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

B.J. Oommen and Murali Thiyagarajah

TR-99-01 January 1999

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

Rectangular 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

Current database systems utilize histograms to approximate frequency distributions of attribute values of relations. These are used to efficiently estimate query result sizes and access plan costs. 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 this paper, we introduce a new histogram-like approximation strategy, called the Rectangular Attribute Cardinality Map (R-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 rectangular cell, where the height of the cell is obtained so as to guarantee that the actual probability density differs from the approximated one by a maximum of a user-specified tolerance, τ. Furthermore, unlike the two traditional histogram types, namely equi-width and equi-depth, the R-ACM is neither equi-width nor equi-depth. Analytically, we show that for the R-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 R-ACM scheme (which is essentially histogram-like) is much more accurate than the traditional histograms for query result size estimation. Due to its high accuracy and low construction costs, we hope that it 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 [6] 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 [3, 4, 9, 10, 12, 14, 15]. 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 this paper we introduce a new catalogue-based non-parametric statistical model called the *Rectangular Attribute Cardinality Map* (R-ACM) that can be used to obtain more accurate estimation results than the currently known estimation techniques. We argue that the R-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 R-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 [16]. However, to present our results in the right perspective a brief survey is given.

Equi-width histograms for single-attributes were considered by Christodoulakis [2] and Kooi [9]. 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 [14]. This was later extended as multi-dimensional equi-depth histograms to represent multiple attribute values by Muralikrishna and Dewitt [12].

Ioannidis and Christodoulakis took a different approach by grouping attribute values based on their frequencies [6, 7, 5]. 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 [6, 7]. The serial histograms provide optimal results for equality join estimations, but less than optimal results for range queries. Faloutsos *et al* [4] proposed using a multi-fractal assumption for real-data distribution as opposed to the uniformity assumptions made within current histograms.

Ioannidis and Poosala [8] 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 Rectangular Attribute Cardinality Map

The Rectangular Attribute Cardinality Map (R-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 subdivisions for the range of values assumed by that attribute. The R-ACM is, in fact, a modified form of the histogram. But unlike the two major forms of histograms, namely, the equi-width histogram, where all the sector widths are equal, and the equi-depth histogram, where the number of tuples in each histogram bucket is equal, the R-ACM has a variable sector width, and has varying number of tuples in each sector. The sector widths or subdivisions of the R-ACM are generated according to a rule that aims at minimizing the estimation error within each subdivision.

Before we proceed, we present the notations that will be used throughout this paper in Table 2.

The R-ACM can be either a one-dimensional or a multi-dimensional depending on the

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

number of attributes being mapped. To introduce the concepts formally, we shall deal with the one-dimensional case in this paper.

Definition 1 A One dimensional Rectangular 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*[1] *of an attribute* X *in relation* R. *Let the value set* $\mathcal{V}$ *be subdivided into* s *number of sector widths according to the range partitioning rule described below. Then the Rectangular Attribute Cardinality Map of attribute* X *is an integer array in which the* j^{th} *index maps the number of tuples in the* j^{th} *value range of the set* $\mathcal{V}$ *for all* j, $1 < j \leq s$.

Rule 1 Range Partitioning Rule: *Given a desired tolerance value* τ *for the R-ACM, the sector widths,* $l_j, 1 \leq j \leq s$, *of the R-ACM should be chosen such that for any attribute value* X_i, *its frequency* x_i *does not differ from the running mean of the frequency of the sector by more than the tolerance value* τ, *where the running mean is the mean of the frequency values examined so far in the current sector.*

For example, consider the frequency set $\{8, 6, 9, 7, 19, 21, 40\}$ corresponding to the attribute values $\{X_0, X_1, X_2, X_3, X_4, X_5, X_6\}$ of an attribute X. Using a tolerance value $\tau = 2$, the attribute value range will be partitioned into the three sectors, $\{8, 6, 9, 7\}, \{19, 21\}, \{40\}$ with sector widths of 4, 2, and 1 respectively..

3.1 Generating the Rectangular ACM

Using the *range partitioning rule*, Algorithm `Generate_R-ACM` partitions the value range of the attribute X into s variable width sectors of the R-ACM.

[1] In this work, only ordinal numerical values are considered for attributes. It is possible to convert non-numeric attributes (symbolic, scalar-typed, fuzzy etc.) into ordinal numbers using a mapping function. This includes most type of attribute values. For non-ordinal data (for example, "real valued" data), we have proposed another type of structure called the Trapezoidal Attribute Cardinality Map (T-ACM). Results pertaining to the T-ACM are currently being compiled for publication.

Algorithm 1 Generate_R-ACM

```
Input:  tolerance τ, frequency distrib.  of X as A[0...L-1]
Output:  R-ACM
begin
   Initialize_ACM;      /* set all entries in ACM to zero */
   current_mean := A[1]; j := 0;
   ACM[j] := A[1];
   for i:=1 to L-1 do      /* for every attribute value */
      if abs(A[i]- current_mean) < τ
         ACM[j] := ACM[j] + A[i];
         current_mean := (current_mean*i + A[i])/(i+1); /* running mean */
      else begin
         l_j := i - 1;      /* set the sector width */
         j++;      /* move to next sector */
         current_mean := A[i];
         ACM[j] := A[i];
      end;
end;
end Algorithm.
```

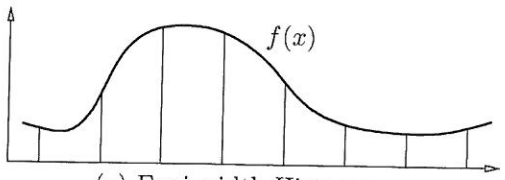
The input to the algorithm are the tolerance value τ for the ACM and the actual frequency distribution of the attribute X. The frequency distribution is assumed to be available in an integer array A, which has a total of L entries for each of the L distinct values of X. For simplicity reasons, we assume that the attribute values are ordered integers from 0 to $L-1$. The output of the algorithm is the R-ACM for the given attribute value set.

It is obvious that the algorithm, `Generate_R-ACM` generates the R-ACM corresponding to the given frequency value set.

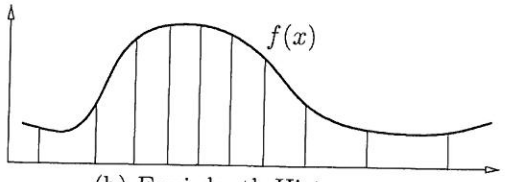
Assuming that the frequency distribution of X is already available in array A, the running time of the algorithm `Generate_R-ACM` is $O(L)$ where L is the number of distinct attribute values.

The reader will observe that we have assumed that the tolerance value, τ, is an input to the above algorithm. The question of how to determine an "optimal" tolerance value for an R-ACM remains open. We are currently investigating the use of adaptive techniques (involving possibly, learning automata) to solve this.

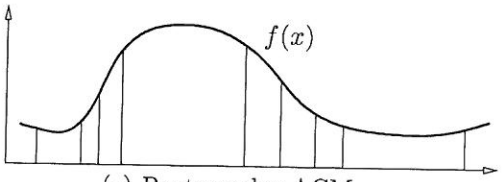
Since the ACM only stores the count of the tuples and not the actual data, it does not incur the usually high I/O cost of having to access the base relations from secondary storages. Secondly, unlike the histogram-based or other parametric and probabilistic counting estimation methods in use currently [10], ACM does not use sampling techniques to approximate the data distribution. Each cell of the ACM maintains the *actual* number of tuples that fall between the boundary values of that cell, and thus, although this leads to an approximation of the density function, there is no approximation of the number of tuples in the data distribution.



(a) Equi-width Histogram



(b) Equi-depth Histogram



(c) Rectangular ACM

Figure 1: R-ACM and Traditional Histograms. Note in (b), the areas of the sectors are equal.

The one-dimensional R-ACM as defined above can be extended to a multi-dimensional one easily to map an entire multi-attribute relation. A multi-dimensional ACM can also be used to store the multi-dimensional attributes that commonly occur in geographical, image, and design databases.

Before we derive any of the analytical properties of the R-ACM, it would be fitting to compare the three methodologies from a common conceptual perspective.

3.2 Rationale for the Rectangular ACM

Without loss of generality, let us consider an arbitrary continuous frequency function $f(x)$. Figure 1 shows the histogram partitioning of $f(x)$ under the traditional equi-width, equi-depth methods and the R-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 R-ACM method,

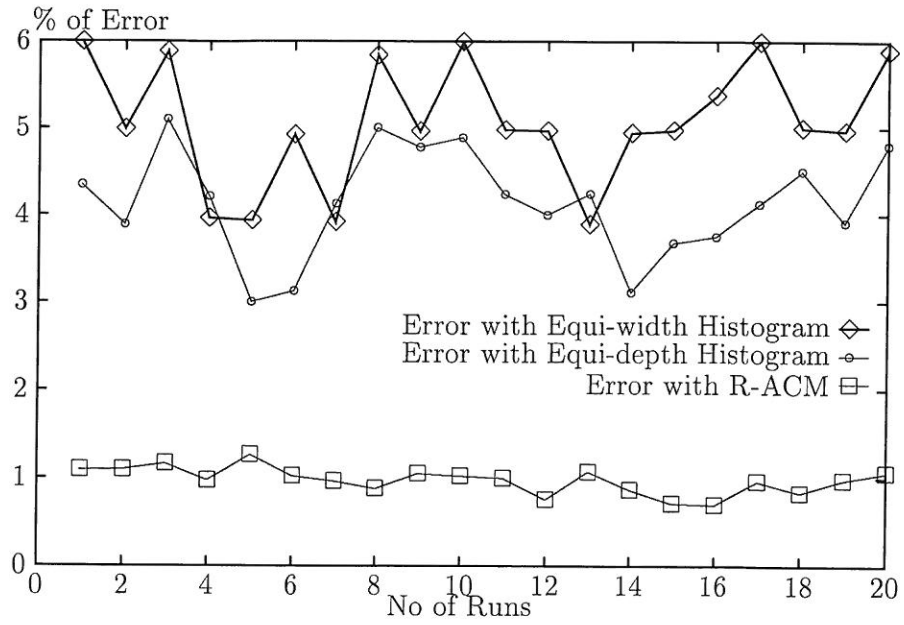


Figure 2: Comparison of Equi-width, Equi-depth Histograms and the R-ACM for Probability Estimation: Each experiment was run 500,000 times to get the average percentage of errors in the estimated occurrence of the attribute values. Estimation errors are given for the exact match on a random distribution with 100,000 tuples and 1000 distinct values. For the R-ACM, the tolerance was $\tau = 3$.

we note that whenever there is a steep frequency changes, the corresponding sector widths proportionally decrease (or in other words, the number of sectors proportionally increases). Hence the actual frequencies of all the attribute values within a sector are assured to be closer to the average frequency of that sector. This partitioning strategy obviously increases the estimation accuracy. Figure 2 shows a comparison of probability estimation errors obtained on all three estimation methods on synthetic data.

The rationale for partitioning the attribute value range using a tolerance value 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 $Var(X_k) = E[(x_k - \mu_k)^2]$, forcing the difference between the frequency of a given value and the running mean of the frequencies to be less than the tolerance τ, (i.e: $|x_k - \mu_k| \leq \tau$, will ensure 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 5). 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.

To demonstrate the relationship between the selection (random match) estimation error and the variance of the ACM, we in all brevity, mention results of an experiment in which we changed the sector widths of the ACM and computed the corresponding variance (i.e:

$Var(ACM)$). The errors between the estimated and actual size of random matches are plotted against the computed variance of the ACM, and shown in Figure 3. The advantages of using the R-ACM are obvious. Indeed, more detailed expressions and experimental results will later strengthen this *initial* claim.

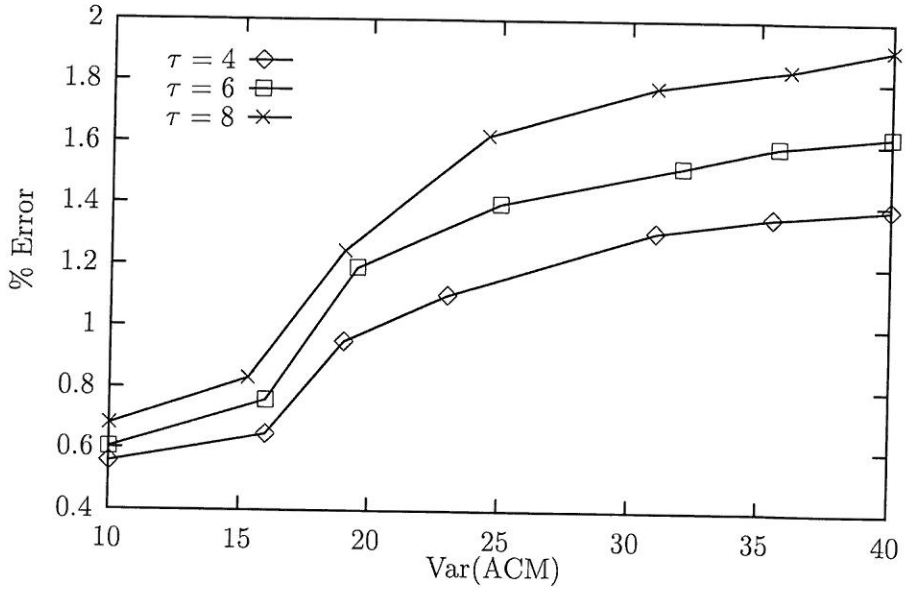


Figure 3: Estimation Error vs ACM Variance: The ACM sectors were randomly partitioned without using a tolerance value and the resulting ACM variances were computed. Using random selection queries (matches), the errors between the actual and expected frequencies where obtained.

4 Density Estimation Using Rectangular ACM

We shall now study the properties of the R-ACM with regard to density approximation. To do this we consider a one-dimensional Rectangular ACM sector of width l with n tuples. Also to render the analysis feasible, we make the following assumption.

Assumption 1 *The attribute values within an R-ACM sector are uniformly distributed.*

Rationale: Since the sector widths of the R-ACM are chosen so that the frequency of the values within the sectors do not differ by more than the allowed tolerance, τ, we can see that these frequencies are guaranteed to be close to each other. The original System R research work [15] relied on the (often erroneous) assumption that the frequencies of an attribute value are uniformly distributed across the *entire attribute value domain.* With the adoption of the equi-width and equi-depth histograms in the modern-day database systems (see Table 1), this was improved by making the uniformity assumptions only within the histogram buckets. Our uniformity assumption within an R-ACM sector is a much weaker assumption

than that used in any other previous works on histograms, due to its partitioning strategy. Indeed, as we shall show from the experimental results in Section 7, this assumption is satisfied with almost no additional approximation.

Using the above uniformity assumption, we shall now derive the probability mass distribution for the R-ACM.

Lemma 1 *The probability mass distribution for the frequencies of the attribute values in an R-ACM is a Binomial distribution with parameters* $(n, \frac{1}{l})$.

Proof: Since there are l distinct values in the sector, the probability of any of these l values, say X_i, occurring in a random assignment of that value to the sector is equal to $\frac{1}{l}$. (See Figure 4).

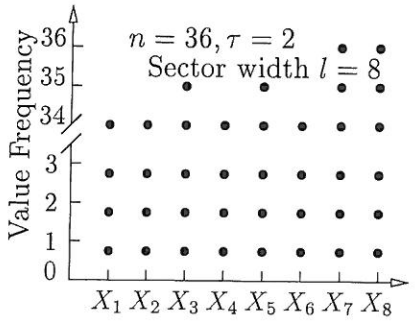


Figure 4: Distribution of Values in an ACM Sector

Consider an arbitrary permutation (or arrangement) of the n tuples in the sector. Suppose the value X_i occurs exactly x_i times. This means all the other $(l-1)$ values must occur a combined total of $(n-x_i)$ times. Since the probability of X_i occurring once is $\frac{1}{l}$, the probability of it not occurring is $(1-\frac{1}{l})$. Hence the probability of an arbitrary permutation of the n tuples, where the value X_i occurs exactly x_i times and the other values collectively occur $n - x_i$ times is,

$$\left(\frac{1}{l}\right)^{x_i} \left(\frac{l-1}{l}\right)^{n-x_i}. \tag{1}$$

Clearly there are $\binom{n}{x_i}$ different permutations of the n tuples in the sector where the above condition is satisfied. Hence we find that the total probability that an arbitrary value X_i occurs exactly x_i times is,

$$p_{X_i}(x_i) = \binom{n}{x_i} \left(\frac{1}{l}\right)^{x_i} \left(\frac{l-1}{l}\right)^{n-x_i} \tag{2}$$

which is exactly the Binomial distribution with parameters $(n, \frac{1}{l})$. This proves the lemma. □

5 Maximum Likelihood Estimate Analysis for the R-ACM

In the previous section we showed that the frequency distribution for a given attribute value in the R-ACM obeys a Binomial distribution. With this as a background, we shall now derive a maximum likelihood estimate for the frequency of an arbitrary attribute value in an R-ACM sector. In classical statistical estimation theory, we are usually interested in estimating the parameters (such as the mean or other unknown characterizing parameters) of the distribution of one or more random variables. In our problem (which is like the inverse version of a traditional statistical estimation 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 (the total number of tuples, n and the width of the R-ACM sector, l). 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 rectangular ACM, the maximum likelihood estimate of the number of tuples for a given value X_i of attribute X is given by,*

$$\hat{x}_{ML} = \frac{n}{l}$$

where n is the number of tuples in the sector containing the value X_i and l is the width of that sector.

Proof: We know from Lemma 2 that the frequency distribution of a given attribute value in an R-ACM sector is a Binomial distribution. So the probability mass function of the frequency distribution of an attribute value $X = X_\alpha$ in an R-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,

$$\begin{aligned} \ln \mathcal{L}(x) &= \ln n! - \ln x! - \ln(n-x)! + x \ln p + (n-x)\ln(1-p) \\ &= \ln \Gamma(n+1) - \ln \Gamma(x+1) - \ln \Gamma(n-x+1) + \\ &\quad x \ln p + (n-x)\ln(1-p) \end{aligned} \tag{3}$$

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

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

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 3, 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 earlier that, due to the uniformity assumption within an R-ACM sector, $p = \frac{1}{l}$. So we have,

$$\hat{x}_{ML} = \frac{n}{l}.$$

Hence the theorem. □

The maximum likelihood estimate, $\hat{x}_{ML} = np$, which we derived using the Gamma function above is, most of the time, 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 rectangular ACM, the maximum likelihood estimate of the number of tuples for a given value X_i of attribute X falls within the range of,*

$$\frac{(n+1)}{l} - 1 \leq \hat{x}_{ML} \leq \frac{(n+1)}{l},$$

where n is the number of tuples in the 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.

$$f(x) - f(x+1) > 0 \quad (4)$$
$$f(x) - f(x-1) > 0 \quad (5)$$

From Equation (4), 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 (5), by using similar algebraic manipulation, we get,

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

Since $p = \frac{1}{l}$, the theorem follows. □

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

The maximum likelihood estimate of the frequency of a given attribute value tells us that the attribute value would have a frequency of $\hat{x}_{ML}$ with maximum 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 when we need to find the worst-case and average-case errors for our result size estimations, we need to obtain another estimate which includes all these possible frequency values. One such estimate is the expected value of the frequency of a given attribute value. We use our Binomial model to find the expected value of the frequency of an attribute value as given in the following lemma and develop a sequence of results regarding the corresponding estimates.

Lemma 2 *For a one-dimensional rectangular ACM, the expected number of tuples for a given value X_i of attribute X is $E(X_i) = n/l$, where n is the number of tuples in the sector containing the value X_i and l is the width of that sector.*

Proof: From Equation (2), the probability that the value X_i occurs exactly k times is,

$$p_{X_i}(k) = \binom{n}{k} \left(\frac{1}{l}\right)^k \left(\frac{l-1}{l}\right)^{n-k}$$

which is a Binomial distribution with parameters $(n, \frac{1}{l})$. The result follows directly from the fact that the mean of the binomial distribution, *Binomial* (n,p), is np, where p is the probability of success. □

The above result is very useful in estimating the results of selection and join operations.

6.1 Estimation Error with Rectangular 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 [6]. 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 R-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,

$$Var(X_k) = \sum_{i=0}^{n_j} x_i^2 \left(\frac{1}{l_j}\right)^i \left(1 - \frac{1}{l_j}\right)^{n_j - i} - \left(\frac{n_j}{l_j}\right)^2 \tag{6}$$

Lemma 3 *The variance of the frequency of an attribute value X in sector j of an R-ACM is,*

$$Var(X) = \frac{n_j(l_j - 1)}{l_j^2} \tag{7}$$

Proof: It is well known that the variance of a Binomial distribution with parameters (n, p) is $np(1-p)$. Hence using the property of the Binomial distribution, the expression for the variance given in Equation 6 can be reduced to the one given in the lemma. □

Lemma 4 *The sector variance of the j^{th} rectangular ACM sector is,*

$$Var_j = \frac{n_j(l_j - 1)}{l_j} \tag{8}$$

Proof: We note that $Var(X_k)$ is same for all $k, 1 \leq k \leq l_j$, in a given sector. Since the random variables are independent[2], summing up the variances of all the values in the sector will give us an upper bound for the estimation error or variance of the sector. The result follows. □

Similarly, summing up the variances of all the sectors, we obtain an expression for the variance of the entire ACM, which is given in the following lemma.

Lemma 5 *The variance of an R-ACM is given by,*

$$Var(ACM) = \sum_{i=1}^{s} Var_i \tag{9}$$

where s is the number of sectors in the ACM.

Proof: The result follows directly from the fact that the variances in each sector are independent, thus summing up the sector variances will yield the variance of the entire ACM. □

6.2 Error Estimates and Self-Joins

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 [11] 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.

[2]We really don't need independence for this. Uncorrelatedness is sufficient.

Lemma 6 *The error, ϵ, resulting from a self-join[3] of relation R on attribute X using a rectangular ACM is given by,*

$$\epsilon = Var(ACM) + \sum_{j=1}^{s} \left\{ \sum_{k=1}^{l_j} x_k^2 - \frac{n_j^2 + n_j l_j - n_j}{l_j} \right\}.$$

Proof: The proof is a little cumbersome and not included here in the interest of clarity and brevity. It is found in [13].

Theorem 3 *The variance of a rectangular ACM corresponding to attribute X is,*

$$Var(ACM) = N - \sum_{j=1}^{s} \frac{n_j}{l_j}. \tag{10}$$

Proof: From Lemma 5, the variance of an R-ACM is given by $Var(ACM) = \sum_{j=1}^{s} Var_j$, where Var_j is the variance of the j^{th} sector. But from Lemma 4, $Var_j = n_j(l_j - 1)/l_j$. Hence,

$$\begin{aligned} Var(ACM) &= \sum_{j=1}^{s} \frac{n_j(l_j - 1)}{l_j} = \sum_{j=1}^{s} n_j - \sum_{j=1}^{s} \frac{n_j}{l_j} \\ &= N - \sum_{j=1}^{s} \frac{n_j}{l_j}. \end{aligned}$$

Hence the theorem follows. □

6.3 Worst Case Error with the Rectangular ACM

As mentioned earlier, forcing a frequency value to be within a given tolerance τ to the running mean ensures that the frequency distribution within an R-ACM sector is very close to uniform. We note that whenever every frequency value is always consistently smaller (or always consistently greater) than the current mean by the tolerance value τ, the resulting sectors will be far from uniform. So we have the following definition.

Definition 2 *A distribution is said to be "least uniform" if for every attribute value of X_i, the frequency x_i attains the value $x_i = \mu_{i-1} - \tau$, if x_i is decreasing or $x_i = \mu_{i-1} + \tau$ if x_i is increasing, where μ_{i-1} is the mean of the first $(i-1)$ frequencies. A sector of the former type is called a monotonically decreasing R-ACM sector. Similarly a sector of the latter type is called a monotonically increasing R-ACM sector.*

[3] In one of his recent works [8], considering a V-optimal histogram (defined in his work), Poosala has claimed that the error resulting from a self-join is equal to the histogram variance. Indeed, although his result is basically true from an order-notation point of view, the more exact expression is given in this lemma.

The motivation for the above definition comes from the following observation. Assume that during the process of constructing an R-ACM sector, the next value x_i is smaller than the current mean μ_{i-1}. We note that if $x_i < \mu_{i-1} - \tau$ then, we will have generated a new sector. Hence the smallest value that x_i can assume is $x_i = \mu_{i-1} - \tau$. The resulting distribution is shown in Figure 5(a). This is formally given by the following lemma. This lemma is given for the case when the sector is a monotonically decreasing R-ACM sector, or in other words, when every frequency value is always smaller than the previous running mean. The case for the increasing R-ACM sector is proved in an analogous way.

Lemma 7 *A decreasing R-ACM sector is "least uniform", if and only if*

$$x_k = a - \sum_{i=1}^{k-1} \frac{\tau}{i} \qquad \text{for } 1 \leq k \leq l_j.$$

Proof: Note that in this case, least uniformity occurs when the quantity $E_S(X) - x_i$ assumes its largest possible value (or when x_i assumes its minimum value), where $E_S(X)$ is the expected value of X in a sector. We assume that the frequency of the first value is $x_1 = a$.

Basis: $x_1 = a$: Since the current mean is $E_S(X) = a$, and the sector is a decreasing R-ACM sector, the minimum possible value for x_2 is obtained from $E_S(X) - x_2 = \tau$. This is achieved when $x_2 = a - \tau$.

Inductive hypothesis: Assume the statement is true for $n = k$. So the sector is the least uniform for the first k values and the frequencies take the following values:

$$\begin{aligned} x_1 &= a \\ x_2 &= a - \tau \\ x_3 &= a - \frac{3\tau}{2} \\ &\vdots \quad \vdots \quad \vdots \\ x_k &= a - \sum_{i=1}^{k-1} \frac{\tau}{i} \end{aligned}$$

For $n = k + 1$: The minimum possible value for x_{k+1} without creating a new sector is obtained when $E_S(X) - x_{k+1} = \tau$. This is achieved when $x_{k+1} = E_S(X) - \tau$. So,

$$\begin{aligned} x_{k+1} &= \frac{x_1 + x_2 + \ldots + x_k}{k} - \tau \\ &= a - \frac{\tau + \frac{3\tau}{2} + \ldots + \sum_{i=1}^{k-1} \frac{\tau}{i}}{k} - \tau \\ &= a - \frac{(k-1)\tau + (k-2)\frac{\tau}{2} + (k-3)\frac{\tau}{3} + \ldots + \frac{\tau}{k-1}}{k} - \tau \\ &= a - \frac{k\tau(\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{k-1}) - (k-1)\tau}{k} - \tau = a - \sum_{i=1}^{k} \frac{\tau}{i} \end{aligned}$$

This proves the lemma. □

Lemma 8 *An increasing R-ACM sector is "least uniform", if and only if*

$$x_k = a + \sum_{i=1}^{k-1} \frac{\tau}{i} \qquad \text{for } 1 \leq k \leq l_j.$$

Proof: The proof is analogous to that of Lemma 7 and omitted in the interest of brevity. □

We shall now present a tight bound for the frequency x_i of an arbitrary attribute value X_i in the following theorem.

Theorem 4 *If the value X_i falls in the j^{th} sector of an R-ACM, then the number of occurrences of X_i is,*

$$\frac{n_j}{l_j} - \left|\tau\left[\ln\left(\frac{l}{i-1}\right) - 1\right]\right| \leq x_i \leq \frac{n_j}{l_j} + \left|\tau\left[\ln\left(\frac{l}{i-1}\right) - 1\right]\right|$$

where n_j and l_j are the number of tuples and the sector width of the j^{th} sector.

Proof: Consider a sector from an R-ACM. Let the frequency of the first value x_1 be a. We note that the R-ACM sector will become a "skewed" one, if the subsequent values are **all** smaller or **all** greater than the previous running mean by τ. From Lemmas 7 and 8, it is obvious that such sectors are the "least uniform" ones in an R-ACM, and consequently the largest estimation error occurs in such a "skewed" sector. Assume that the frequency values from x_1 to x_{l_j} decrease monotonically in this manner. In other words, if the mean of the first k values is μ_k, then the next value will take its lowest allowable frequency, $\mu_k - \tau$. The resulting distribution is shown in Figure 5(a). From Lemma 7, the frequency of an arbitrary attribute value X_i is given by,

$$x_i = a - \sum_{k=1}^{i-1} \frac{\tau}{k}$$

The expected value $E(X_i)$ is the mean frequency for the entire sector. So,

$$\begin{aligned} E(X_i) &= \frac{\sum_{k=1}^{l_j} x_k}{l_j} \\ &= a - \tau\left(1 + \frac{1}{2} + \ldots + \frac{1}{l_j - 1} - \frac{l_j - 1}{l_j}\right) \end{aligned}$$

But the frequency of an arbitrary value X_i is,

$$x_i = a - \tau\left(1 + \frac{1}{2} + \ldots + \frac{1}{i-1}\right)$$

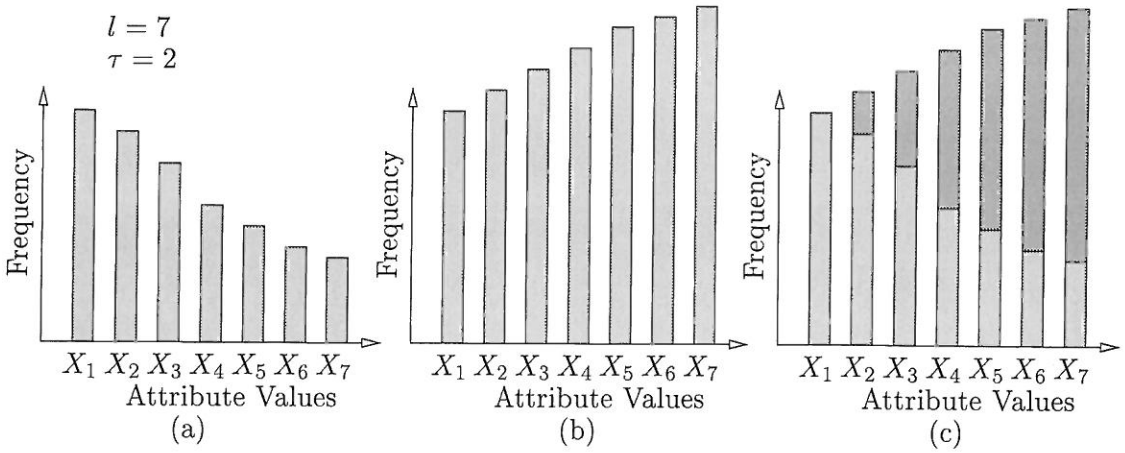


Figure 5: (a) A decreasing R-ACM sector; (b) An increasing R-ACM sector; (c) Darkly shaded area represents the likely frequencies of attribute values given by Theorem 4.

So the estimation error is,

$$x_i - E(X_i) = \tau \left(\sum_{k=i}^{l_j} \frac{1}{k} - 1 \right) \leq \left| \tau \left[\ln \left(\frac{l}{i-1} \right) - 1 \right] \right|$$

Hence,

$$x_i \geq \frac{n_j}{l_j} - \left| \tau \left[\ln \left(\frac{l}{i-1} \right) - 1 \right] \right|$$

Similarly using symmetry, for an R-ACM sector with monotonically increasing frequency value (see Figure 5(b)), we can show that,

$$x_i \leq \frac{n_j}{l_j} + \left| \tau \left[\ln \left(\frac{l}{i-1} \right) - 1 \right] \right|.$$

The theorem follows. □

The reader should note that the composite effect of the monotonically decreasing frequency sequence and the monotonically increasing frequency sequence restricts the set of frequency values which the attributes can take as shown in Figure 5(c). The following example illustrates the implications of the above theorem.

Example 1 *Consider an R-ACM sector of width 10 containing 124 tuples, where the R-ACM is partitioned using a tolerance value $\tau = 3$. Let us attempt to find the estimated frequency ranges for the attribute values (a) X_3 and (b) X_6.*

(a). The frequency range of X_3 is,

$$\begin{aligned} 12.4 - |3(\ln 5 - 1)| \leq \quad & x_3 \quad \leq 12.4 + |3(\ln 5 - 1)| \\ 10.57 \leq \quad & x_3 \quad \leq 14.23 \end{aligned}$$

(b). The frequency range of X_6 is,

$$\begin{aligned} 12.4 - |3(\ln 2 - 1)| \leq \quad & x_6 \quad \leq 12.4 + |3(\ln 2 - 1)| \\ 11.48 \leq \quad & x_6 \quad \leq 13.32 \end{aligned}$$

Notice that in both the above cases, the possible frequency values from an equi-width or equi-depth histograms are $0 \leq x \leq 124$, where $x = x_3$ or x_6. The power of the R-ACM in the estimation is obvious!

6.4 Worst-Case Error in Range Select Queries

Unlike the previous case, when estimating the sum of frequencies in an attribute value range, we have to consider three distinct cases. These (See Figure 6) are namely the cases when,

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

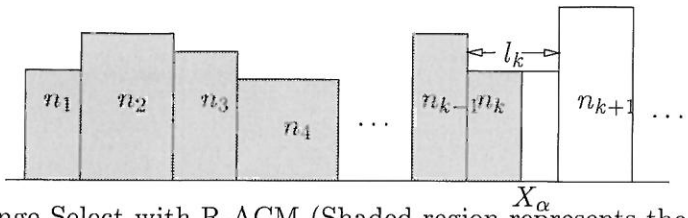
In the first case, estimation using the R-ACM gives the accurate result (n_j) and there is no estimation error. The estimation error in the second case is given by the theorem below (See Figure 7 (a)). The estimation error in the third case can be obtained by noting that it is in fact the combination of the first and second cases.

Theorem 5 *An estimate for the worst-case error, ϵ, in estimating the sum of frequencies of X_i in the attribute value range, $X_\alpha \leq X_i \leq X_\beta$, when both X_α and X_β fall completely within an R-ACM sector is given by,*

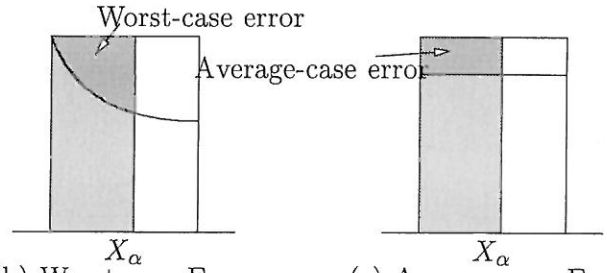
$$\epsilon = \left| \ln \left\{ l_j^{(\beta-\alpha+1)\tau} \frac{(\beta-1)!}{(\alpha-2)!} \right\} \right| - (\beta - \alpha + 1)\tau$$

where $\beta > \alpha$.

Proof: Using Theorem 4, we can compute the error resulting from this result size estimation by summing up the worst-case error for each of the attribute values in the given value range.

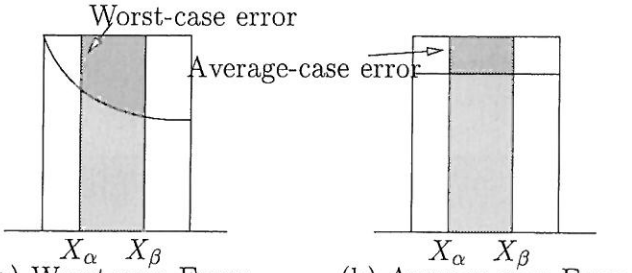


(a) Range Select with R-ACM (Shaded region represents the result)



(b) Worst-case Error (c) Average-case Error

Figure 6: Estimation of Range-select Using the R-ACM



(a) Worst-case Error (b) Average-case Error

Figure 7: Estimation of a Range Completely within an R-ACM Sector

Thus we have the following cumulative error.

$$\begin{aligned}
\epsilon &= \sum_{i=\alpha}^{\beta} \left\{ \tau \left| \ln\left(\frac{l_j}{i-1}\right) - 1 \right| \right\} \\
&= \tau \sum_{i=\alpha}^{\beta} \ln\left(\frac{l_j}{i-1}\right) - (\beta - \alpha + 1)\tau \\
&= \left| \ln\left\{ l_j^{(\beta-\alpha+1)\tau} \frac{(\beta-1)!}{(\alpha-2)!} \right\} \right| - (\beta - \alpha + 1)\tau.
\end{aligned}$$

Hence the theorem follows. □

6.5 Average Case Error with Rectangular ACM

The previous theorem gives the bounds for the worst-case estimation error by considering the "least uniform" sector. But what about the estimation error on the average? Average case error occurs in a truly random sector. In a random sector, frequency values do not monotonically increase or decrease as in the least uniform case we considered before. Instead, they take on a random value around the mean bounded by the tolerance value. In other words, if the current mean is μ, then the next frequency value is a random variable between $\mu - \tau$ and $\mu + \tau$. Whenever the next frequency value falls outside the range of $[\mu - \tau, \mu + \tau]$, then a new sector is generated.

Theorem 6 *The average case error in estimating the frequency of an arbitrary attribute value with a rectangular ACM is bounded by* 2τ.

Proof: To compute the average case error, we are required to compute the error at every attribute value and then take its expected value by weighting it with the probability of the corresponding error. However, since the frequency of an attribute value can vary only in the range of $[\mu - \tau, \mu + \tau]$, the maximum variation the frequency can have is 2τ. It follows that the maximum error is always bounded by 2τ, and the result follows. □

6.6 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 9 *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 R-ACM and T-ACM.

Corollary 1 *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 R-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. □

Since the worst-case error in estimating the selection queries $\sigma_{X=X_i}$ and $\sigma_{Y=Y_j}$ is dependent on the positions of the attribute values X_i and Y_j within the corresponding R-ACM sectors, we note that the worst-case error in the above join is also dependent on the positions of the attribute values being joined. Figure 8 shows the relationship of the worst-case join estimation error and the positions i, j of the attribute values X_i and Y_j within the R-ACM sectors. Note that the join estimation has the lowest worst-case error when both X_i and Y_j are the last attribute values in their corresponding sectors.

We compare the worst-case and average-case estimation errors for equality-match queries in the traditional equi-width, equi-depth histograms and the new R-ACM method in Table 3 and note that the R-ACM provides much smaller estimation errors than the traditional histograms.

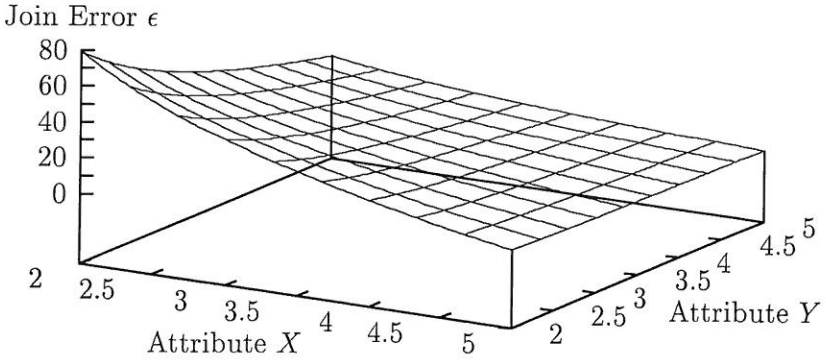


Figure 8: Join Estimation Error and the Positions of Attribute Values

Histogram	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	$\frac{2}{3l_j}$	$\frac{1}{2l_j}$
R-ACM	$\tau\left\lvert\ln\left(\frac{l_j}{i-1}\right) - 1\right\rvert$	2τ

Table 3: Comparison of Histogram and R-ACM Errors. Note that τ is much smaller than l_j.

7 Experimental Results

We conducted an extensive array of experiments, involving a number of real-world databases, to compare the performance of the R-ACM to the traditional equi-width and equi-depth histograms currently being used by most commercial database systems. Tables 4 and 5 show the experimental results conducted on some of the data set generated from the U.S. CENSUS database[4][1]. Since it is impossible to list all of our experimental results here, we refer the reader to [16] for more detailed information.

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

[4]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.

Operation	Actual Size	Equi-width		Equi-depth		R-ACM	
		Size	Error	Size	Error	Size	Error
Equi-select	1796	1279.1	28.8%	1365.6	23.4%	1702.3	5.23%
Range-select	32109	30008.3	6.5%	31214.9	2.8%	32319.2	-0.65%
Equi-join	720988	543908	24.6%	610482	15.3%	660183	8.43%

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

Operation	Actual Size	Estimated Result			Percentage Error		
		$\tau = 4$	$\tau = 6$	$\tau = 8$	$\tau = 4$	$\tau = 6$	$\tau = 8$
Equi-select	1435	1338.5	1292.1	1143.6	6.75%	9.97%	20.34%
Range-select	26780	26219.1	24918.3	22098.9	2.09%	6.95%	17.48%
Equi-join	563912	610180	680953	719320	-8.2%	-20.8%	27.56%

Table 5: Result Estimation Using R-ACM: Data - U.S. CENSUS Database

The first group of experiments were conducted on equi-width, equi-depth histograms and the R-ACM. In each of our experimental runs, we chose different build-parameters for the histograms and the R-ACM. The build-parameters for the equi-width and equi-depth histograms are the sector width and the number of tuples within a sector respectively. The build-parameter for the R-ACM is the tolerance value, τ.

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 compared the result estimates from the R-ACM for three different tolerance values. Again we considered (a) exact match select queries (b) range select queries and (c) equi-join queries and obtained the average estimation errors by comparing the estimates to the actual result sizes. The results are given in Table 5.

7.1 Analysis of the Results

The results from the first set of experiments show that the estimation error resulting from the R-ACM is **consistently** much lower than the estimation error from the equi-width and equi-depth histograms. This is consequent to the fact the frequency distribution of an attribute value within an R-ACM is guaranteed to be close to the sector mean since the partitioning of the sectors is based on a user-specified tolerance value, and the deviation from the running sector mean. Thus, for example, we see that in the equi-select operation in Table 4, the percentage estimation error from the R-ACM is only 5.23%, which is significantly smaller

than that obtained by the equi-width and equi-depth histograms which are 28.8% and 23.4% respectively. This indeed demonstrates an order of magnitude of superior performance. Such results are typical with both synthetic data and real-world data (See [16] for more detailed experimental results). The power of the R-ACM is obvious!

The second set of experiments illustrates that the accuracy of the R-ACM falls in an inversely proportional manner to the tolerance value. Since smaller tolerance values result in a proportionally larger number of sectors in the R-ACM, there is obviously a trade-off of estimation accuracy and the storage requirements of the R-ACM. From Table 5, we see that for the tolerance value $\tau = 4$, the percentage error for the range-select query is only 2.09%. As opposed to this, when the tolerance value is increased to $\tau = 8$, the percentage error for the same operation becomes 17.48%. Such results are again typical.

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 R-ACM is superior to that of the traditional equi-width and equi-depth histograms.

8 Conclusion

In this paper we have introduced a new histogram-like approximation technique, called the Rectangular Attribute Cardinality Map, for query result size estimation. Since this strategy is based on a user-defined tolerance value for partitioning the sectors, its worst-case and average-case errors are assured to be within a desired bound, and thus it is more accurate than the traditional histograms. By proving a Binomial distribution to represent frequency variations within sectors, we have first of all presented a maximum likelihood analysis, and thereafter provided theoretical results to compare the accuracy of the R-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 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] U.S. Census Bureau. U.S. Census database. 1997.

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

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

[4] 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.

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

[6] 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.

[7] Yannis Ioannidis and Stavros Christodoulakis. Optimal histograms for limiting worst-case error propagation in the size of join results. In *ACM TODS*, 1992.

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

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

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

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

[12] 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.

[13] B. John Oommen and Murali Thiyagarajah. The rectangular attribute cardinality map: A new histogram-like techniques for query optimization. Technical report, School of Computer Science, Carleton University, Ottawa, Canada, Oct 1998.

[14] 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.

[15] 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.

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