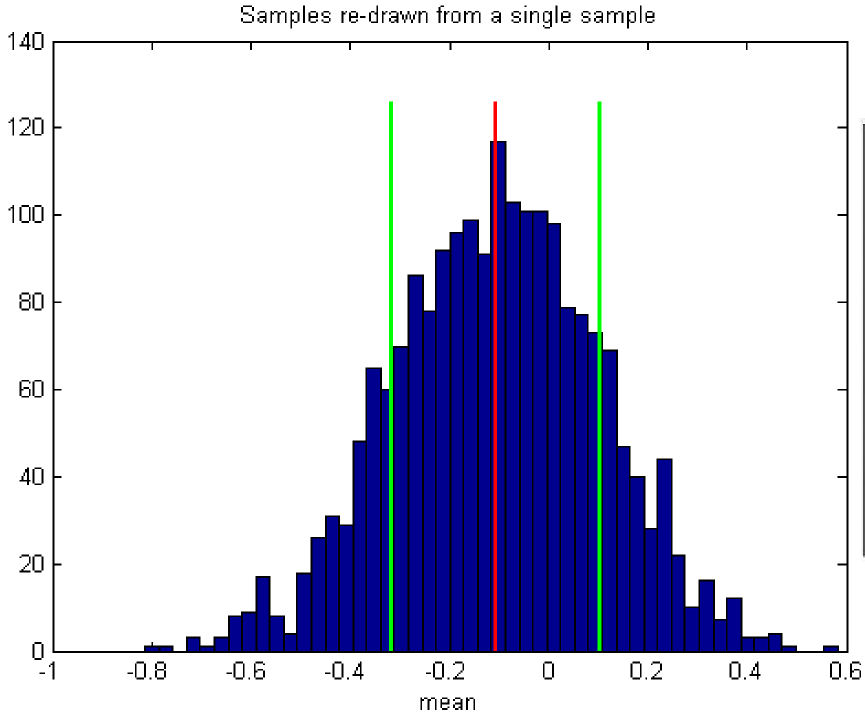
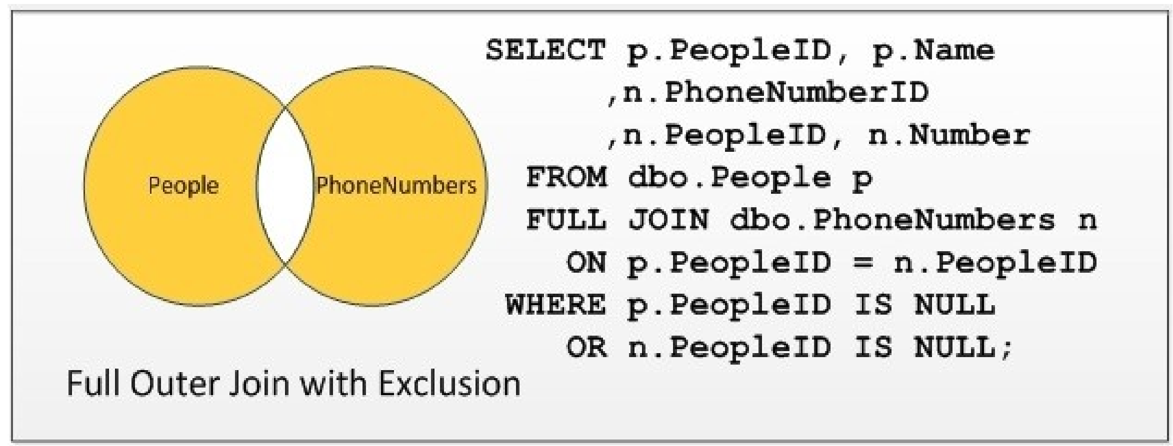
A Sampling Algebra for Aggregate Estimation





(presented by Philipp Moritz)

Motivation

- Joins and Sampling do not commute → cannot "push down" the sampling operator to the leaves of the query
- Sampling is an important operation:
  - For efficient subsampling algorithms
  - Bootstrap in Statistics for Confidence Intervals
- Ideally: Subsampling + Confidence Intervals
- Implementations for restricted Joins known (single table, AQUA)

Generalized Uniform Sampling

- Approach: Do not generate iid samples, rather calculate the samples quantity + approximation to confidence intervals directly

DEFINITION 1 (GUS SAMPLING [12]). *A randomized selection process $\mathcal{G}_{(a,\bar{b})}$ which gives a sample $\mathcal{R}$ from $\mathbf{R} = R_1 \times R_2 \times \cdots \times R_n$ is called Generalized Uniform Sampling (GUS) method, if, for any given tuples $t = (t_1, \ldots, t_n), t' = (t'_1, \ldots, t'_n) \in \mathbf{R}$, $P(t \in \mathcal{R})$ is independent of t, and $P(t, t' \in \mathcal{R})$ depends only on $\{i : t_i = t'_i\}$. In such a case, the GUS parameters a, $\bar{b} = \{b_T | T \subset \{1 : n\}\}$ are defined as:*

$$a = P[t \in \mathcal{R}]$$

$$b_T = P[t \in \mathcal{R} \wedge t' \in \mathcal{R} | \forall i \in T, t_i = t'_i, \forall j \in T^C, t_j \neq t'_j].$$

Second Order Equivalence

- An equivalence relation for transforming statements involving GUS quasioperators
- SOE equivalence is equivalent to first and second order probabilities P(t \in E(R)) and P(t, u \in E(R)) agreeing

Definition 2 (SOA-equivalence). *Given (possibly randomized) expressions* $\mathcal{E}(R)$ *and* $\mathcal{F}(R)$*, we say*

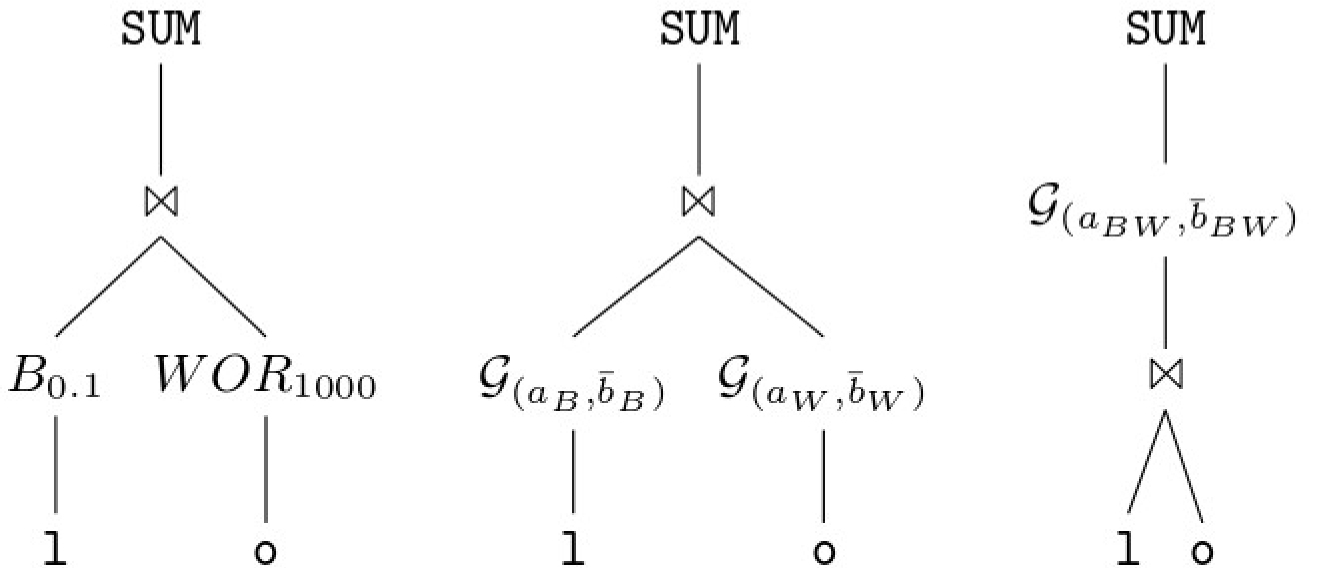
$$\mathcal{E}(R) \overset{SOA}{\Longleftrightarrow} \mathcal{F}(R)$$

if for ***any arbitrary*** *SUM-aggregate* $\mathcal{A}_f(S) = \sum_{t \in S} f(t)$,

$$E[\mathcal{A}_f(\mathcal{E}(R))] = E[\mathcal{A}_f(\mathcal{F}(R))]$$
$$Var[\mathcal{A}_f(\mathcal{E}(R))] = Var[\mathcal{A}_f(\mathcal{F}(R))].$$

Transforming into standard form

- Using a simple set of rules, the GUS quasioperator can be pushed up in the query tree, just below the aggregating operation



(a) (b) (c)

Computing sampling approximation and confidence intervals

- Once the transformation on the last slide has been performed, the following theorem allows the calculation of subsampled aggregate results:

THEOREM 1. *[12] Let $f(t)$ be a function/property of $t \in R$, and $\mathcal{R}$ be the sample obtained by a GUS method $\mathcal{G}_{(a,\bar{b})}$. Then, the aggregate $\mathcal{A} = \sum_{\mathbf{t} \in R} f(\mathbf{t})$ and the sampling estimate $X = \frac{1}{a} \sum_{\mathbf{t} \in \mathcal{R}} f(\mathbf{t})$ have the property:*

$$E[X] = \mathcal{A}$$

$$\sigma^2(X) = \sum_{S \subset \{1:n\}} \frac{c_S}{a^2} y_S - y_\phi \tag{1}$$

with

$$y_S = \sum_{t_i \in R_i | i \in S} \left(\sum_{t_j \in R_j | j \in S^C} f(t_i, t_j) \right)^2$$

$$c_S = \sum_{T \in \mathcal{P}(n)} (-1)^{|T|+|S|} \, b_T.$$

Efficient Implementation

- Can use further subsampling to compute Y_S:

$$\hat{Y}_S = \frac{1}{c_{S,\emptyset}} \left(Y_S - \sum_{T \subset S^C, T \neq \emptyset} c_{S,T} \hat{Y}_{S \cup T} \right)$$

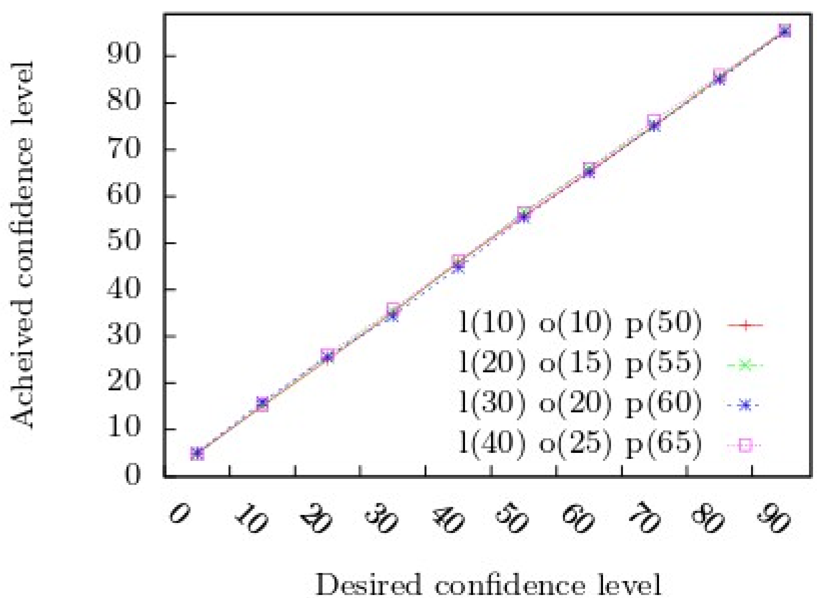
where

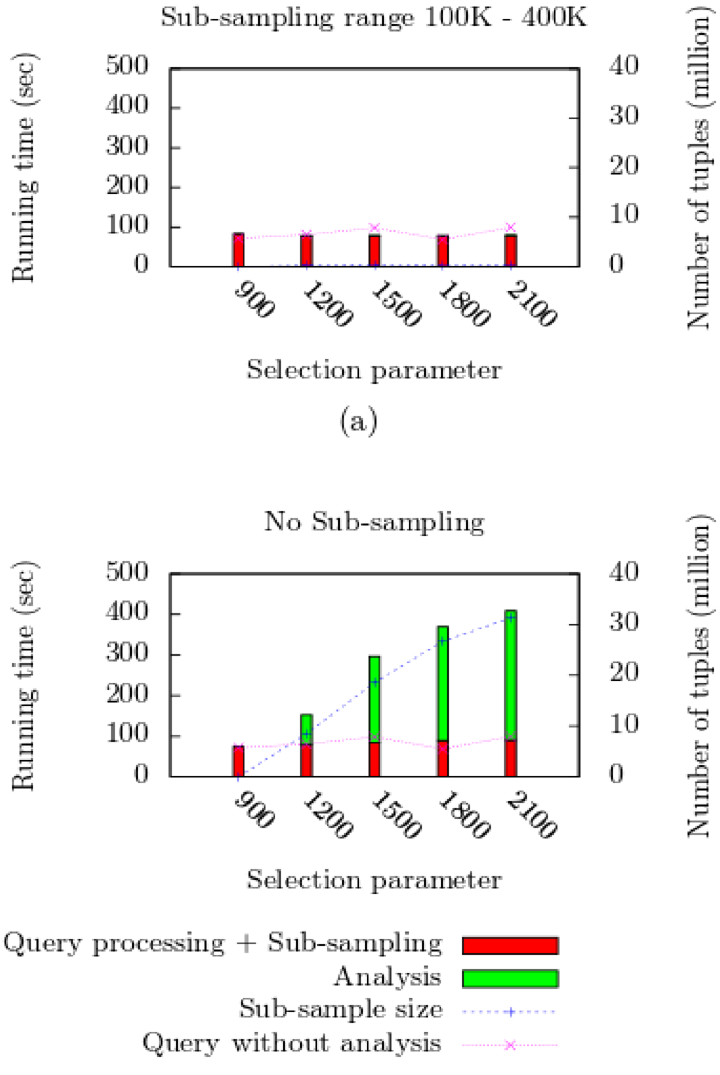
$$c_{S,T} = \sum_{U \subset T} (-1)^{|U|+|S|} \, b_{S \cup U}.$$

Experimental validation

- Running the following query on a 1TB dataset:

```
SELECT SUM(l_discount*(1.0-l_tax))
FROM lineitem TABLESAMPLE (x PERCENT),
  orders TABLESAMPLE(y ROWS),
  part TABLESAMPLE(z PERCENT)
WHERE l_orderkey  = o_orderkey AND
   l_partkey = p_partkey AND o_totalprice < q AND
   p_retailprice < r;
```





Discussion

- For query on n columns, need 2^n operations → too large?
- Using only variance for constructing confidence intervals can lead to far too tight or loose intervals
- In certain cases, can do explicit sampling + statistical bootstrap (or Bag of Little Bootstraps)