Query Optimization
Exercise Session 11

Andrey Gubichev

January 12, 2015

Today

A: A branch of mathematics concerning the study of finite or countable discrete structures.

Q: What is ?

Today

A: A branch of mathematics concerning the study of finite or countable discrete structures.

Q: What is combinatorics?

Combinatorics 101

Given a set of n elements, how many distinct k-element subsets can be formed?

Combinatorics 101

Given a set of n elements, how many distinct k-element subsets can be formed?

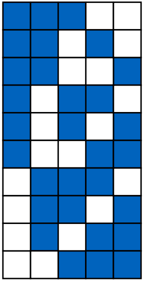
$$\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$$

Combinatorics 101

Given a set of n elements, how many distinct k-element subsets can be formed?

$$\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$$

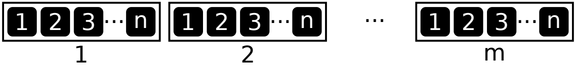
Example: Choose 3 out of 5: $\binom{5}{3} = \frac{5!}{2! \cdot 3!} = \frac{120}{2 \cdot 6} = 10$



Direct, Uniform, Distinct

Waters/Yao Bottom-Up

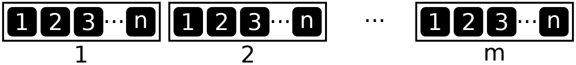
Given m pages with n tuples on each page, e.g. a total of $N = m \cdot n$ tuples:



- How many distinct subsets of size k exist?

Waters/Yao Bottom-Up

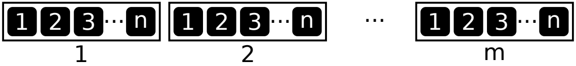
Given m pages with n tuples on each page, e.g. a total of $N = m \cdot n$ tuples:



- How many distinct subsets of size k exist? $\binom{N}{k}$
- How many distinct subsets of size k exist, where a page does not contain any chosen tuples? Choose k from all but one page, i.e. from $N - n$ tuples:

Waters/Yao Bottom-Up

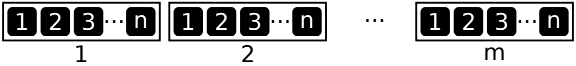
Given m pages with n tuples on each page, e.g. a total of $N = m \cdot n$ tuples:



- How many distinct subsets of size k exist? $\binom{N}{k}$
- How many distinct subsets of size k exist, where a page does not contain any chosen tuples? Choose k from all but one page, i.e. from $N - n$ tuples: $\binom{N-n}{k}$
  So the probability that a page contains none of the k tuples is

Waters/Yao Bottom-Up

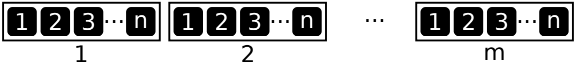
Given m pages with n tuples on each page, e.g. a total of $N = m \cdot n$ tuples:



- How many distinct subsets of size k exist? $\binom{N}{k}$
- How many distinct subsets of size k exist, where a page does not contain any chosen tuples? Choose k from all but one page, i.e. from $N - n$ tuples: $\binom{N-n}{k}$
  So the probability that a page contains none of the k tuples is
  $$p := \frac{\binom{N-n}{k}}{\binom{N}{k}}$$
- What is the probability that a certains page contains at least one tuple?

Waters/Yao Bottom-Up

Given m pages with n tuples on each page, e.g. a total of $N = m \cdot n$ tuples:



- How many distinct subsets of size k exist? $\binom{N}{k}$
- How many distinct subsets of size k exist, where a page does not contain any chosen tuples? Choose k from all but one page, i.e. from $N - n$ tuples: $\binom{N-n}{k}$
  So the probability that a page contains none of the k tuples is

$$p := \frac{\binom{N-n}{k}}{\binom{N}{k}}$$

- What is the probability that a certains page contains at least one tuple? $1 - p \ldots$ unless all pages have to be involved $(k > N - n)$.
- Multiplied by the number of pages, we get the number of qualifying pages, denoted $\overline{\mathcal{Y}}_n^{N,m}(k)$.

Approximation

Let $m = 50$, $n = 1000 \Rightarrow N = 50k$, $k = 100$

Yao (exact) : $p = \frac{\binom{N-n}{k}}{\binom{N}{k}} = \prod_{i=0}^{k-1} \frac{N-n-i}{N-i} = \prod_{i=0}^{99} \frac{49k-i}{50k-i} = 13.2\%$

Waters : $p \approx (1 - \frac{k}{N})^n$

Approximation

Let $m = 50$, $n = 1000 \Rightarrow N = 50k$, $k = 100$

Yao (exact) : $p = \frac{\binom{N-n}{k}}{\binom{N}{k}} = \prod_{i=0}^{k-1} \frac{N-n-i}{N-i} = \prod_{i=0}^{99} \frac{49k-i}{50k-i} = 13.2\%$

Waters : $p \approx (1 - \frac{k}{N})^n \approx 13.5\%$

Direct, Uniform, Non-Distinct

Combinatorics 101 revisited

- Now *with replacement*: How many distinct *multisets* exist chosing k from n?

Combinatorics 101 revisited

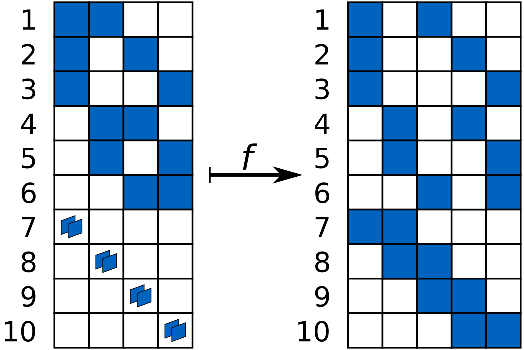
- Now *with replacement*: How many distinct *multisets* exist chosing k from n?
  As many as there are distinct sets chosing k from $n + k - 1$!

Combinatorics 101 revisited

- Now *with replacement*: How many distinct *multisets* exist chosing k from n?
  As many as there are distinct sets chosing k from $n + k - 1$!
- Bijection between multisets and sets. From multiset to set:
  $f : (x_1, x_2, \ldots, x_k) \mapsto (x_1 + 0, x_2 + 1, \ldots, x_k + (k - 1))$

Combinatorics 101 revisited

- Now *with replacement*: How many distinct *multisets* exist chosing k from n?
  As many as there are distinct sets chosing k from $n+k-1$!
- Bijection between multisets and sets. From multiset to set: $f : (x_1, x_2, \ldots, x_k) \mapsto (x_1 + 0, x_2 + 1, \ldots, x_k + (k-1))$
- Example: Choose 2 from 4
  - # sets: $\binom{4}{2}$
  - # multisets: $\binom{4+2-1}{2}$

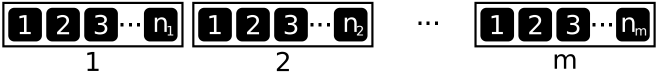


Cheung

- Like Yao, but not necessarily distinct
- Same formula as Yao, but:
  - We don't need to distinguish cases when computing the probability that a bucket contains at least one item
  - We substitue N by $N + k - 1$ to compute $\tilde{p}$

Direct, Non-Uniform, Distinct

Direct, Non-Uniform, Distinct



Assume that $n_j > 0\ \forall j \in [1, m]$, then the expected number of qualifying pages is

$$\sum_{j=1}^{m} \left(1 - \frac{\binom{N-n_j}{k}}{\binom{N}{k}} \right)$$

With $N = \sum_{j=1}^{m} n_j$.

Distribution Function

- The number of possibilities to select x $(x \leq n_j)$ items from bucket j is $\binom{n_j}{x}$.
- The number of possibilities to draw the remaining $k - x$ items from other buckets is $\binom{N-n_j}{k-x}$.
- Recall: The number of possibilities to draw k items from N is $\binom{N}{k}$.

$\Rightarrow$ The probability that x items qualify from bucket j is

$$\frac{\binom{n_j}{x}\binom{N-n_j}{k-x}}{\binom{N}{k}}$$

Sequential, Uniform, Distinct

Sequential, Uniform, Distinct

- Estimate the distribution of distance between two qualifying tuples
- Bitvector B, b bits are set to 1
- First, let's find the distribution of number of zeros
  - before first 1
  - between two consecutive 1s
  - after last 1
- $B - j - 1$ positions for i
- every bitvector has $b - 1$ sequences of a form $10 \ldots 01$
- $\frac{(B-j-1)\binom{B-j-2}{b-2}}{(b-1)\binom{B}{b}} = \frac{\binom{B-j-1}{b-1}}{\binom{B}{b}}$
- now, the expected number of 0s: $\frac{B-b}{b+1}$
- then, the expected total number of bits between first and last 1:

Sequential, Uniform, Distinct

- Estimate the distribution of distance between two qualifying tuples
- Bitvector B, b bits are set to 1
- First, let's find the distribution of number of zeros
  - before first 1
  - between two consecutive 1s
  - after last 1
- $B - j - 1$ positions for i
- every bitvector has $b - 1$ sequences of a form $10 \ldots 01$
- $\frac{(B-j-1)\binom{B-j-2}{b-2}}{(b-1)\binom{B}{b}} = \frac{\binom{B-j-1}{b-1}}{\binom{B}{b}}$
- now, the expected number of 0s: $\frac{B-b}{b+1}$
- then, the expected total number of bits between first and last 1: $B - \frac{B-b}{b+1} = \frac{Bb+b}{b+1}$

Histograms

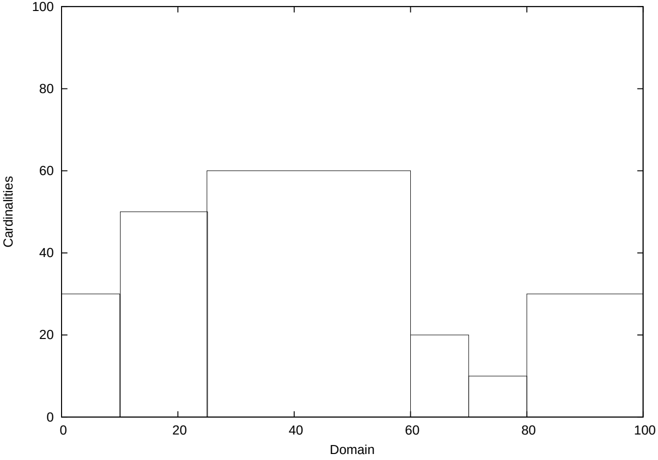
A histogram $H_A : B \to \mathbb{N}$ over a relation R partitions the domain of the aggregated attribute A into disjoint buckets B, such that

$$H_A(b) = |\{r | r \in R \wedge R.A \in b\}|$$

and thus $\sum_{b \in B} H_A(b) = |R|$.

Histograms

A rough histogram might look like this:



Using Histograms (3)

Given a histogram, we can approximate the selectivities as follows:

$$A = c \qquad \frac{\sum_{b \in B : c \in b} H_A(b)}{\sum_{b \in B} H_A(b)}$$

$$A > c \qquad \frac{\sum_{b \in B : c \in b} \frac{\max(b) - c}{\max(b) - \min(b)} H_A(b) + \sum_{b \in B : \min(b) > c} H_A(b)}{\sum_{b \in B} H_A(b)}$$

$$A_1 = A_2 \qquad \frac{\sum_{b_1 \in B_1, b_2 \in B_2, b' = b_1 \cap b_2 : b' \neq \emptyset} \frac{\max(b') - \min(b')}{\max(b_1) - \min(b_1)} H_{A_1}(b_1) \frac{\max(b') - \min(b')}{\max(b_2) - \min(b_2)} H_{A_2}(b_2)}{\sum_{b_1 \in B_1} H_{A_1}(b_1) \sum_{b_2 \in B_2} H_{A_2}(b_2)}$$

Info

- Exercises due January 19, 2015.