Direct, Non-Uniform, Distinct

So far:

1. every page contains the same number of records, and
2. every record is accessed with the same probability.

Now:

> *Model the distribution of items to buckets by m numbers n_i (for $1 \leq i \leq m$) if there are m buckets.*
> *Each n_i equals the number of records in some bucket i.*

Direct, Non-Uniform, Distinct (2)

The following theorem is a simple application of Yao's formula:

Theorem (Yao/Waters/Christodoulakis)

Assume a set of m buckets. Each bucket contains $n_j > 0$ items $(1 \leq j \leq m)$. The total number of items is $N = \sum_{j=1}^{m} n_j$. If we lookup k distinct items, then the probability that bucket j qualifies is

$$\mathcal{W}_{n_j}^{N}(k,j) = [1 - \frac{\binom{N-n_j}{k}}{\binom{N}{k}}] \quad (= \mathcal{Y}_{n_j}^{N}(k)) \tag{28}$$

and the expected number of qualifying buckets is

$$\overline{\mathcal{W}}_{n_j}^{N,m}(k) := \sum_{j=1}^{m} \mathcal{W}_{n_j}^{N}(k,j) \tag{29}$$

Direct, Non-Uniform, Distinct (3)

The product formulation in Eq. 20 of Theorem 2 results in a more efficient computation:

Corollary

If we lookup k distinct items, then the expected number of qualifying buckets is

$$\overline{\mathcal{W}}_{n_j}^{N,m}(k) = \sum_{j=1}^{m}(1 - p_j) \tag{30}$$

with

$$p_j = \begin{cases} \prod_{i=0}^{n_j-1} \frac{N-k-i}{N-i} & k \leq n_j \\ 0 & N - n_j < k \leq N \end{cases} \tag{31}$$

Direct, Non-Uniform, Distinct (4)

If we compute the p_j after we have sorted the n_j in ascending order, we can use the fact that

$$p_{j+1} = p_j * \prod_{i=n_j}^{n_{j+1}-1} \frac{N-k-i}{N-i}.$$

Direct, Non-Uniform, Distinct (5)

Many buckets: statistics too big. Better: Histograms

Corollary

For $1 \leq i \leq L$ let there be l_i buckets containing n_i items. Then, the total number of buckets is $m = \sum_{i=1}^{L} l_i$ and the total number of items in all buckets is $N = \sum_{i=1}^{L} l_i n_i$. For k randomly selected items the number of qualifying buckets is

$$\overline{\mathcal{W}}_{n_j}^{N,m}(k) = \sum_{i=1}^{L} l_i \mathcal{Y}_{n_j}^{N}(k) \tag{32}$$

Direct, Non-Uniform, Distinct (6)

Distribution function. The probability that $x \leq n_j$ items in a bucket j qualify, can be calculated as follows:

- The number of possibilities to select x items in bucket n_j is

$$\binom{n_j}{x}$$

- The number of possibilites to draw the remaining $k - x$ items from the other buckets is

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

- The total number of possibilities to distributed k items over the buckets is

$$\binom{N}{k}$$

This shows the following:

Direct, Non-Uniform, Distinct (7)

Theorem

Assume a set of m buckets. Each bucket contains $n_j > 0$ items $(1 \leq j \leq m)$. The total number of items is $N = \sum_{j=1}^{m} n_j$. If we lookup k distinct items, then the probability that x items in bucket j qualify is

$$\mathcal{X}_{n_j}^{N}(k,x) = \frac{\binom{n_j}{x} \quad \binom{N-n_j}{k-x}}{\binom{N}{k}} \tag{33}$$

Further, the expected number of qualifying items in bucket j is

$$\overline{\mathcal{X}}_{n_j}^{N,m}(k) = \sum_{x=0}^{\min(k,n_j)} x \mathcal{X}_{n_j}^{N}(k,x) \tag{34}$$

In standard statistics books the probability distribution $\mathcal{X}_{n_j}^{N}(k,x)$ is called *hypergeometric distribution*.

Direct, Non-Uniform, Distinct (8)

Let us consider the case where all n_j are equal to n. Then, we can calculate the average number of qualifying items in a bucket. With $y := \min(k, n)$ we have

$$
\begin{aligned}
\overline{\mathcal{X}}^{N,m}_{n_j}(k) &= \sum_{x=0}^{\min(k,n)} x\mathcal{X}^N_n(k,x) \\
&= \sum_{x=1}^{\min(k,n)} x\mathcal{X}^N_n(k,x) \\
&= \frac{1}{\binom{N}{k}} \sum_{x=1}^{y} x\binom{n}{x}\binom{N-n}{k-x}
\end{aligned}
$$

Direct, Non-Uniform, Distinct (9)

$$
\begin{aligned}
\overline{\mathcal{X}}^{N,m}_{n_j}(k) &= \frac{1}{\binom{N}{k}} \sum_{x=1}^{y} x \binom{n}{x} \binom{N-n}{k-x} \\
&= \frac{1}{\binom{N}{k}} \sum_{x=1}^{y} \binom{x}{1} \binom{n}{x} \binom{N-n}{k-x} \\
&= \frac{1}{\binom{N}{k}} \sum_{x=1}^{y} \binom{n}{1} \binom{n-1}{x-1} \binom{N-n}{k-x} \\
&= \frac{\binom{n}{1}}{\binom{N}{k}} \sum_{x=0}^{y-1} \binom{n-1}{0+x} \binom{N-n}{(k-1)-x} \\
&= \ldots
\end{aligned}
$$

(cont.)

Direct, Non-Uniform, Distinct (10)

$$\begin{aligned}
\overline{\mathcal{X}}_{n_j}^{N,m}(k) &= \dots \\
&= \frac{\binom{n}{1}}{\binom{N}{k}} \binom{n-1+N-n}{0+k-1} \\
&= \frac{\binom{n}{1}}{\binom{N}{k}} \binom{N-1}{k-1} \\
&= n\frac{k}{N} \quad = \quad \frac{k}{m}
\end{aligned}$$

Direct, Non-Uniform, Distinct (11)

Let us consider the even more special case where every bucket contains a single item. That is, $N = m$ and $n_i = 1$. The probability that a bucket contains a qualifying item reduces to

$$\begin{aligned}
\mathcal{X}_1^N(k,x) &= \frac{\binom{1}{x} \quad \binom{N-1}{k-1}}{\binom{N}{k}} \\
&= \frac{\binom{N-1}{k-1}}{\binom{N}{k}} \\
&= \frac{k}{N} \quad (= \frac{k}{m})
\end{aligned}$$

Since x can then only be zero or one, the average number of qualifying items a bucket contains is also $\frac{k}{N}$.

Sequential: Vector of Bits

When estimating seek costs, we need to calculate the probability distribution for the distance between two subsequent qualifying cylinders. We model the situation as a bitvector of length B with b bits set to one. Then, B corresponds to the number of cylinders and a one indicates that a cylinder qualifies.
[Later: Vector of Buckets]

Sequential: Vector of Bits (2)

Theorem

Assume a bitvector of length B. Within it b ones are uniformly distributed. The remaining $B - b$ bits are zero. Then, the probability distribution of the number j of zeros

1. *between two consecutive ones,*
2. *before the first one, and*
3. *after the last one*

is given by

$$\mathcal{B}_b^B(j) = \frac{\binom{B-j-1}{b-1}}{\binom{B}{b}} \tag{35}$$

Sequential: Vector of Bits (3)

Proof:
To see why the formula holds, consider the total number of bitvectors having a one in position i followed by j zeros followed by a one.
This number is

$$\binom{B-j-2}{b-2}$$

We can chose $B-j-1$ positions for i.
The total number of bitvectors is

$$\binom{B}{b}$$

and each bitvector has $b-1$ sequences of the form that a one is followed by a sequence of zeros is followed by a one.

Sequential: Vector of Bits (4)

Hence,

$$\begin{aligned}\mathcal{B}_b^B(j) &= \frac{(B-j-1)\binom{B-j-2}{b-2}}{(b-1)\binom{B}{b}} \\ &= \frac{\binom{B-j-1}{b-1}}{\binom{B}{b}}\end{aligned}$$

Part (1) follows.
To prove (2), we count the number of bitvectors that start with j zeros before the first one.
There are $B-j-1$ positions left for the remaining $b-1$ ones.
Hence, the number of these bitvectors is $\binom{B-j-1}{b-1}$ and part (2) follows.
Part (3) follows by symmetry.

Sequential: Vector of Bits (5)

We can derive a less expensive way to calculate formula for $\mathcal{B}_b^B(j)$ as follows.
For $j = 0$, we have $\mathcal{B}_b^B(0) = \frac{b}{B}$.
If $j > 0$, then

$$\begin{aligned}
\mathcal{B}_b^B(j) &= \frac{\binom{B-j-1}{b-1}}{\binom{B}{b}} \\
&= \frac{\frac{(B-j-1)!}{(b-1)!((B-j-1)-(b-1))!}}{\frac{B!}{b!(B-b)!}} \\
&= \frac{(B-j-1)! \;\; b!(B-b)!}{(b-1)!((B-j-1)-(b-1))! \;\; B!}
\end{aligned}$$

Sequential: Vector of Bits (6)

$$
\begin{aligned}
\mathcal{B}_b^B(j) &= \frac{(B-j-1)! \;\; b!(B-b)!}{(b-1)!((B-j-1)-(b-1))! \;\; B!} \\
&= b\frac{(B-j-1)! \;\; (B-b)!}{((B-j-1)-(b-1))! \;\; B!} \\
&= b\frac{(B-j-1)! \;\; (B-b)!}{(B-j-b)! \;\; B!} \\
&= \frac{b}{B-j}\frac{(B-j)! \;\; (B-b)!}{(B-b-j)! \;\; B!} \\
&= \frac{b}{B-j}\prod_{i=0}^{j-1}(1-\frac{b}{B-i})
\end{aligned}
$$

This formula is useful when $\mathcal{B}_b^B(j)$ occurs in sums over j.

Sequential: Vector of Bits (7)

Corollary

Using the terminology of Theorem 8, the expected value for the number of zeros

1. *before the first one,*
2. *between two successive ones, and*
3. *after the last one*

is

$$\overline{\mathcal{B}}_b^B = \sum_{j=0}^{B-b} j\mathcal{B}_b^B(j) = \frac{B-b}{b+1} \tag{36}$$

Sequential: Vector of Bits (8)

Proof:

$$
\begin{aligned}
\sum_{j=0}^{B-b} j\binom{B-j-1}{b-1} &= \sum_{j=0}^{B-b}(B-(B-j))\binom{B-j-1}{b-1} \\
&= B\sum_{j=0}^{B-b}\binom{B-j-1}{b-1} - \sum_{j=0}^{B-b}(B-j)\binom{B-j-1}{b-1} \\
&= B\sum_{j=0}^{B-b}\binom{b-1+j}{b-1} - b\sum_{j=0}^{B-b}\binom{B-j}{b} \\
&= B\sum_{j=0}^{B-b}\binom{b-1+j}{j} - b\sum_{j=0}^{B-b}\binom{b+j}{b}
\end{aligned}
$$

Sequential: Vector of Bits (9)

$$\begin{aligned}
\sum_{j=0}^{B-b} j\binom{B-j-1}{b-1} &= B\sum_{j=0}^{B-b}\binom{b-1+j}{j} - b\sum_{j=0}^{B-b}\binom{b+j}{b} \\
&= B\binom{(b-1)+(B-b)+1}{(b-1)+1} - b\binom{b+(B-b)+1}{b+1} \\
&= B\binom{B}{b} - b\binom{B+1}{b+1} \\
&= (B - b\frac{B+1}{b+1})\binom{B}{b}
\end{aligned}$$

With

$$\begin{aligned}
B - b\frac{B+1}{b+1} &= \frac{B(b+1)-(Bb+b)}{b+1} \\
&= \frac{B-b}{b+1}
\end{aligned}$$

the claim follows

Sequential: Vector of Bits (10)

Corollary

Using the terminology of Theorem 8, the expected total number of bits from the first bit to the last one, both included, is

$$\overline{\mathcal{B}}_{tot}(B, b) = \frac{Bb + b}{b + 1} \tag{37}$$

Sequential: Vector of Bits (11)

Proof:
We subtract from B the average expected number of zeros between the last one and the last bit:

$$\begin{aligned} B - \frac{B-b}{b+1} &= \frac{B(b+1)}{b+1} - \frac{B-b}{b+1} \\ &= \frac{Bb+B-B+b}{b+1} \\ &= \frac{Bb+b}{b+1} \end{aligned}$$

Sequential: Vector of Bits (12)

Corollary

Using the terminology of Theorem 8, the number of bits from the first one and the last one, both included, is

$$\overline{\mathcal{B}}_{\textit{1-span}}(B, b) = \frac{Bb - B + 2b}{b + 1} \tag{38}$$

Sequential: Vector of Bits (13)

Proof (alternative 1):
Subtract from B the number of zeros at the beginning and the end:

$$\begin{aligned}\overline{\mathcal{B}}_{\text{1-span}}(B,b) &= B - 2\frac{B-b}{b+1}\\ &= \frac{Bb+B-2B+2b}{b+1}\\ &= \frac{Bb-B+2b}{b+1}\end{aligned}$$

Sequential: Vector of Bits (14)

Proof (alternative 2):
Add the number of zeros between the first and the last one and the number of ones:

$$\begin{aligned}
\overline{\mathcal{B}}_{\text{1-span}}(B,b) &= (b-1)\overline{\mathcal{B}}_b^B + b \\
&= (b-1)\frac{B-b}{b+1} + \frac{b(b+1}{b+1} \\
&= \frac{Bb-b^2-B+b+b^2+b}{b+1} \\
&= \frac{Bb-B+2b}{b+1}
\end{aligned}$$

Sequential: Applications for Bitvector Model

- If we look up one record in an array of B records and we search sequentially, how many array entries do we have to examine on average if the search is successful?
- Let a file consist of B consecutive cylinders. We search for k different keys all of which occur in the file. These k keys are distributed over b different cylinders. Of course, we can stop as soon as we have found the last key. What is the expected total distance the disk head has to travel if it is placed on the first cylinder of the file at the beginning of the search?
- Assume we have an array consisting of B different entries. We sequentially go through all entries of the array until we have found all the records for b different keys. We assume that the B entries in the array and the b keys are sorted. Further all b keys occur in the array. On the average, how many comparisons do we need to find all keys?

Sequential: Vector of Buckets

Theorem (Yao)

Consider a sequence of m buckets. For $1 \leq i \leq m$, let n_i be the number of items in a bucket i. Then there is a total of $N = \sum_{i=1}^{m} n_i$ items. Let $t_i = \sum_{l=0}^{i} n_i$ be the number of items in the first i buckets. If the buckets are searched sequentially, then the probability that j buckets that have to be examined until k distinct items have been found is

$$\mathcal{C}_{n_i}^{N,m}(k,j) = \frac{\binom{t_j}{k} - \binom{t_{j-1}}{k}}{\binom{N}{k}} \tag{39}$$

Thus, the expected number of buckets that need to be examined in order to retrieve k distinct items is

$$\overline{\mathcal{C}}_{n_i}^{N,m}(k) = \sum_{j=1}^{m} j\mathcal{C}_{n_i}^{N,m}(k,j) = m - \frac{\sum_{j=1}^{m} \binom{t_{j-1}}{k}}{\binom{N}{k}} \tag{40}$$

Sequential: Vector of Buckets (2)

The following theorem is very useful for deriving estimates for average sequential accesses under different models [Especially: the above theorem follows].

Theorem (Lang/Driscoll/Jou)

Consider a sequence of N items. For a batched search of k items, the expected number of accessed items is

$$A(N,k) = N - \sum_{i=1}^{N-1} Prob[Y \leq i] \tag{41}$$

where Y is a random variable for the last item in the sequence that occurs among the k items searched.