

The background of the slide is a photograph of a historic building with a prominent tower and a group of students sitting on a green lawn in front of it. The building has a mix of red and grey stone. The students are gathered in a circle, some looking at a laptop. The sky is blue with some clouds.

INFOH417 Database System Architectures


Mahmoud SAKR <mahmoud.sakr@ulb.be>

École polytechnique de Bruxelles

2022

Statistics for Cost Estimation

Statistical Information for Cost Estimation

- n_r : number of tuples in a relation r .
- b_r : number of blocks containing tuples of r .
- l_r : size of a tuple of r .
- f_r : blocking factor of r — i.e., the number of tuples of r that fit into one block.
- $V(A, r)$: number of distinct values that appear in r for attribute A ; same as the size of $\Pi_A(r)$. 
- If tuples of r are stored together physically in a file, then:

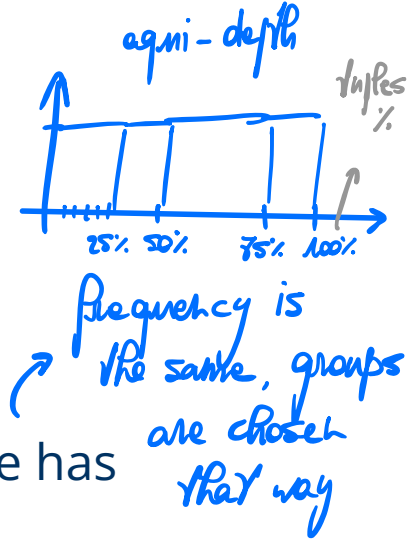
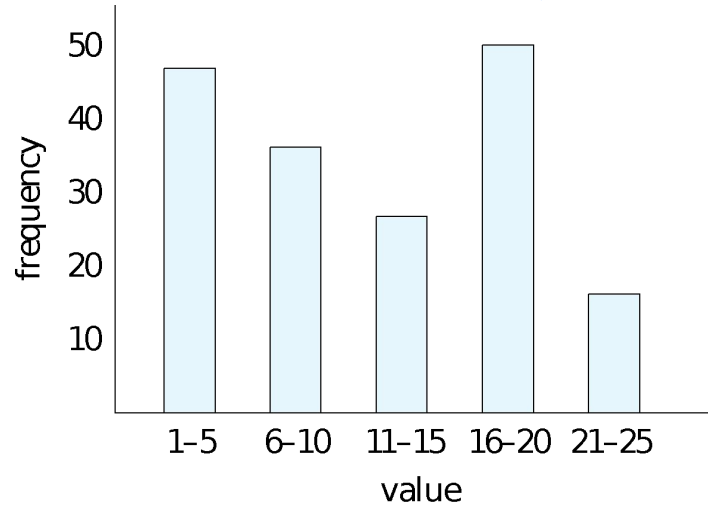
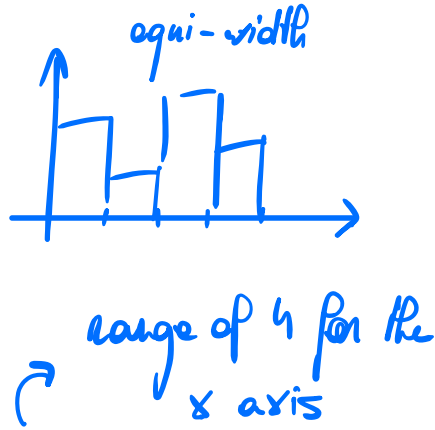
$$b_r = \left\lceil \frac{n_r}{f_r} \right\rceil$$

Histograms

one for each column

better understanding of the distribution of the attributes

- Histogram on attribute age of relation *person*



- Equi-width** histograms
- Equi-depth** histograms break up range such that each range has (approximately) the same number of tuples
 - E.g. (4, 8, 14, 19)
- Many databases also store *n* **most-frequent values** and their counts
 - Histogram is built on remaining values only

Histograms (cont.)

- Histograms and other statistics usually computed based on a **random sample**
- Statistics may be out of date
 - Some database require a **analyze (vacuum)** command to be executed to update statistics
 - Others automatically recompute statistics
 - e.g., when number of tuples in a relation changes by some percentage

only if data has changed

- $\sigma_{A=v}(n)$: if v is a primary key $\rightarrow c=1$
 else $c = \frac{n_n}{v(A,n)}$

- $\sigma_{A < v}(n)$: if $v < \min(A,n) \rightarrow c=0$
 $v > \max(A,n) \rightarrow c = n_n$

else $c = n_n \times \frac{v - \min(A,n)}{\max(A,n) - \min(A,n)}$

can be implemented using the histogram

$\hookrightarrow c = \frac{n_n}{2}$ if no stats linear interpolation

- $\sigma_{\theta_1 \dots \theta_n}(n)$: $c = n_n \times \frac{s_1 \times \dots \times s_n}{n_n^n}$ (assumption of independence)

where $\frac{s_i}{n_n}$ is the selectivity of θ_i

matching tuples

- $\sigma_{\theta_1 \dots \theta_n}(n)$: $c = n_n \times \left(1 - \left(1 - \frac{s_1}{n_n}\right) \times \dots \times \left(1 - \frac{s_n}{n_n}\right)\right)$

- $\sigma_{\neg \theta}(n)$: $c = 1 - \frac{s_\theta}{n_n}$ or $n_n - |\sigma_\theta(n)|$

Join: • $c(r \times s) = n_r \cdot n_s$

- $c(r \times s) \leq n_s$ or primary key of r

- $c(r \times s) = n_s$ or foreign key of s referencing r

- $c(r \times s) = \min \left\{ \frac{n_r \cdot n_s}{v(A,r)}, \frac{n_r \cdot n_s}{v(A,s)} \right\}$ or common attribute not a key

Postgres Statistics

Catalogue (table that stores statistics on the actual tables of the database)

Table 51.89. pg_stats Columns

Column	Type	Description
schemaname	name	(references <code>pg_namespace.nspname</code>) Name of schema containing table
tablename	name	(references <code>pg_class.relname</code>) Name of table
attname	name	(references <code>pg_attribute.attname</code>) Name of the column described by this row
inherited	bool	If true, this row includes inheritance child columns, not just the values in the specified table
null_frac	float4	Fraction of column entries that are null
avg_width	int4	Average width in bytes of column's entries
n_distinct	float4	If greater than zero, the estimated number of distinct values in the column. If less than zero, the negative of the number of distinct values divided by the number of rows. (The negated form is used when <code>ANALYZE</code> believes that the number of distinct values is likely to increase as the table grows; the positive form is used when the column seems to have a fixed number of possible values.) For example, -1 indicates a unique column in which the number of distinct values is the same as the number of rows.

`most_common_vals` anyarray

A list of the most common values in the column. (Null if no values seem to be more common than any others.)

`most_common_freqs` float4[]

A list of the frequencies of the most common values, i.e., number of occurrences of each divided by total number of rows. (Null when `most_common_vals` is.)

`histogram_bounds` anyarray → *pen equi-depth histogram*

A list of values that divide the column's values into groups of approximately equal population. The values in `most_common_vals`, if present, are omitted from this histogram calculation. (This column is null if the column data type does not have a < operator or if the `most_common_vals` list accounts for the entire population.)

`correlation` float4

Statistical correlation between physical row ordering and logical ordering of the column values. This ranges from -1 to +1. When the value is near -1 or +1, an index scan on the column will be estimated to be cheaper than when it is near zero, due to reduction of random access to the disk. (This column is null if the column data type does not have a < operator.)

`most_common_elems` anyarray

A list of non-null element values most often appearing within values of the column. (Null for scalar types.)

`most_common_elem_freqs` float4[]

A list of the frequencies of the most common element values, i.e., the fraction of rows containing at least one instance of the given value. Two or three additional values follow the per-element frequencies; these are the minimum and maximum of the preceding per-element frequencies, and optionally the frequency of null elements. (Null when `most_common_elems` is.)

`elem_count_histogram` float4[]

A histogram of the counts of distinct non-null element values within the values of the column, followed by the average number of distinct non-null elements. (Null for scalar types.)

Selection Size Estimation

$\sigma_{A=v}(r)$  *Expected size?*

- $n_r / V(A, r)$: number of records that will satisfy the selection
- Equality condition on a key attribute: *size estimate = 1*

↳ 0 or 1 time maximum in the relation

$\sigma_{A \leq v}(r)$ (case of $\sigma_{A \geq v}(r)$ is symmetric)

- Let c denote the estimated number of tuples satisfying the condition.
- If $\min(A, r)$ and $\max(A, r)$ are available in catalog

- $c = 0$ if $v < \min(A, r)$

- $c = n_r \cdot \frac{v - \min(A, r)}{\max(A, r) - \min(A, r)}$) *interpolation*

*compare v with the max
and min to see if we
are even going to get
something back
or if we
retrieve
everything*

- If histograms available, can refine above estimate
 - In absence of statistical information c is assumed to be $n_r / 2$.

Size Estimation of Complex Selections

selectivity of 0.1 means
that 10% of the tuples are
getting returned

- The **selectivity** of a condition θ_i is the probability that a tuple in the relation r satisfies θ_i .
 - If s_i is the number of satisfying tuples in r , the selectivity of θ_i is given by s_i/n_r .
- Conjunction:** $\sigma_{\theta_1 \wedge \theta_2 \wedge \dots \wedge \theta_n}(r)$. Assuming independence, estimate of

tuples in the result is:
$$n_r * \frac{s_1 * s_2 * \dots * s_n}{n_r^n}$$

not always true, but we
want estimates

- Disjunction:** $\sigma_{\theta_1 \vee \theta_2 \vee \dots \vee \theta_n}(r)$. Estimated number of tuples:

$$n_r * \left(1 - \left(1 - \frac{s_1}{n_r} \right) * \left(1 - \frac{s_2}{n_r} \right) * \dots * \left(1 - \frac{s_n}{n_r} \right) \right)$$

we expect
it to be the
"not" of the
conjunction

⇒ 1-selectivities
multiplied by
each other

- Negation:** $\sigma_{\neg \theta}(r)$. Estimated number of tuples:

$n_r - \text{size}(\sigma_{\theta}(r))$
or 1-selectivity

Join Operation: Running Example

→ depending on what we are talking about

Running example: $student \bowtie takes$

Catalog information for join examples:

- $n_{student} = 5,000$. $f_{student} = 50$, which implies that $b_{student} = 5000/50 = 100$.
- $n_{takes} = 10000$. $f_{takes} = 25$, which implies that $b_{takes} = 10000/25 = 400$.
- $V(ID, takes) = 2500$, which implies that on average, each student who has taken a course has taken 4 courses.
 - Attribute ID in $takes$ is a foreign key referencing $student$.
 - $V(ID, student) = 5000$ (primary key!) → doesn't repeat

```
create table student
(ID          varchar(5),
 name        varchar(20) not null,
 dept_name   varchar(20),
 tot_cred    numeric(3,0) check (tot_cred >= 0),
 primary key (ID),
 foreign key (dept_name) references department (dept_name)
 on delete set null
);
```

```
create table takes
(ID          varchar(5),
 course_id   varchar(8),
 sec_id      varchar(8),
 semester    varchar(6),
 year        numeric(4,0),
 grade       varchar(2),
 primary key (ID, course_id, sec_id, semester, year),
 foreign key (course_id, sec_id, semester, year) references section
(course_id, sec_id, semester, year)
 on delete cascade,
 foreign key (ID) references student (ID)
 on delete cascade
);
```

Estimation of the Size of Joins

- The Cartesian product $r \times s$ contains $n_r \cdot n_s$ tuples; each tuple occupies $s_r + s_s$ bytes.
- If $R \cap S = \emptyset$, then $r \bowtie s$ is the same as $n_r \times n_s$.
- If $R \cap S$ is a key for R , then a tuple of s will join with at most one tuple from r
 - therefore, the number of tuples in $r \bowtie s$ is no greater than the number of tuples in s .
- If $R \cap S$ in S is a foreign key in S referencing R , then the number of tuples in $r \bowtie s$ is exactly the same as the number of tuples in s .
 - The case for $R \cap S$ being a foreign key referencing S is symmetric.
- In the example query $student \bowtie takes$, ID in $takes$ is a foreign key referencing $student$
 - hence, the result has exactly n_{takes} tuples, which is 10000

Estimation of the Size of Joins (Cont.)

- If $R \cap S = \{A\}$ is not a key for R or S .
If we assume that every tuple t in R produces tuples in $R \bowtie S$, the number of tuples in $R \bowtie S$ is estimated to be:

$$\frac{n_r * n_s}{V(A,s)}$$

If the reverse is true, the estimate obtained will be:

$$\frac{n_r * n_s}{V(A,r)}$$

- The lower of these two estimates is probably the more accurate one.
- Can improve on above if histograms are available
 - Use formula similar to above, for each cell of histograms on the two relations

Estimation of the Size of Joins (Cont.)

- Compute the size estimates for *student* ⋈ *takes* without using information about foreign keys:
 - $V(ID, takes) = 2500$, and
 $V(ID, student) = 5000$
 - The two estimates are $5000 * 10000/2500 = 20,000$ and $5000 * 10000/5000 = 10000$
 - We choose the lower estimate, which in this case, is the same as our earlier computation using foreign keys.

The Internals of PostgreSQL

Chapter 3

Query Processing

<https://www.interdb.jp/pg/pgsql03.html>

Postgres optimizer code snippets

Postgres genetic query optimizer

<https://www.postgresql.org/docs/13/geqo-intro.html>

https://doxygen.postgresql.org/geqo_8h_source.html

var=const selectivity

https://doxygen.postgresql.org/selfuncs_8h.html#a31ee9824c23028c56ca3d6ca92c39a7e

Range typanalyze

https://doxygen.postgresql.org/rangetypes_typanalyze_8c_source.html

Range overlap

https://github.com/postgres/postgres/blob/cd3f429d9565b2e5caf0980ea7c707e37bc3b317/src/include/catalog/pg_operator.dat#L3110

rangesel

https://doxygen.postgresql.org/rangetypes_selfuncs_8c.html#a632d39f45c72d18cf792fb33014155ee

Selectivity Estimation of Inequality Joins In Databases

[Diogo Repas](#), [Zhicheng Luo](#), [Maxime Schoemans](#), [Mahmoud Sakr](#)

<https://arxiv.org/abs/2206.07396>

Credits

Many slides in this lecture are taken from:

- Avi Silberschatz, Henry F. Korth, S. Sudarshan. Database System Concepts

Recommended reading

- The Internals of PostgreSQL (<https://www.interdb.jp/pg/>)

Selectivity Estimation of Inequality Joins In Databases

Diogo Repas

Université libre de Bruxelles (ULB)
Brussels, Belgium
diogo.seca.repas.goncalves@ulb.be

Maxime Schoemans

Université libre de Bruxelles (ULB)
Brussels, Belgium
maxime.schoemans@ulb.be

Zhicheng Luo

Université libre de Bruxelles (ULB)
Brussels, Belgium
zhicheng.luo@ulb.be

Mahmoud Sakr

Université libre de Bruxelles (ULB)
Brussels, Belgium
Ain Shams University
Cairo, Egypt
mahmoud.sakr@ulb.be

ABSTRACT

Selectivity estimation refers to the ability of the SQL query optimizer to estimate the size of the results of a predicate in the query. It is the main calculation, based on which the optimizer can select the cheapest plan to execute. While the problem is known since the mid 70s, we were surprised that there are no solutions in the literature for the selectivity estimation of inequality joins. By testing four common database systems: Oracle, SQL-Server, PostgreSQL, and MySQL, we found that the open-source systems PostgreSQL and MySQL lack this estimation. Oracle and SQL-Server make fairly accurate estimations, yet their algorithms are secret. This paper thus proposes an algorithm for inequality join selectivity estimation. The proposed algorithm has been implemented in PostgreSQL and sent as a patch to be included in the next releases.

1 INTRODUCTION

Query optimization is the overall process of generating the most efficient query plan given an SQL statement. The query optimizer, responsible for this process, applies equivalence rules to reorganize and merge the operations in the query to find the fastest execution plan and feeding it to the executor. It examines multiple access methods, such as sequential table scans or index scans, different join methods such as nested loops and hash joins, different join orders, sub-query normalization, materialized views, and other possible transformations. Starting with a naively generated query plan, the optimizer generates a set of equivalent plans. To choose the most efficient plan, almost all systems adopt a cost-based approach, which roots back in the architecture of System R [2] and Volcano/Cascades [5, 6].

In cost-based query optimization, the optimizer estimates the cost of the alternative query plans and chooses the plan with minimum cost. The cost is estimated in terms of the CPU and I/O resources that the query plan will use. A central component in cost estimation is the Selectivity Estimation (SE). SE collects statistics for all attributes in a relation, such as data distribution histograms, most common values, null percentage, etc. These statistics are then used during planning time to estimate the number of tuples generated by a predicate in the query. A smaller selectivity value means a smaller size for intermediate results, which is favorable for a more efficient execution. The cost-based optimizer thus reorders the selection and join predicates to quickly reduce the sizes of intermediate results.

Since, in general, the cost of each operator depends on the size of its input relations, it is important to provide good estimations of their selectivity, that is, of their result size, to the query optimizer [16]. Inaccurate selectivity estimations can lead to inefficient query plans being chosen, sometimes leading to orders of magnitude longer execution times [13].

There is a trade-off between the size of the stored statistics and the complexity of the estimation algorithm on the one hand, and the estimation accuracy on the other. Recent research thus focuses on using machine learning methods to capture the data distribution into compact models. While there are good results in this research direction [23], common relational database systems continue to use traditional statistics structures, mostly based on histograms. A histogram can be used as a discrete approximation of the probability density function of an attribute.

Despite the popularity of histograms, there is a lack of theory on how to use them in estimating inequality join selectivity. This paper aims at filling this gap, and presents the following main contributions:

- A novel algorithm for join selectivity estimation of inequality operators using histogram statistics
- The implementation of this algorithm in PostgreSQL both for scalar inequality joins, as well as for multiple operators of range types
- An extension of the algorithm that also takes advantage of additional statistics, when available.
- The proposed algorithm has been implemented in PostgreSQL and submitted as a patch for inclusion in a future release¹

Section 2 starts by reviewing existing work in selectivity estimation. A running example to be used throughout the paper is given in Section 3.1. Then, some definitions, notation and terminology are introduced in Section 3.2. The algorithm presented in this paper consists of mapping the problem of selectivity estimation to a probability theory problem, as described in Section 4. The algorithm of join selectivity estimation is developed in Section 5. The section also develops a way to incorporate null values and Most Common Values (MCV) statistics in the estimation model. A mapping of this model for several range operators is also presented. Finally, an implementation in PostgreSQL and the experimental evaluation are provided in Section 6.

¹<https://github.com/DRepas/postgres/tree/rangejoinsel>

2 REVIEW OF RELATED WORK

A survey on DBMS query optimizer has been proposed in 2021 [13], which categorizes cardinality estimation methods into synopsis-based methods, sampling-based methods, and learning-based methods. Many learning-based methods [11, 12, 24] have been proposed in recent years and show better accuracy than traditional methods. But there are still many missing parts to be solved to put them into real systems, such as the cost of model training and updating, and the black-box property of learning algorithms [23]. Sampling-based methods estimate selectivity by executing a (sub)query on samples collected from tables, whose accuracy depends on the degree to which the samples fit the original data distribution [13]. These methods, however, suffer from a high cost of storage and retrieval time, especially when the tables are very large. Another limitation of sampling-based methods is that they currently only support equality join selectivity estimation [13]. Histograms, as a form of synopsis-based methods, have been extensively studied [11, 13] and are widely adopted in common database systems [3] for the purpose of selectivity estimation, including MySQL, PostgreSQL, Oracle, and SQL Server [10, 15, 19, 20].

MySQL uses two histogram types for selectivity estimation [20]. One is the singleton histogram which stores the distinct values and their cumulative frequency. Another is the equi-depth histogram, called equi-height in MySQL documentation. This histogram stores the lower and upper bounds, cumulative frequency, and the number of distinct values for each bucket. However, the usage of histograms is limited to restriction selectivity estimation [20], i.e., the selection operator. For join selectivity estimation, MySQL naively returns a constant: 0.1 for equality joins and 0.3333 for inequality joins [22]. The following is the excerpt of the MySQL source code in which these constants are defined [21]:

```
/// Filtering effect for equalities : col1 = col2
#define COND_FILTER_EQUALITY 0.1f
/// Filtering effect for inequalities : col1 > col2
#define COND_FILTER_INEQUALITY 0.3333f
```

PostgreSQL also uses histogram as optimizer statistics [10]. By analyzing its manual [7], as well as its source code, it uses equi-depth histograms. In contrast to MySQL, the number of distinct values in each bucket is not stored. For this reason, PostgreSQL does not use these histogram statistics in estimating equi-join selectivity. It rather uses a singleton histogram of Most Common Values (MCV) [7]. As for inequality join selectivity estimation ($<$, \leq , $>$, \geq), a default constant value of 0.3333 is returned [8].

The following is an excerpt of the PostgreSQL source code in which these constants are defined [9]:

```
/* default selectivity estimate for equalities such as "A = b" */
#define DEFAULT_EQ_SEL 0.005
/* default selectivity estimate for inequalities such as "A < b" */
#define DEFAULT_INEQ_SEL 0.3333333333333333
```

Oracle Database uses three types of histograms to capture the data distribution of a relation's attribute [15]: singleton histograms (referred to as frequency histogram and top frequency histogram in the official documentation), equi-depth histogram (referred to as height-balanced in the official documentation), and hybrid histogram (a combination of equi-depth and frequency histograms).

The type of histogram is determined based on specific criteria to fit different situations. The official documentation [15] also states some factors behind their selectivity estimation algorithms, such as endpoint numbers (the unique identifier of a bucket, e.g., the cumulative frequency of buckets in frequency and hybrid histograms) and values (the highest value in a bucket), and whether column values are popular (an endpoint value that appears multiple times in a histogram) or non-popular (every column value that is not popular). However, the details of these estimation algorithms are not published. Few online articles, in the form of hacker blogs, did experimental analyses to guess how selectivity estimation works in Oracle Database but didn't yield a clear algorithm, [4, 14].

SQL-Server is another popular closed-source DBMS. Due to its proprietary nature, implementation details are scarce. According to the official documentation [19], a proprietary kind of histogram with a density vector associated is built in three steps for each attribute. The official documentation [17, 18] describes four core assumptions for the selectivity estimation: independence when no correlation information is available, uniformity in histogram bins, inclusion when filtering a column with a constant, and containment when joining distinct values from two histograms [1]. Although the white paper [17] is a publication from SQL Server that deals with the problem of selectivity estimation, it does not explain the algorithm used for join selectivity. Similar to Oracle, the implemented algorithm is a secret.

To identify if any informed selectivity estimation is taking place when performing inequality joins in SQL-Server and Oracle, we have performed the following experiment. Two different attributes, T1 and T2, with 1000 and 200 rows respectively, were randomly generated by sampling the range [0, 100] uniformly. They were then joined using the $<$ (less than) operator. Both databases made a quite accurate selectivity estimation of this inequality join. Oracle Database had an estimation error of 3%, and SQL Server had a smaller error of just 0.29%. As such, we know that both systems implement good estimation algorithms.

In conclusion, although learning-based methods have become a popular research direction for selectivity estimation in recent years, histograms are still the most commonly used statistics in existing DBMS for this purpose. The recurring types of used histogram statistics are equi-depth histograms approximating the distribution of values, and singleton histograms of Most Common Values. As our investigation indicates, MySQL and PostgreSQL don't have algorithms implemented for join selectivity estimation, and they use predefined constants. On the other hand, popular commercial DBMS (SQL-Server and Oracle) have implemented some algorithms based on the histograms, but we couldn't find any source describing them. This paper addresses this gap by proposing such an algorithm.

3 PRELIMINARIES

This paper presents a formal model to reason about two different selectivity estimation types:

- Restriction selectivity estimation: when one of the sides of the operator is an attribute of a relation and the other is a constant value.
 - Example: `SELECT * FROM R1 WHERE R1.X < 100`

- Join selectivity estimation: when both sides of the operator are attributes of different relations.
– Example: `SELECT * FROM R1, R2 WHERE R1.X < R2.Y`

Selectivity estimation of operations where both sides of the operator are attributes of the same relation, with no join or Cartesian product involved (Example: `SELECT * FROM R1 WHERE R1.x < R1.y`), is not addressed by this paper.

The selectivity of an operator is the fraction of values kept in the result after selection. In the case of restriction selectivity, the denominator is the input relation's size. In the case of join selectivity, the denominator is the input relations' Cartesian product size (their sizes multiplied). This fraction can be interpreted as the probability that a given randomly selected tuple from the input relation, or from the Cartesian product of input relations in the case of joins, is selected by the operator being considered.

The focus of the next sections will be on the restriction selectivity estimation of the *less than* ($<$) operator. The restriction selectivity estimation of all scalar inequality operators will be derived from this initial estimation. We will also build/generalize on it to develop the join selectivity estimation. This restriction selectivity estimation in the next section is already implemented by all common database systems, thus not a novel contribution of this work. We however formulate it as a probability problem, and develop the join selectivity estimation on top of it, to maximize the code reuse in these systems.

The attributes being restricted or joined will be treated as random variables that follow a distribution modeled by a Probability Density Function (PDF) and/or a Cumulative Distribution Function (CDF).

3.1 Running Example

For demonstration purposes, relations R1 and R2 will be used throughout this paper. For each relation, 12 integers were manually selected to cover as many corner cases as possible when using equi-depth histograms (introduced in section 3.2), such as skew and common bin boundaries.

$R1.X = \{10, 11, 12, 20, 21, 22, 24, 25, 30, 35, 38, 45\}$

$R2.Y = \{15, 16, 17, 20, 30, 35, 38, 39, 40, 42, 45, 50\}$

3.2 Histogram Statistics

Histograms are commonly used to approximate the PDF of an attribute by grouping the values in uniform bins. Each bin is an interval of the form $B_j =]hist_j, hist_{j+1}]$, where $hist_i$ are values from the domain of the attribute. It is important to note that the side on which the interval is open or closed is not relevant for the purposes of this paper, as all estimations correspond to integration over a continuous domain, where singular points do not affect the final result. By defining a bin this way (using intervals), this paper is restricting itself to domains where total order exists. This excludes categorical data, for which the idea of an equi-depth histogram does not apply.

Let $H_j(X) = P(X \in B_j)$ be the fraction of values of the attribute X that is represented by the histogram bin B_j , i.e., the height/depth of the histogram bin:

- *Singleton histograms* are such that each bin refers to the frequency of a single element. Typically used to collect Most Common Values statistics (introduced in section 5.3).

- *Equi-width histograms* are such that each bin has the same width. That is, $hist_{j+1} - hist_j$ is constant.
- *Equi-depth histograms* are such that each bin has the same depth (height) but varying width. That is $\forall j, P(X \in B_j) = H_j(X) = \frac{1}{n}$, where n is the total number of bins.

For selectivity estimation, equi-depth histograms are favoured because the resolution of the histogram adapts to skewed value distributions. Typically, the histogram is smaller than the original data. Thus, it cannot represent the true distribution entirely and some assumptions are induced, e.g., uniformity on a single attribute, and independence assumption among different attributes. The use of equi-depth histograms is also motivated by their prevalence in current RDBMS. These are usually constructed through the use of random sampling of the original relations.

For demonstration purposes, Figure 1 shows a graphical representation of equi-depth histograms for the attributes $R1.X$ and $R2.Y$ of the running example. For both histograms, there are three bins, meaning that the fraction accounted for by each bin is $\frac{1}{3}$. For attribute $R1.X$, $hist_X = [10, 20, 25, 45]$, which means that $B_{X0} = [10, 20]$, $B_{X1} =]20, 25]$, and $B_{X2} =]25, 45]$. For attribute Y , $hist_Y = [15, 20, 38, 50]$, which means that $B_{Y0} = [15, 20]$, $B_{Y1} =]20, 38]$, and $B_{Y2} =]38, 50]$.

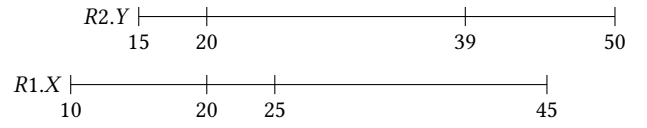


Figure 1: Equi-depth histograms of $R1.X$ and $R2.Y$ with 3 bins each.

Using histograms as a statistical representation of attributes involves the following implicit assumptions:

- The data is distributed uniformly inside each bin
- The histograms are complete (they account for all the data points), that is:
 - $hist_0 = \min(X)$
 - $hist_n = \max(X)$

In practice, these two assumptions do not strictly hold. The data is usually not uniformly distributed inside each bin. The more bins used in the histograms, the smaller the error introduced by this assumption. Database systems, e.g., PostgreSQL, typically create the histogram using a random sample of the attribute values, especially when the number of tuples is too large. The assumption of completeness of the histogram might be broken in the presence of sampling. When the sample is representative of the underlying data, the estimation is still fairly accurate.

Given the equi-depth histogram of an attribute X , with n bins, one can derive its approximate PDF and CDF as shown next. Let $f_X(c)$ and $F_X(c)$ denote the PDF and the CDF of X , respectively, at a given value c then:

$$f_X(c) = \begin{cases} 0 & c < hist_0 \\ \frac{1}{n(hist_{j+1} - hist_j)} & c \in B_j, 0 \leq j < n \\ 0 & c \geq hist_n \end{cases} \quad (1)$$

$$F_X(c) = \begin{cases} 0 & c < hist_0 \\ \frac{1}{n}(j + \frac{c - hist_j}{hist_{j+1} - hist_j}) & c \in B_j, 0 \leq j < n \\ 1 & c \geq hist_n \end{cases} \quad (2)$$

When $c \in B_j, 0 \leq j < n$, formula 1 is derived from the definition of equi-depth histogram, where each bin represents $H_j(X) = \frac{1}{n}$ of the data, spread over a width of $hist_{j+1} - hist_j$.

Formula 2 is derived from the following:

$$F_X(c) = \int_{-\infty}^c f_X(x) dx = \int_{-\infty}^{hist_0} f_X(x) dx + \sum_{i=0}^{j-1} \int_{hist_i}^{hist_{i+1}} f_X(x) dx + \int_{hist_j}^c f_X(x) dx$$

where,

$$\begin{aligned} \int_{-\infty}^{hist_0} f_X(x) dx &= 0 \\ \sum_{i=0}^{j-1} \int_{hist_i}^{hist_{i+1}} f_X(x) dx &= \sum_{i=0}^{j-1} H_j(X) = \sum_{i=0}^{j-1} \frac{1}{n} = \frac{j}{n} \\ \int_{hist_j}^c f_X(x) dx &= H_j(X) \frac{c - hist_j}{hist_{j+1} - hist_j} dx = \frac{1}{n} \frac{c - hist_j}{hist_{j+1} - hist_j} \end{aligned}$$

This last formula performs linear interpolation within the bin where c is contained, thus assuming a uniform distribution of values within the bin. The assumption that the histogram is complete is reflected in substituting the infinite bounds by $hist_0, hist_n$. In the running example, the PDF and CDF of $R1.X$ can be derived from the formulas presented in this section as follows:

$$f_X(c) = \begin{cases} 0 & c < 10 \\ \frac{1}{30} & 10 \leq c < 20 \\ \frac{1}{15} & 20 \leq c < 25 \\ \frac{1}{60} & 25 \leq c < 45 \\ 0 & 45 \leq c \end{cases}$$

$$F_X(c) = \begin{cases} 0 & c < 10 \\ \frac{1}{30}c - \frac{1}{3} & 10 \leq c < 20 \\ \frac{1}{15}c - 1 & 20 \leq c < 25 \\ \frac{1}{60}c + \frac{1}{4} & 25 \leq c < 45 \\ 1 & 45 \leq c \end{cases}$$

4 A FORMAL MODEL FOR SELECTIVITY ESTIMATION

We first start by formalizing the problem of restriction selectivity estimation for the *Less Than* ($<$) operator. Suppose the goal of estimating the selectivity of the following operation (expressed in SQL):

```
SELECT *
FROM R1
WHERE R1.X < c
```

where c is a constant. Treating the attribute $R1.X$ as a random variable X , estimating the selectivity of the above operation is equivalent to finding $P(X < c)$. Given the PDF or the CDF of X , f_X or F_X , respectively, the selectivity of the operation above can be formalized as:

$$P(X < c) = \int_{-\infty}^c f_X(x) dx = F_X(c) \quad (3)$$

Suppose now that the goal is estimation the join selectivity for the *Less Than* ($<$) operator. That is, we want to estimate the selectivity of the following operation (expressed in SQL):

```
SELECT *
FROM R1, R2
WHERE R1.X < R2.Y
```

Treating the attributes $R1.X$ and $R2.Y$ as random variables X and Y , respectively, estimating the selectivity of the above operation can be formulated as finding $P(X < Y)$.

Consider the joint distribution of X and Y , $P(X, Y)$. The probability that a sample (a, b) taken at random from the Cartesian product of the values in $R1.X$ and $R2.Y$ can be defined as follows:

$$\forall a \in R1.X, b \in R2.Y, P(X = a, Y = b) = P(X = a) \times P(Y = b)$$

or equivalently:

$$P(X, Y) = P(X)P(Y)$$

which is the definition of independent random variables. Note that when a Cartesian product is involved, either explicitly as in the SQL statement above, or implicitly through a join clause, the two variables are independent.

Given a joint PDF of X and Y , $f_{X,Y}$. With X and Y being independent, it is known that $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, with f_X and f_Y the PDFs of X and Y , respectively. Considering F_X to be the CDF of X , the selectivity of the *less than* ($<$) operator can be formalized as follows:

$$\begin{aligned} P(X < Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^y f_{X,Y}(x, y) dx dy = \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^y f_X(x)f_Y(y) dx dy = \\ &= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^y f_X(x) dx \right) f_Y(y) dy = \\ &= \int_{-\infty}^{+\infty} F_X(y)f_Y(y) dy \end{aligned} \quad (4)$$

This formula thus presents a solution for estimating the join selectivity estimation. Next, we discuss how to translate it into an algorithm.

5 IMPLEMENTATION IN A DATABASE SYSTEM

In RDBMS implementations, histograms are used as a discrete approximation of the PDF and CDF of attributes. This section maps the theory above into an implementable solution in databases using equi-depth histograms.

5.1 Selectivity Estimation

Restriction selectivity estimation. Recall that restriction selectivity estimation is about estimating the selectivity of a predicate in the following form:

```
SELECT *
FROM R1
WHERE R1.X < c
```

As described in section 4, restriction selectivity estimation can be calculated using the CDF of X (equation 3). Deriving the CDF from an equi-depth histograms is presented in equation 2. To find the bin, B_j , where c is contained, one can perform a binary search over the histogram boundaries.

Algorithm 1 illustrates the estimation of restriction selectivity. The *hist* array represents the equi-depth histogram, stored as an ordered array of bin boundaries. The function *binary_search* returns the greatest index in this array that is less than or equal to a given constant, effectively finding the bin where such constant falls into. The rest of the algorithm computes the CDF at c using equation 2.

Algorithm 1: Restriction selectivity estimation for the expression $R1.X < c$

```
Input: hist an array of length  $n$  representing the
        equi-depth histogram of  $R1.X$ ,  $c$  the scalar literal
        in the query
Output: the estimated selectivity
/* Identify preceding whole bins */
j ← binary_search(hist, c);
/* Corner cases */
if j < 0 then
  return 0
if j ≥ n - 1 then
  return 1
/* Estimate using preceding bins */
selectivity ← j / (n - 1);
/* Adjust using linear interpolation */
selectivity += (c - hist[j]) / (hist[j+1] - hist[j]) / (n - 1);
return selectivity
```

Using the running example, and the following query:

```
SELECT *
FROM R1
WHERE R1.X < 30
```

this query yields 8 rows, which corresponds to a selectivity of $\frac{2}{3}$. Using Algorithm 1 with the histograms presented in the running example, the number 30 will be found in B_{X2} , meaning that the estimated selectivity will be $\frac{2}{3} + \frac{1}{3} \frac{30-25}{20} = \frac{9}{12}$. After multiplying by the attribute's cardinality (12), we get the estimated row count of 9, which is a close estimate to the actual result size.

Figure 2 shows a graphical depiction of the PDF of the $R1.X$, which is directly obtainable from its equi-depth histogram. The integral in equation 2, as well as the estimation calculated above using Algorithm 1, correspond to the highlighted area in the figure.

Join selectivity estimation. Join selectivity estimation is mapped into a double integral involving the two PDFs. Equation 4 illustrates

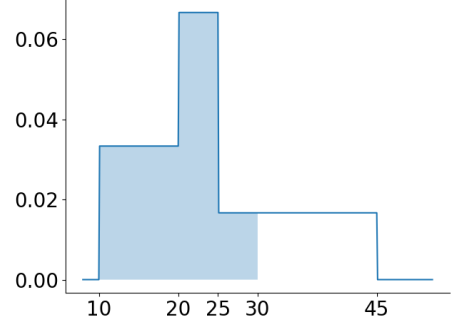


Figure 2: Restriction Selectivity Estimation of $R1.X < 30$

that join selectivity can be estimated by using the CDF of X and the PDF of Y . These can be calculated using equations 1 and 2.

The CDF of X is linear piece-wise, each piece is defined in a bin of X 's histogram. The PDF of Y is a step function, i.e., constant piece-wise, where each piece is defined in a bin of Y 's histogram. This leads to the conclusion that their product, which is needed in equation 4, is a linear piece-wise function, with every piece being defined in an intersection of X and Y 's bins (see Figure 3 for a graphical depiction of this using the running example introduced in section 3.1). By merging the bounds of the two histograms of X and Y in a single sorted array, *sync*, the intersections of X and Y 's bins will be of the form $[sync_j, sync_{j+1}]$.

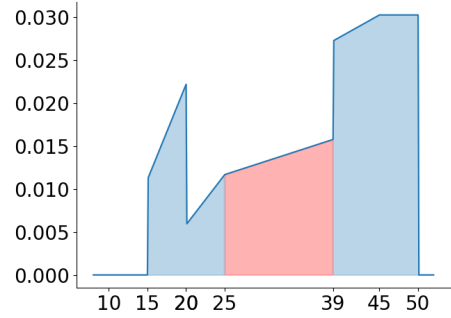


Figure 3: Piece of product $F_X \cdot f_Y$ used for Join Selectivity Estimation

From equation 1, we know that f_Y is 0 for all values until first value of $hist_Y$, and it is 0 after the last value of $hist_Y$. From equation 2, we know that F_X is 0 for all values until first value of $hist_X$, and it is 1 after the last value of $hist_X$.

Analyzing the product $F_X \cdot f_Y$ piece-wise:

$$(F_X \cdot f_Y)(c) = \begin{cases} 0 & c \leq \max(hist_{X0}, hist_{Y0}) \\ F_X(c) \cdot f_Y(c) & \max(hist_{X0}, hist_{Y0}) < c \\ < \min(hist_{X_{n_X-1}}, hist_{Y_{n_Y-1}}) \\ f_Y(c) & hist_{X_{n_X-1}} \leq c < hist_{Y_{n_Y-1}} \\ 0 & c \geq \max(hist_{X_{n_X-1}}, hist_{Y_{n_Y-1}}) \end{cases} \quad (5)$$

Given that the product, $F_X \cdot f_Y$, in equation 4, is linear in each interval of the form $[sync_j, sync_{j+1}]$, equation 4 can be discretized as follows:

$$\begin{aligned}
 P(X < Y) &= \int_{-\infty}^{+\infty} (F_X \cdot f_Y)(y) dy = \\
 &= \sum_{k=0}^{n_X+n_Y-1} \int_{sync_k}^{sync_{k+1}} (F_X \cdot f_Y)(y) dy = \\
 &= \sum_{k=0}^{n_X+n_Y-1} \frac{(F_X \cdot f_Y)(sync_k) + (F_X \cdot f_Y)(sync_{k+1})}{2} (sync_{k+1} - sync_k)
 \end{aligned} \tag{6}$$

To maximize code re-use, it is possible to reorganize this equation into a sum of the CDFs of both X and Y , i.e., so that we can reuse Algorithm 1. The following is derived directly from equation 3:

$$\begin{aligned}
 f_Y(sync_k)(sync_{k+1} - sync_k) &= \\
 f_Y(sync_{k+1})(sync_{k+1} - sync_k) &= \\
 \int_{sync_k}^{sync_{k+1}} f_Y(y) dy &= \\
 F_Y(sync_{k+1}) - F_Y(sync_k)
 \end{aligned} \tag{7}$$

the first two steps rely on the fact that f_Y is constant in the interval $[sync_k, sync_{k+1}]$.

Equation 6 can now be re-written using only CDFs of X and Y as follows:

$$\begin{aligned}
 P(X < Y) &= \\
 \sum_{k=0}^{n_X+n_Y-1} \frac{(F_X \cdot f_Y)(sync_k) + (F_X \cdot f_Y)(sync_{k+1})}{2} (sync_{k+1} - sync_k) &= \\
 \frac{1}{2} \sum_{k=0}^{n_X+n_Y-1} (F_X(sync_k) + F_X(sync_{k+1}))(F_Y(sync_{k+1}) - F_Y(sync_k))
 \end{aligned} \tag{8}$$

Algorithm 2 illustrates the estimation of join selectivity of the *less than* ($<$) operator using this equation. It thus has the advantage of re-using algorithm 1 to compute F_X and F_Y . The creation of the *sync* array in Algorithm 2 comes at the expense of time and space of $O(n_X + n_Y)$, for duplicating and merging the two sorted histograms. To optimize this, the two histograms can be scanned in parallel, without the need to materialize the *sync* array. This also allows for further optimization. The algorithm only needs to iterate over the overlapping region of both histograms. All the partial sums before that will be zero, as can be verified in Figure 4. After the overlapping region the remaining partial sums are equal to what is left of the histogram of Y , $1 - cur_{F_Y}$, because all remaining values of Y will be greater than the maximum value in X . We adopt these optimizations in our implementation, yet we omit them here for the clarity of presentation.

Algorithm 2: Join selectivity estimation algorithm for the *less than* ($<$) operator re-using Algorithm 1 as F_X and F_Y

Input:

histogram of X , $hist_X$, is an array of length n_X

histogram of Y , $hist_Y$, is an array of length n_Y

Output: Join selectivity estimation of $X < Y$

selectivity $\leftarrow 0$;

sync \leftarrow merge_sorted($hist_X$, $hist_Y$);

cur_F_X $\leftarrow F_X(sync[0])$; // always zero

cur_F_Y $\leftarrow F_Y(sync[0])$; // always zero

for $k \leftarrow 1$ **to** $n_X + n_Y - 1$ **do**

 next_F_X $\leftarrow F_X(sync[k])$; // using Algorithm 1

 next_F_Y $\leftarrow F_Y(sync[k])$; // using Algorithm 1

 selectivity $+= (cur_F_X + next_F_X) * (next_F_Y - cur_F_Y)$;

 cur_F_X $\leftarrow next_F_X$;

 cur_F_Y $\leftarrow next_F_Y$;

return selectivity / 2

The goal of Algorithm 2 is to calculate the area under the curve of the product $F_X \cdot f_Y$, represented in Figure 4. Taking the 3-bin histograms calculated in section 3.2 from the running example attributes $R1.X$ and $R2.Y$, a materialized *sync* array would have the values [10, 15, 20, 25, 39, 45, 50] after merging, sorting and removing duplicates. These correspond to the boundaries of the pieces in which the product $F_X \cdot f_Y$ is linear. Stepping through Algorithm 2 with k from 1 to 5, corresponding to the 6 pieces represented by the *sync* array, we arrive at the following sum:

$$\begin{aligned}
 P(R1.X < R2.Y) &= \frac{1}{2} (\\
 &= (0 + \frac{1}{6}) \times (0 - 0) + (\frac{1}{6} + \frac{1}{3}) \times (\frac{1}{3} - 0) + \\
 &= (\frac{1}{3} + \frac{2}{3}) \times (\frac{8}{19} - \frac{1}{3}) + (\frac{2}{3} + \frac{9}{10}) \times (\frac{2}{3} - \frac{8}{19}) + \\
 &= (\frac{9}{10} + 1) \times (\frac{28}{33} - \frac{2}{3}) + (1 + 1) \times (1 - \frac{28}{33}) \\
 &= \frac{24221}{37620} \approx 0.643833
 \end{aligned}$$

The correct result will have 95 rows, which corresponds to a selectivity of $\frac{95}{144} \approx 0.659722$. After multiplying by the cardinality of the Cartesian product of both attributes (144) and rounding the result to the nearest integer, we get the estimated row count of $92.712 \approx 93$, which is close to the correct result size.

Figure 4 shows a graphical depiction of the product of the CDF of $R1.X$ and the PDF of $R1.Y$, which is directly obtainable from their equi-depth histograms. The integral in equation 6, as well as the estimation calculated above using Algorithm 2, correspond to the area under the curve in the figure.

Figure 4 illustrates the multiplication of the CDF($R1.X$) and the PDF($R2.Y$). The integral in equation 4, as well as its estimation in Algorithm 4, correspond to calculating the area under the curve in the figure.

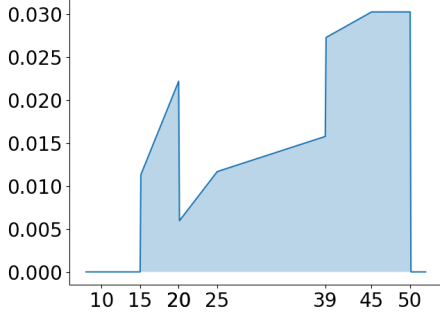


Figure 4: Product $F_X \cdot f_Y$ used for Join Selectivity Estimation

Note that the code re-use in algorithm 2 has a small performance impact. This algorithm has a time complexity of $O((n_X + n_Y) \log(n_X + n_Y))$ since it performs a binary search (twice) for each element of each histogram. This binary search is not necessary since the two histograms are being scanned sequentially and the current indices are known at each iteration. One way to avoid this overhead would be to optionally specify j as an input parameter of algorithm 1, thus reducing the time complexity to $O(n_X + n_Y)$.

5.2 Extending to all scalar inequality operators

Given the restriction and join selectivity estimators for the less than inequality, all scalar inequality operators can be implemented by noting the following equivalences:

Restriction selectivity:

- $P(X \geq c) = 1 - P(X < c)$
- $P(X > c) = 1 - P(X < c) - P(X = c)$
- $P(X \leq c) = P(X < c) + P(X = c)$

Join selectivity:

- $P(X \geq Y) = 1 - P(X < Y)$
- $P(X > Y) = P(Y < X)$
- $P(X \leq Y) = P(X < Y) + P(X = Y)$

Estimators for equality selections and joins are already implemented by almost all common systems. In case they are missing, one could assume $P(X = c)$ and $P(X = Y)$ to be zero, thus leading to under-/over-estimate the selectivity.

5.3 Making Use of Other Statistics

Typically, RDBMS will collect statistics about nulls, in the form of a fraction of null values, and Most Common Values (MCV), in the form of a singleton histogram. Histograms will thus be constructed for the remaining part of the data. When the histogram statistics only refer to a fraction of the data, the methods described up to this point only provide an estimation for this fraction. The final estimation must thus take nulls and MCV into account.

As a general way to integrate such other statistics in the estimation besides the histograms, we note the following: given a non-overlapping partitioning of the data, if each partition j corresponds to a fraction p_j of the original data, and selectivity within that partition is s_j , the final selectivity can be calculated by the inner product $p \cdot s$.

Figure 5: Nine cases for join selectivity estimation

	Y is null	Y in MCV	Y in histogram
X is null	case1	case2	case3
X in MCV	case4	case5	case6
X in histogram	case7	case8	case9

Since a value is either null, a most common value, or accounted for by the histogram, the overall restriction selectivity can be calculated by the following formula:

$$p \cdot s = p_{null}s_{null} + p_{mcv}s_{mcv} + p_{hist}s_{hist} \quad (9)$$

Null Values. All inequality operators are strict, this means that the selectivity of null values is 0. For this reason, the first term in equation 9 is also 0.

Most Common Values. MCV statistics maintain pairs of values and their frequencies in the table. They are maintained for the top k frequent values, where k is a statistics collection parameter. Since MCV represent the data in its original form, it is possible to accurately compute the selectivity for these values.

To estimate the restriction selectivity of an operator using the most common values, algorithm 3 can be used. This algorithm computes the selectivity of a Boolean operator on a list of most common values by adding the frequencies of the most common values satisfying this Boolean condition.

Algorithm 3: `mcv_selectivity(values, fractions, n, op)`, estimates the restriction selectivity, using only the MCV statistics, for a given Boolean operator `op`

Input:

MCV statistics of X (array of *values* and corresponding array of *fractions*)

Length of MCV arrays, n

constant, c

operator, op

Output: Restriction selectivity estimation of $X <op> c$

selectivity $\leftarrow 0$;

for $i \leftarrow 0$ **to** n **do**

if $op(values[i], c)$ **then**

 selectivity $+=$ fractions[i] ;

return selectivity

For join selectivity estimation, since there is a need to combine statistics from null values, MCV, and equi-depth histograms for both X and Y , there are 9 cases that need consideration, depending on the combination of values of X and Y , as shown in Figure 5.

For strict operators, such as inequalities, only cases 5, 6, 8, and 9 need to be calculated. This is because nulls result in empty joins. Case 9 has already been handled in Algorithm 2. For case 5, we iterate over the values in the MCV of Y . For each value, we multiply the fraction represented by that value in Y times the MCV restriction selectivity of the operator in question with the current value of the MCV of Y as the constant. This process is described in algorithm 4.

For cases 6 and 8, Algorithm 5 is used, by swapping the arguments. For each common value in the statistics of X , multiply its

Algorithm 4: Join selectivity estimation algorithm for any binary Boolean operator re-using algorithm 3 as `mcv_selectivity(values, fractions, n, op)`

Input:

MCV statistics of X (array of *values_X* and corresponding array of *fractions_X*)
 Length of MCV arrays of X, n_X
 MCV statistics of Y (array of *values_Y* and corresponding array of *fractions_Y*)
 Length of MCV arrays of Y, n_Y
 operator, *op*

Output: Join selectivity estimation of $X \langle op \rangle Y$

selectivity $\leftarrow 0$;

for $i \leftarrow 0$ **to** n **do**

selectivity += fractions_Y[i] *
 mcv_selectivity(values_X, fractions_X, values_Y[i],
 op);

return selectivity

fraction by the histogram restriction selectivity of Y using the current value of X as the constant.

Algorithm 5: Join selectivity estimation algorithm for the *less than* ($<$) operator re-using algorithm 1 as F_Y

Input:

MCV statistics of X (array of *values* and corresponding array of *fractions*)
 Length of MCV arrays, n_X
 histogram of Y, *hist_Y*
 operator, *op*

Output: Join selectivity estimation of $X < Y$

selectivity $\leftarrow 0$;

for $i \leftarrow 0$ **to** n_X **do**

selectivity += fractions[i] * $F_Y(\text{values}[i])$;

return selectivity

Given algorithms 4 and 5, the selectivity of the *less than* ($<$) operator considering histograms and most common values can be estimated by the following formula:

$$\begin{aligned} \text{hist_mcv_selectivity} = & p_{\text{hist}_X} p_{\text{hist}_Y} s_{\text{hist}_X \times \text{hist}_Y} \\ & + p_{\text{hist}_X} p_{\text{mcv}_Y} s_{\text{hist}_X \times \text{mcv}_Y} \\ & + p_{\text{mcv}_X} p_{\text{hist}_Y} s_{\text{mcv}_X \times \text{hist}_Y} \\ & + p_{\text{mcv}_X} p_{\text{mcv}_Y} s_{\text{mcv}_X \times \text{mcv}_Y} \end{aligned}$$

The final selectivity taking null values into account can be estimated as follows:

$$\text{selectivity} = (1 - p_{\text{null}}) \times \text{hist_mcv_selectivity} \quad (10)$$

5.4 Implementation for Ranges and Multi-Ranges

The algorithms described above are for scalar types. An advanced type, which is implemented by many database systems is the range type. A range type is a tuple (left, right, lc, rc), where left \leq right are two values of a domain with a total order. lc and rc specify whether respectively the left and right bounds are included in the range. The range type can be parameterized by the type of its bounds, e.g., range(float), range(timestamp), etc. In this section, we describe how the selectivity estimation in previous sections can be applied to the range type and the respective operators.

PostgreSQL, as an example of DBMS that has range types, collects the statistics for range attributes in the form of two equi-depth histograms: one for the lower bounds of the ranges, and one for the upper bounds. In the following, let X, Y be attributes of the same range type. Also, let $X.lower$ be the variable that represents all the lower bounds of X , $X.upper$ be the variable that represents all the upper bounds of X , and similarly for Y . Then it is possible to estimate the selectivity of the different range operators as follows:

- $P(X << Y) = P(X.upper < Y.lower)$, where $<<$, reads strictly left of, yields true when X ends before Y starts
- $P(X >> Y) = P(X.lower > Y.upper)$, where $>>$, reads strictly right of, yields true when X starts after Y ends
- $P(X \< Y) = P(X.upper < Y.upper)$, where $\<$, reads X does not extend to the right of Y , yields true when X ends before the end of Y
- $P(X \> Y) = P(X.lower < Y.lower)$, where $\>$, reads X does not extend to the left of Y , yields true when X starts before Y starts
- $P(X \&\& Y) = 1 - P(X << Y) - P(X >> Y)$, where $\&\&$ indicates the overlapping between X and Y
- and so on

It is however not possible to accurately estimate the join selectivity of the operators that express total or partial containment. This is mainly because the lower and upper histograms assume Independence between the range bounds. For containment operators, we need to relate the two bounds, which explicitly breaks this assumption.

Another consideration in estimating the selectivity of range operators is the fraction of empty ranges since these are not accounted for by the histograms. Depending on the operator, empty ranges are either always included or always excluded when compared to non-empty ranges and similarly when compared to other empty ranges.

6 EXPERIMENTS

This section evaluates the selectivity estimation accuracy of the proposed algorithm, and its relation to the size of the histogram statistics, i.e., the number of bins. Firstly, the proposed algorithm has been implemented in PostgreSQL 14, including support for range operators as described in section 5.4. We prepared a patch, and it is currently under review for inclusion in the next release of PostgreSQL. The batch is also included as an artifact with this paper.

The experiment described in this section is thus run using our implementation in PostgreSQL 15-develop on a Debian virtual Debian machine with 32GB of Disk and 8GB of RAM. We created two relations, R1 and R2, with range attributes R1.X and R2.Y, and cardinalities of 20390 and 20060 rows respectively. Note that for range types, we use the very same algorithm as for scalar types, so the results of this experiment hold for both.

The range values in the two relations were generated to cover a mixture of short, medium, and long ranges. We also included corner cases such as ranges with infinite bounds, empty ranges, and null values in the two relations. The following query was executed varying the number of histogram bins:

```
SELECT *
FROM R1, R2
WHERE x << y
```

Note that $<<$ denotes the "strictly left of" operator, which returns true if and only if the upper bound of x is less than the lower bound of y . PostgreSQL collects bounds histogram statistics for range attributes. To estimate the selectivity of the query above, histograms of the upper bound of X and lower bound of Y are used as input for Algorithm 2.

The optimizer statistics collector of PostgreSQL has a parameter called *statistics target*, that controls the number of bins in the equi-depth histograms. The default value is 100, and it can be increased up to 10000. The experiments described in this section were run by incrementing the statistics target by steps of 100 starting with the default value till the maximum value.

In the experiment, we observe two quantities: (1) the planning time, which is the time taken by the query optimizer to enumerate the alternative query plans, and estimate their costs, and (2) the cost estimation error defined as the absolute difference between the estimated and the actual number of rows returned by the query, divided by the cardinality of the Cartesian product of the two relations, which is 409013259.

Figure 6 shows the change in planning time (in milliseconds) as the statistics target increases. Apart from two outliers, the planning time shows approximately linear behavior, indicating that the binary search does not have a significant impact on it in the allowed range of statistics targets. Recall that the analytical complexity is $O((n_X + n_Y) \log(n_X + n_Y))$.

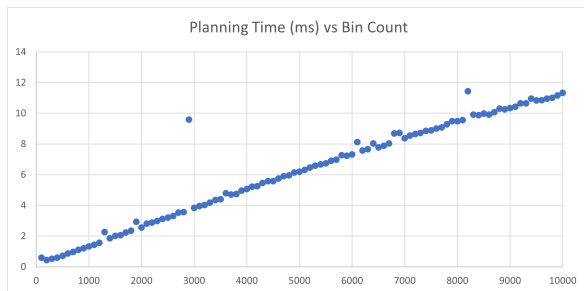


Figure 6: Query planning time (ms) V.S. the number of histogram bins in algorithm 2

Figure 7 shows the estimation error (in a logarithmic scale) against the number of histogram bins, i.e., by varying the statistics

target. As expected, using more bins leads to a lower estimation error. The largest error observed, when using only 100 histogram bins, was 1,112%. The error then drops rapidly to less than 0,002% at 900 bins, which corresponds to less than 5% of each relation size. The error stays consistently around this value for the histograms bigger than 900 bins.

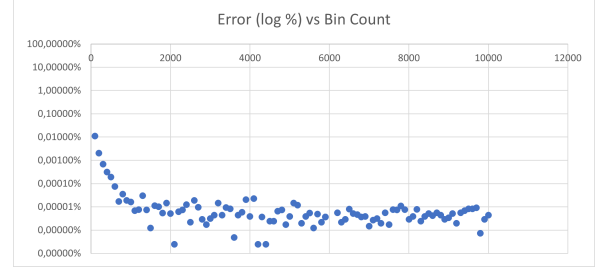


Figure 7: Log selectivity estimation error V.S. the number of histogram bins

Figure 8 plots the selectivity estimation error against the planning time in milliseconds. The significance of this figure is to illustrate the relation of the expenditure in terms of planning time versus the gain in terms of reduced error. This figure shows that, for the relations used, the planning time does not need to exceed 2 milliseconds to obtain extremely accurate estimations.

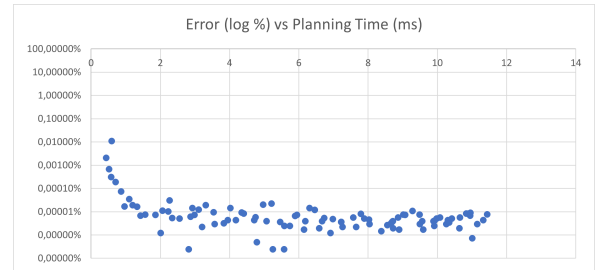


Figure 8: Log selectivity estimation error V.S. planning time (ms)

7 CONCLUSIONS

This paper proposed an algorithm for estimating the selectivity of inequality join predicates. It is a fundamental problem in databases that has not been solved before up to our knowledge. Common open-source databases, PostgreSQL and MySQL, lack implementations for this functionality. We have implemented and pushed the proposed algorithm as a patch to be included in PostgreSQL. Proprietary databases, Oracle and SQL-Server, return fairly accurate estimations, but their algorithms are not known. Our experiments show that the proposed algorithm provides comparable estimation accuracy, slightly more accurate. To produce these estimations, the algorithm uses equi-depth histogram statistics, which are adopted in all these systems. In a practical setting, the planning time remains within 2 milliseconds, which is the accepted norm in common databases.

REFERENCES

- [1] Nicolas Bruno and Surajit Chaudhuri. 2002. Exploiting statistics on query expressions for optimization. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 263–274.
- [2] Donald Chamberlin, Morton Astrahan, Mike Blasgen, Jim Gray, W. III, Bruce Lindsay, Raymond Lorie, James Mehl, Thomas Price, Gianfranco Putzolu, Patricia Selinger, Mario Schkolnick, Donald Slutz, Irving Traiger, Bradford Wade, and Robert Yost. 1981. A History and Evaluation of System R. *Commun. ACM* 24 (1981), 632–646.
- [3] Graham Cormode, Minos Garofalakis, Peter J. Haas, and Chris Jermaine. 2012. Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches. *Foundations and Trends in Databases* 4 (2012), 1–294.
- [4] Alberto Dell’Era. 2008. *Join Over Histograms*. https://www.adellera.it/static_html/investigations/join_over_histograms/JoinCardinalityEstimationWithHistogramsExplained.pdf
- [5] Goetz Graefe. 2014. The Cascades Framework for Query Optimization. *IEEE Data Eng. Bull.* 18 (2014), 19–29.
- [6] Goetz Graefe and William McKenna. 1991. The Volcano Optimizer Generator. (12 1991), 21.
- [7] PostgreSQL Global Development Group. 1996–2022. *Row Estimation Examples*. <https://www.postgresql.org/docs/14/row-estimation-examples.html>
- [8] PostgreSQL Global Development Group. 1996–2022. *Selectivity functions for standard operators*. https://github.com/postgres/postgres/blob/REL_14_STABLE/src/include/utls/selfuncs.h
- [9] PostgreSQL Global Development Group. 1996–2022. *Source Code File selfuncs.h*. <https://github.com/postgres/postgres/blame/0107855b1480d381f28f935e279ec3b64f410ef7/src/include/utls/selfuncs.h#L33>
- [10] PostgreSQL Global Development Group. 1996–2022. *Statistics Used by the Planner*. <https://www.postgresql.org/docs/14/planner-stats.html>
- [11] Shohedul Hasan, Saravanan Thirumuruganathan, Jeess Augustine, Nick Koudas, and Gautam Das. 2020. Deep Learning Models for Selectivity Estimation of Multi-Attribute Queries. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD ’20)*. Association for Computing Machinery, New York, NY, USA, 1035–1050. <https://doi.org/10.1145/3318464.3389741>
- [12] Andreas Kipf, Thomas Kipf, Bernhard Radke, Viktor Leis, Peter A. Boncz, and Alfons Kemper. 2018. Learned Cardinalities: Estimating Correlated Joins with Deep Learning. *CoRR abs/1809.00677* (2018). arXiv:1809.00677 <http://arxiv.org/abs/1809.00677>
- [13] Hai Lan, Zhifeng Bao, and Yuwei Peng. 2021. A Survey on Advancing the DBMS Query Optimizer: Cardinality Estimation, Cost Model, and Plan Enumeration. *Data Science and Engineering* 6 (2021), 86–101.
- [14] Jonathan Lewis. 2006. *Join Cardinality*. Apress, Berkeley, CA, 265–305. https://doi.org/10.1007/978-1-4302-0087-1_10
- [15] Oracle. 2021. *SQL Tuning Guide*. <https://docs.oracle.com/en/database/oracle/oracle-database/21/tgsql/histograms.html>
- [16] Evaggelia Pitoura. 2009. *Selectivity Estimation*. Springer US, 2548–2548.
- [17] Joseph Sack. 2014. Optimizing Your Query Plans with the SQL Server 2014 Cardinality Estimator. (04 2014).
- [18] SQL Server. 2021. *Cardinality Estimation (SQL Server) - SQL server*. <https://docs.microsoft.com/en-us/sql/relational-databases/performance/cardinality-estimation-sql-server?view=sql-server-ver15>
- [19] SQL Server. 2022. *Statistics - SQL server*. <https://docs.microsoft.com/en-us/sql/relational-databases/statistics/statistics?view=sql-server-ver15>
- [20] MySQL team at Oracle. [n.d.]. *MySQL 8.0 Reference Manual*. <https://dev.mysql.com/doc/refman/8.0/en/optimizer-statistics.html>
- [21] MySQL team at Oracle. [n.d.]. *Source Code File item.h*. <https://github.com/mysql/mysql-server/blob/8d8c986e5716e38cb776b627a8ee9e92241b4ce/sql/item.h#L100>
- [22] MySQL team at Oracle. 2000. *MySQL Server*. <https://github.com/mysql/mysql-server/tree/mysql-8.0.29>
- [23] Xiaoying Wang, Changbo Qu, Weiyuan Wu, Jiannan Wang, and Qingqing Zhou. 2021. Are We Ready for Learned Cardinality Estimation? *Proc. VLDB Endow.* 14, 9 (may 2021), 1640–1654. <https://doi.org/10.14778/3461535.3461552>
- [24] Zongheng Yang, Eric Liang, Amog Kamsetty, Chenggang Wu, Yan Duan, Xi Chen, Pieter Abbeel, Joseph M. Hellerstein, Sanjay Krishnan, and Ion Stoica. 2019. Selectivity Estimation with Deep Likelihood Models. *CoRR abs/1905.04278* (2019). arXiv:1905.04278 <http://arxiv.org/abs/1905.04278>