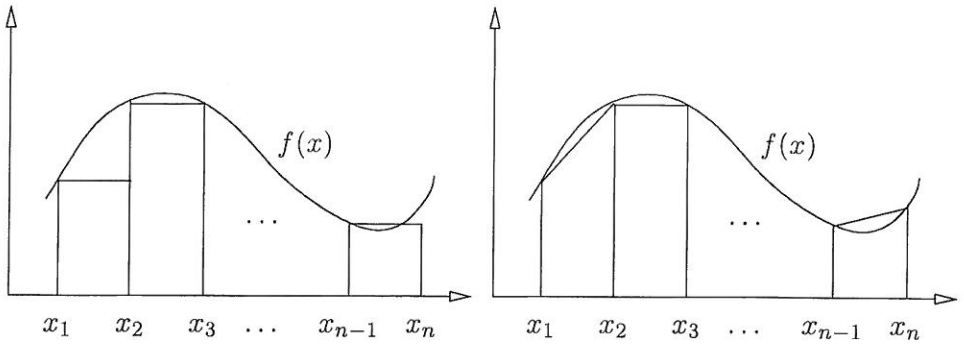


Figure 8: Comparison of Histogram and Trapezoidal ACM

Lemma 11 *If the estimation error for computing the sum of frequencies of all the attribute values falling between $X = a$ and $X = b$, using the trapezoidal ACM is $Error_T$ and that of using the histogram method is $Error_H$, then $Error_T < Error_H$.*

Proof: Without loss of generality, let us consider a continuous function $f(x)$. (See Figure 8.) Let A^* and A^{**} be the approximations to the area under the function $f(x)$ between $x = a$ and $x = b$ by these two methods respectively. Also let ϵ_1 and ϵ_2 be the errors introduced by the trapezoidal ACM and histogram methods in the estimated areas A^* and A^{**} respectively. Hence,

$$\begin{aligned} \epsilon_1 &= A^* - \int_a^b f(x)dx \\ \text{and } \epsilon_2 &= A^{**} - \int_a^b f(x)dx. \end{aligned}$$

The histogram method, also known as the rectangular rule, and the trapezoidal ACM, also known as the trapezoidal rule are two well known numerical integration techniques to approximate the area under a curve. It can be shown [8] that using the trapezoidal rule, the estimation error, ϵ_1, has the following bounds.

$$\frac{(b-a)^3}{12n^2}M_2^* \leq \epsilon_1 \leq \frac{(b-a)^3}{12n^2}M_2.$$

Similarly it can be shown that using the rectangular rule, the estimation error, ϵ_2, has the following bounds.

$$\frac{(b-a)^3}{6n^2}M_2^* + \frac{b^2}{2n}M_1^* \leq \epsilon_2 \leq \frac{(b-a)^3}{6n^2}M_2 + \frac{b^2}{2n}M_1.$$

In both bounds, the quantities M_1^*, M_1 are the smallest and largest values for the first derivative of f and M_2^*, M_2 are the smallest and largest values for the second derivative of f between $x = a$ and $x = b$. Hence it follows that $Error_T < Error_H$. □

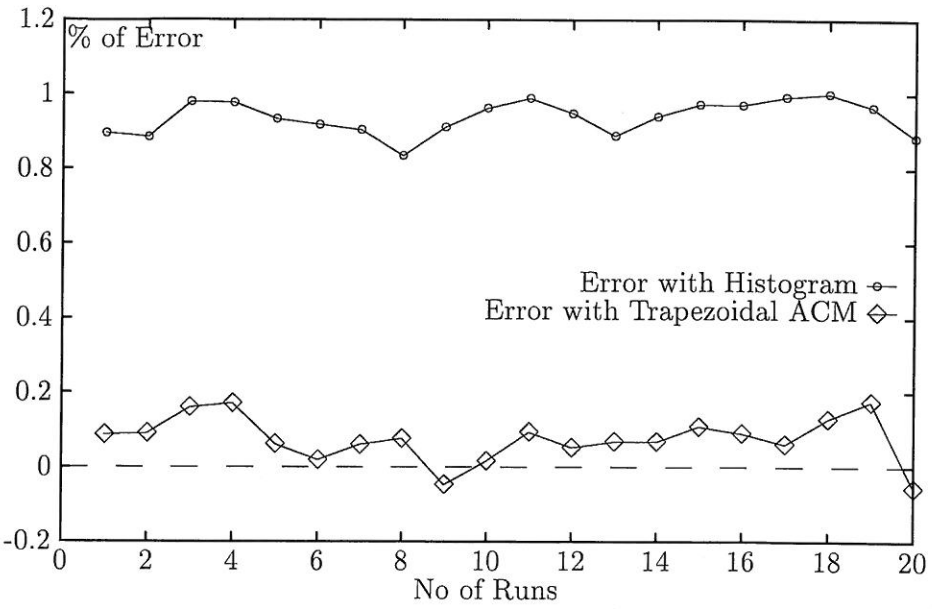


Figure 9: Comparison of Histogram and the T-ACM for Probability Estimation: Each experiment was run 100,000 times to get the average percentage of errors in the estimated occurrence of the attribute values. Estimation errors are given for exact match on a random distribution with 100,000 tuples and 1000 distinct values. Both histogram and T-ACM were of equi-width type with a sector width of 5 and no of sectors equal to 200.

Claim 1 *If the frequency estimation error for an arbitrary attribute value using the trapezoidal ACM is* $Error_T$ *and that of using the equi-width histogram with the same number of sectors is* $Error_H$*, then* $Error_T < Error_H$.

Rationale: Assume the actual frequency of an arbitrary attribute value X_i is ξ. Let the frequencies computed by an equi-width histogram and a T-ACM with the same number of sectors be $\hat{\xi}_H$ and $\hat{\xi}_T$ respectively. We use $E[(\xi - \hat{\xi})^2]$ as the measure for comparing the errors resulting from the histogram and the T-ACM. So we have,

$$\begin{aligned} E_H[(\xi - \hat{\xi}_H)^2] &= E_H\left[\left(\xi - \frac{n}{l}\right)^2\right] \\ &= E(\xi^2) - \left(\frac{n}{l}\right)^2 \\ &= \frac{\sum_{k=0}^{l-1}\sum_{i=0}^{n}\binom{n}{x_i}x_i^2 p_H^{x_i}(1-p_H)^{n-x_i}}{l} - \left(\frac{n}{l}\right)^2 \end{aligned}$$

Histogram Type	Worst-case Error	Average-case Error
Equi-width	$\max\left(n_j - \frac{n_j}{l}, \frac{n_j}{l}\right)$	$\max\left(n_j - \frac{n_j}{l}, \frac{n_j}{l}\right)$
Equi-depth	$\frac{2}{3l_j}$	$\frac{1}{2l_j}$
T-ACM	$\max\left(a_j + \frac{2(n_j - a_j l)}{l(l-1)}i, n_j - a_j - \frac{2(n_j - a_j l)}{l(l-1)}i\right)$	$a_j + \frac{2(n_j - a_j l)}{l(l-1)}i - \frac{n_j}{l}$

Table 3: Comparison of Histogram Errors

where p_H is the probability of selecting one of the l attribute values in the histogram sector and is equal to $\frac{1}{l}$. Similarly,

$$\begin{aligned} E_T[(\xi - \hat{\xi}_T)^2] &= E(\xi^2) - [E(\hat{\xi}_T)]^2 \\ &= \frac{\sum_{k=0}^{l-1}\sum_{i=0}^{n}\binom{n}{x_i}x_i^2 p_{T_k}^{x_i}(1-p_{T_k})^{n-x_i}}{l} - \left(\frac{\sum_{k=0}^{l-1} a + \frac{2(n-al)}{l(l-1)}.k}{l}\right)^2 \\ &= \frac{\sum_{k=0}^{l-1}\sum_{i=0}^{n}\binom{n}{x_i}x_i^2 p_{T_k}^{x_i}(1-p_{T_k})^{n-x_i}}{l} - \left(\frac{n}{l}\right)^2 \end{aligned}$$

where p_{T_k} is the probability of selecting the k^{th} attribute value in a trapezoidal sector and is equal to $\frac{a}{n} + \frac{2(n-al)}{nl(l-1)}.k$.

Analytically proving that

$$\sum_{k=0}^{l-1}\sum_{i=0}^{n}\binom{n}{x_i}x_i^2 p_H^{x_i}(1-p_H)^{n-x_i} > \sum_{k=0}^{l-1}\sum_{i=0}^{n}\binom{n}{x_i}x_i^2 p_{T_k}^{x_i}(1-p_{T_k})^{n-x_i}$$

is difficult. But extensive simulations demonstrate that this is true, and we intend to use a symbolic mathematical package such as Mathematica or Maple to show this to be true. However we believe it is true due to the well acclaimed superiority of the trapezoidal rule over the rectangular rule in numerical integration, indicating that $E_H[(\xi - \hat{\xi}_H)^2] > E_T[(\xi - \hat{\xi}_T)^2]$. □

We compare the worst-case and average-case estimation errors of the T-ACM to the traditional equi-width, equi-depth histograms and the R-ACM technique proposed in [12] in Table 3.

8 Experimental Results

We conducted an extensive array of experiments, using a number of real-world databases, to compare the performance of the T-ACM to the traditional equi-width and equi-depth histograms currently being used by most commercial database systems. Tables 4 and 5

Operation	Actual Size	Equi-width		Equi-depth		T-ACM	
		Size	Error	Size	Error	Size	Error
Equi-select	1721	1250.13	27.36%	1292.30	24.91%	1788.29	3.91%
Range-select	32073	29698	7.40%	30526	4.82%	32614	1.69%
Equi-join	795280	518443	34.81%	580475	27.01%	877034	10.28%

Table 4: Comparison of Equi-width, Equi-depth and T-ACM: U.S. CENSUS Database.

show the experimental results conducted on some of the data set generated from the U.S. CENSUS database[1][16]. Since it is impossible to list all of our experimental results here, we refer the reader to [15] for more detailed information. In the interest of completeness, however, (and to aid other researchers to duplicate our results) we would like to emphasize that the input to the algorithms was the database itself, but the output was the T-ACM obtained after invoking the algorithms `Generate_T-ACM` and `Implement_T-ACM` from Section 3.1.

The queries for our experiments consisted of either (a) equality join (b) equality selection or (c) range selection operators.

The first group of experiments were conducted on equi-width, equi-depth histograms and the T-ACM. In each of our experimental runs, we chose different build-parameters for the histograms and the T-ACM. The build-parameter for the equi-width histogram and the T-ACM is the sector width whereas the build-parameter for the equi-depth histogram is the number of tuples in the sector. The histograms and the T-ACM were constructed for some selected attributes from the U.S. CENSUS database.

We obtained the relative estimation error as a ratio by subtracting the estimated size from the actual result size and dividing it by the actual result size. Obviously, the cases where the actual result sizes were zero were not considered for error estimation. We implemented a simple query processor to compute the actual result size. The results were obtained by averaging the estimation errors over a number of experiments and are shown in Table 4.

In the second group of experiments, we computed the variance of the T-ACM under different build-parameters for equality-select, range-select and equi-join operations and the corresponding percentage estimation errors in this set of experiments. The average percentage estimation errors and the corresponding variance of the T-ACM are given in Table 5. Note that the row numbers I, II, and III correspond to equality-select, range-select and equi-join operations respectively.

[1]The CENSUS database contains information about households and persons in the U.S.A, providing various statistics. The Data Extraction System (DES) at the U.S. Census Bureau allows extracting records and fields from very large public information archives such as governmental surveys and census records. It produces custom extracts in selectable formats that can be later analyzed by statistical packages.

No	Size	Estimated Result			Percentage Error		
		$\mathcal{V} = 1007$	$\mathcal{V} = 1196$	$\mathcal{V} = 1493$	$\mathcal{V} = 1007$	$\mathcal{V} = 1196$	$\mathcal{V} = 1493$
I	72	74.36	75.60	78.66	3.28%	4.99%	9.25%
II	318	297.55	343.79	360.90	6.43%	8.11%	13.49%
III	163	171.67	175.06	323.72	5.32%	7.40%	9.86%

Table 5: Variance of the T-ACM and the Estimation Errors: U.S. CENSUS Database

8.1 Analysis of the Results

The results from the above set of experiments show that the estimation error resulting from the T-ACM is *consistently* much lower (indeed, of an order of magnitude) than the estimation error from the equi-width and equi-depth histograms. This is consequent to the fact that the trapezoidal rule of the numerical integration technique is more accurate than the right-end or left-end histogram approximation techniques. Thus we see in the equi-select operation in Table 4, the percentage estimation error from the T-ACM is only 3.91%, but that of the equi-width and equi-depth histograms are 27.36% and 24.91% respectively, demonstrating an order of magnitude of superior performance. Such results are typical with both synthetic data and real-world data. The power of the T-ACM is obvious!

The results from the second set of experiments show that the estimation accuracy falls in an inversely proportional manner to the variance of the T-ACM for all three types of query operations considered. For example, considering the range-select query in row II of Table 5, we see that the percentage estimation error from the T-ACM with the variance of $\mathcal{V} = 1007$ is only 6.43%, whereas the percentage estimation error from the T-ACM with the variance of $\mathcal{V} = 1493$ is equal to 13.49%, which is more than a 100% increase!

Our results from these two sets of experiments confirm our theoretical results, summarized in Table 3, and clearly demonstrate that the estimation accuracy of the T-ACM is superior to that of the traditional equi-width and equi-depth histograms.

9 Conclusion

In this paper we have introduced a new histogram-like approximation strategy, called the Trapezoidal Attribute Cardinality Map, for query result size estimation. Since this technique is based on the philosophy of numerical integration, it is much more accurate than the traditional histograms. By proving a Binomial distribution (whose parameters vary with the location of the attribute value) to represent frequency variations within sectors, we have provided theoretical results to compare the accuracy of the T-ACM to that of the traditional histograms, both in the average-case and worst-case. We have also conducted extensive experiments using real-world data (U.S. CENSUS, 1997) to support the validity of our theoretical results. We hope that due to its high accuracy and relatively low construction costs, it could prove to be a standard tool for query result size estimation in future database systems.

References

[1] S. Christodoulakis. Estimating selectivities in data bases. In *Technical Report CSRG-136*, Computer Science Dept, University of Toronto, 1981.

[2] S. Christodoulakis. Estimating record selectivities. In *Information Systems*, volume 8, 1983.

[3] Christos Faloutsos, Yossi Matias, and Avi Silberschatz. Modeling skewed distributions using multifractals and the 80-20 law. In *Technical Report*, Dept. of Computer Science, University of Maryland, 1996.

[4] Yannis Ioannidis. Universality of serial histograms. In *Proceedings of the 19th International Conference on Very Large Databases*, Dec 1993.

[5] Yannis Ioannidis and S. Christodoulakis. On the propagation of errors in the size of join results. In *Proceedings of the ACM SIGMOD Conference*, pages 268–277, 1991.

[6] Yannis Ioannidis and Viswanath Poosala. Balancing histogram optimality and practicality for query result size estimation. In *ACM SIGMOD Conference*, pages 233–244, 1995.

[7] R. P. Kooi. *The optimization of queries in relational databases.* PhD thesis, Case Western Reserve University, 1980.

[8] Erwin Kreyszig. *Advanced Engineering Mathematics.* John Wiley & Sons, New York, 6th edition, 1988.

[9] M.V. Mannino, P. Chu, and T. Sager. Statistical profile estimation in database systems. In *ACM Computing Surveys*, volume 20, pages 192–221, 1988.

[10] A.W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications.* Academic Press, New York, 1979.

[11] M. Muralikrishna and David J Dewitt. Equi-depth histograms for estimating selectivity factors for multi-dimensional queries. In *Proceedings of ACM SIGMOD Conference*, pages 28–36, 1988.

[12] B. John Oommen and Murali Thiyagarajah. The Rectangular Attribute Cardinality Map: A New Histogram-like Techniques for Query Optimization. Technical Report TR-99-01, School of Computer Science, Carleton University, Ottawa, Canada, Jan 1999.

[13] Gregory Piatetsky-Shapiro and Charles Connell. Accurate estimation of the number of tuples satisfying a condition. In *Proceedings of ACM SIGMOD Conference*, pages 256–276, 1984.

[14] P. Selinger, D.D. Chamberlin M.M. Astrahan, R.A. Lorie, and T.G. Price. Access path selection in a relational database management system. In *Proceedings of ACM-SIGMOD Conference*, 1979.

[15] Murali Thiyagarajah. PhD Thesis - In preparation, School of Computer Science, Carleton University, Ottawa, Canada.

[16] U.S. Census Bureau. U.S. CENSUS Database. 1997.