Urn Models and Yao's Formula

Danièle Gardy and Laurent Némirovski

Laboratoire PRiSM, Université de Versailles Saint-Quentin, 45 avenue des Etats-Unis, 78035 Versailles, France

Abstract. Yao's formula is one of the basic tools in any situation where one wants to estimate the number of blocks to be read in answer to some query. We show that such situations can be modelized by probabilistic urn models. This allows us to fully characterize the distribution probability of the number of selected blocks under uniformity assumptions, and to consider extensions to non-uniform block probabilities. We also obtain a computationnally efficient approximation of Yao's formula.

1 Introduction

Evaluating the performances of any database is an intricate task, with many intermediate computations. One frequent step consists in evaluating the number of memory blocks that must be read in order to obtain a specified subset of some set, for example by a selection. A first answer was given by Yao in [Ya], at least for the expectation of the number of selected blocks, and under uniformity and independance assumptions.

The mathematical treatment of Yao was based on a very simple expression for the number of ways of choosing i objects among j; this allowed him to obtain the average number of desired blocks, but he did not characterize further the probability distribution. Getting enough information on the probability distribution is important, because the mean value is not enough to characterize a random variable : If substituting the mean value for the random variable itself, when computing some query costs, certainly allows fast evaluation of some costs in a situation where the accent is not on the detailed study of the number of retrieved blocks, but rather on quick computations for choosing an execution plan, at the same time one must have some confidence that using this average value will not induce too much error and lead us to choose a wrong plan!

Extending Yao's approach to try and get more information on the distribution, such as its higher order moments, would quickly lead to intricate computations, which may finally succeed in giving a mathematical expression in the uniform case, but which would probably not express the underlying mathematical phenomenon in an intuitive, easy to understand form, and which would fail when considering non-uniform distributions on blocks. We shall see that, when using the right framework (random allocations and probabilistic urn models), it is easy to obtain all the desired information on the random variable *number of retrieved*

blocks; as a consequence, we can check that in many situations the random variable is quite close to its expectation, which justifies the use of this expectation instead of the random variable itself. More precisely, we give the variance of the number of selected blocks, and we can obtain higher order moments, if desired. However, this seems less useful that getting the approximate distribution, for large numbers of blocks m and of selected objects n : This distribution is Gaussian, and its mean and variance are very easy to compute.[1]

Our first mathematical tool is a random allocation model that is an extension of a well-known occupancy urn model : When allocating a given number of balls into a sequence of urns, what can we say about the random variable *number of empty urns?* Our second tool is a systematic use of the methods of the analysis of algorithms, i.e. generating functions, and asymptotic approximations by complex analysis.

Numerical computations show that our approximation of the expectation is both very close to the exact value and much quicker to compute than Yao's formula : In many situations *it is not worthwhile to use the exact formule* , which differs from our approximation by a negligible amount, and which requires a significant computation time. We compute our approximation of Yao's formula with a constant number of operations, while the exact formule requires $O(n)$ operations. Another interesting consequence is that our mathematical model easily lends itself to extensions such as non-uniform distributions. We consider in this paper piecewise distributions, and show that we can extend our results to obtain again an asymptotic Gaussian limiting distribution, whose mean and variance can be efficiently computed.

The plan of the paper is as follows : We present in Section 2 our urn model for the classical problem (uniform placement of objects in blocks), then study the distribution of the number of selected blocks. We discuss in Section 3 the adequacy of an extension of Yao's formula (presented in [GGT]) to the case where the probabilities of blocks are no longer uniform. This leads us to an extension to piecewise distributions in Section 4. We also give in this section a sketch of our proofs; we refer the interested reader to [GN2] for detailed proofs. We conclude by a discussion of our results and of possible extensions in Section 5. A preliminary version of these results is presented in [GN1].

2 Yao's formula revisited

2.1 Notations and former result.

Consider a set $\mathcal{E}$ of objects, whose representation in memory requires a specified number of pages. A query on the set $\mathcal{E}$ selects a subset $\mathcal{F}$ of $\mathcal{E}$; now assume that the cardinality of the answer set $\mathcal{F}$ is known; how many pages must be read to

[1] We recall that a Gaussian distribution is fully determined by its first two moments, i.e. by its expectation and its variance.

access all the objects of $\mathcal{F}$? It is usual to assume that placement of the objects of $\mathcal{F}$ is done randomly and uniformly : each object of $\mathcal{F}$ has the same probability to be on any page. We shall use the following notations :

- the total number of objects is p, these p objects are on m memory pages; each block contains $b = p/m$ objects (in this section we assume that all the blocks have the same capacity);
- the query selects n objects;
- the number of blocs containing the n objects selected by the query is a random variable with integer values X.

Yao gave in [Ya] the expectation of X, by computing the probability that a given page is *not* selected : This probability is equal to the number of allocations of the n selected objects among $m-1$ pages, divided by the total number of configurations, i.e. by the number of allocations of n objects among m pages; hence

$$E[X] = m\left(1 - \frac{\binom{p-b}{n}}{\binom{p}{n}}\right). \tag{1}$$

2.2 Occupancy urn models.

This familiar problem lends itself to a detailed probabilistic analysis, based on a random allocation model that is a variation of one of the basic urn occupancy models. We refer the reader to the survey book of Johnson and Kotz [JK] for a general presentation of urn models and give directly the urn model that we shall use :

> *Take a sequence of m urns and a set of n balls, then allocate an urn to each ball. The trials are independent of each other; at each trial all urns have the same probability $1/m$ to receive the ball. When the n balls are allocated, define the random variable X as the* number of empty urns.

The translation to our database problem now should be obvious : The m urns are the pages; the n balls are the selected objects; the empty urns are the pages that contain no selected object, i.e. that won't be accessed. A few years before Yao, Cardenas [Ca] gave a formula for the expectation of X that is a direct application of the classical urn model :

$$E[X] = m\left(1 - \left(1 - \frac{1}{m}\right)^n\right). \tag{2}$$

Much is known for this model : exact formulae for the moments of X and for the probability distribution, asymptotic normality and convergence towards a Gaussian process in what Kolchin et al. [KSC] call the *central domain,* where the ratio n/m is constant, or at least belongs to an interval $[a_1, a_2]$, $0 < a_1 \leq a_2 < +\infty$. However, there is a difference between the database situation we want to model and the urn model : We have imposed no limitation on urn capacities, while

pages have a finite capacity b. Assuming that pages have unbounded capacity is not realistic from a database point of view, and we turn now to an extension of the model where *each urn can receive no more than b balls.*

What happens to the assumptions of independence and uniformity when considering bounded urns? It seems obvious that, if urns have a finite number b of cells, each of which can receive exactly one ball, and if at some point an urn U_1 is still empty while an urn U_2 has received one or more balls, the probability that the next ball will be allocated to U_1 is no longer equal to the probability that it will be allocated to U_2. However, we may still consider that, at any time, the *empty cells* have the same probability to receive the next ball; hence the probability of any urn at that point is proportional to the number of its empty cells. Now the independence assumption becomes : The allocation of any ball into a *cell* is independent of former allocations. In other words, independence and uniformity still hold, but for cells instead of urns. Returning to our database formulation, a cell is one of the objects of the set $\mathcal{E}$, and the objects that are selected for the query are chosen independently of each other, and independently of their placement on pages, which is exactly the usual assumption.

2.3 Analysis of the bounded-capacity urn model.

Let X be the random variable *number of selected blocks, or of non empty urns.* We introduce now the generating function enumerating the set of possible allocations of balls into urns, which will give us a tool for studying the probability distribution of X. Define $N_{l,n}$ as the number of allocations of n balls into l of the m urns, in such a way that none of the l urns is empty, and define $F(x,y) = \sum_{l,n} N_{l,n}\ x^l\ y^n/n!$: We use the variables x to "mark" the non empty urns, and y to mark the balls. The probability that an allocation gives l urns with at least one ball, conditioned by the number n of balls, the average number of non empty urns, and its variance all can be obtained by extracting suitable coefficients of the function $F(x,y)$ or of some of its derivatives. We compute $F(x,y)$ by symbolic methods, and obtain, when desired, approximate expressions of coefficients by asymptotic analysis; see for example [FS] for an introduction to these tools. We have that

$$F(x,y) = \left(1 + x((1+y)^b - 1)\right)^m . \tag{3}$$

The probability that l blocks are selected, the average number of selected blocks and its variance are obtained from the coefficients of $F(x,y)$ and of its derivatives :

$$\mathbb{P}\mathrm{r}\ (l/n) = \frac{[x^l y^n]F(x,y)}{[y^n]F(1,y)} = \frac{\binom{m}{l}}{\binom{p}{n}} \sum_{i=0}^{l} (-1)^{l-i} \binom{l}{i} \binom{ib}{n}; \tag{4}$$

$$E[X] = m\left(1 - \frac{\binom{p-b}{n}}{\binom{p}{n}}\right); \tag{5}$$

$$\sigma^2[X] = m(m-1)\frac{\binom{p-2b}{n}}{\binom{p}{n}} - m^2\left(\frac{\binom{p-b}{n}}{\binom{p}{n}}\right)^2 + m\frac{\binom{p-b}{n}}{\binom{p}{n}}. \tag{6}$$

For large m and n, the formulae (5) have an approximate expression in terms of the ratio n/m. Although it is possible to consider an arbitrary relationship between the orders of growth of n and m, we shall limit ourselves to the case where n/m is constant; in other words we are in the central domain of Kolchin et al. We give below approximate values for the moments $E[X]$ and $\sigma^2[X]$; we shall see later on (see Section 2.4) that these approximations are very close to the actual values.

$$E[X] \sim m\left(1 - \left(1 - \frac{n}{bm}\right)^b\right); \tag{7}$$

$$\sigma^2[X] \sim m\left(1 - \frac{n}{bm}\right)^b \left(1 - \frac{bn}{bm-n}\left(1 - \frac{n}{bm}\right)^b\right). \tag{8}$$

What about the limiting distribution in the central domain? Applying former results [Ga], we see that the limiting distribution of the random variable X exists, and is a Gaussian distribution whose mean and variance satisfy the formulae (7) and (8).

2.4 Numerical applications.

We first evaluate the possibility of using the approximate formula (7) instead of Yao's exact formula, whose computation is much more complicated and requires a time that cannot be neglected. We have fixed the total number of blocks at $m = 500$, and varied the size b of the blocks. For comparison purposes, we have also indicated the values corresponding to the infinite case, which give a lower bound for the expected number of blocks. Table 1 gives the numerical results, which are also the basis for the plots of Figure 1. The exact and approximate values are very close to each other; as a consequence, in the sequel we shall no longer bother with the exact values, but rather use the approximate expressions. However, as was already noticed by Yao, the infinite urn case is *not* a good approximation, at least for low sizes b of blocks.

Our next objective was to check the quality of the Gaussian approximation. We give in Figure 2 the exact distribution of the number of selected blocks, as well as the Gaussian distribution with the same mean and variance. As the former values of m and n would have given us some extremely low probabilities, we have chosen here $m = 100$ and $b = 5$, with a total of $n = 200$ balls. For these rather small values of m and n, we can check that the approximation by a Gaussian distribution is already quite good; of course it is still better for larger values of the parameters m and n.

n	Yao (exact)			Yao (approximate)			Unbounded urns
	b=2	b=5	b=10	b=2	b=5	b=10	(Cardenas)
200	180.0	170.5	167.7	180	170.4	167.5	164.9
400	320.1	291.0	282.9	320	290.8	282.8	275.5
600	420.1	373.3	360.9	420	373.2	360.7	349.5
800	480.0	427.4	412.7	480	427.3	412.5	399.2
1000	500	461.2	446.4	500	461.1	446.3	432.4

Table 1. Number of selected blocs, for a uniform distribution and several sizes of blocks

We can now be reasonably sure that the random variable *number of expected blocks* behaves as a Gaussian distribution, which has some important consequences: For a Gaussian distribution, we can quantify precisely the probability that we are at a given distance from the average value E, using the variance σ^2.

It is worth noting that our asymptotic formulae make for very quick computations, in contrast to Yao's formula. For $n = 1000$ the computation of Yao's formula requires 2.5s, the bound obtained by assuming that the urns are infinite is 0.15s, and the asymptotic formula is obtained in at most 0.017s. Our formula need a constant number of operations when n and m increase. Yao's formula requires n divisions and $n - 1$ multiplications for the same computation.

3 How well-fitted to object databases is Yao's formula?

It may happen that the database has different block sizes, or that the objects in the database are clustered according to the values of some attribute; then the probability that a block is selected is no longer uniform, and Yao's formula no longer gives a valid estimation of the average number of selected blocks. An extension of Yao's formula to deal with such a situation was proposed by Gardarin et al. in [GGT], where they consider blocks of different sizes in an object-oriented database, and suggest that Yao's formula be applied on each group of equal-sized blocks:

Let C be a collection of j partitions, where each partition has p_i objects and requires m_i blocks. Assume that n objects are chosen at random in the collection C; then the average number of selected blocks is

$$\sum_{i=1}^{j} m_i \left(1 - \frac{\binom{p_i - p_i/m_i}{n_i}}{\binom{p_i}{n_i}}\right), \text{with } n_i = n * p_i/p. \tag{9}$$

The proof of this formula relies on the assumption that the placement of objects into blocks is uniform, and that each object has the same probability of beeing selected, independently of the others; as a consequence the number n_i of selected objects in the i-th partition is $n * p_i/p$. We argue that this does not always hold, for example in the simple example we give below.

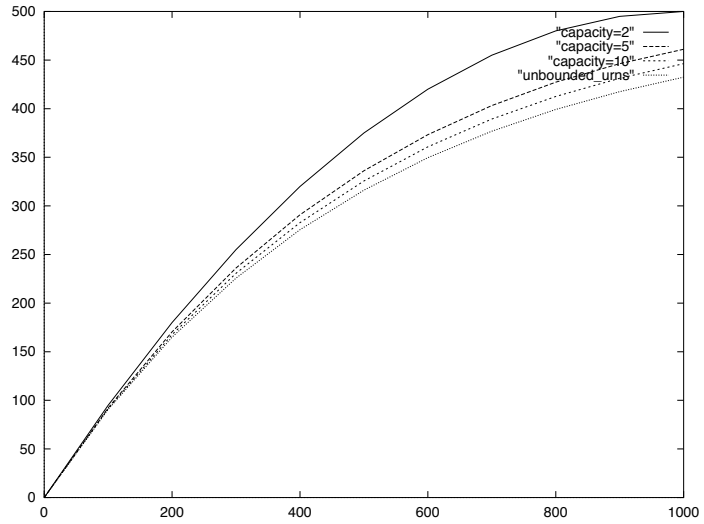

Fig. 1. Number of selected blocks, plotted against the number of selected objects, for several sizes of blocks

Consider a database containing information relative to persons, and the query Q *"Find all people having a salary at least equal to 1000 $"*. If we know the number n of persons satisfying this query, and if the placement of persons on pages (blocks) is random, then the average number of blocks to retrieve is given by Yao's formula. Now assume that the database also contains data relative to cars, that each car has an owner, and that the persons are partitioned according to their owning, or not, a car : some blocks will contain data relative to the persons who own a car, and to their cars, and others blocks will contain data relative to the persons who do not own a car. Then we have introduced a correlation between the salary and ownership of a car : We expect the proportion of people satisfying our query Q to be larger among car owners than among the other people. This means that the blocks containing data on car owners have a greater probability of being selected than the other blocks.

The formula (9) does not take into account this phenomenon, and may lead to biased results, all the more when clustering is done according to an attribute that appears in the query. In the following section, we consider the case where the probability that a block is selected can take a finite number of values, i.e. where we can partition the blocks according to their probability.

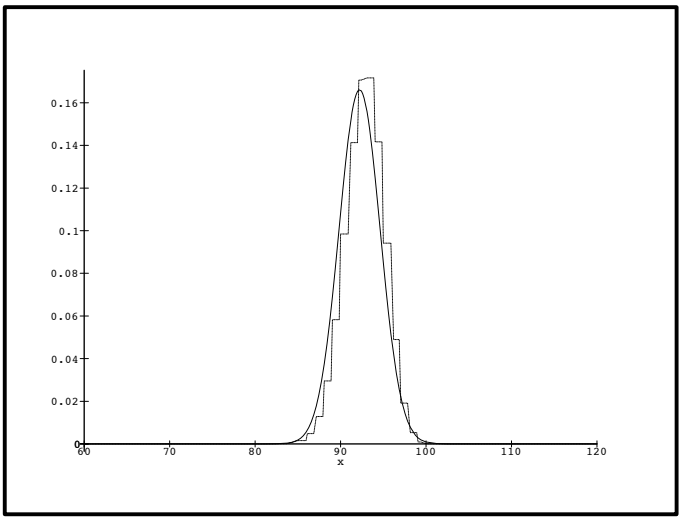

Fig. 2. Probability distribution of the number of selected blocks and of the Gaussian distribution with same mean and variance

4 Piecewise uniform distributions

We now define a model to deal with non-uniformity of database objects or of blocks. We shall assume that *the probability distribution on the blocks is piecewise uniform,* and that *the objects are stocked in blocks that may have different capacities.* This leads us to extend our urn model to allow different probabilities for urns. On each trial, the ball falls into the urn U with a probability $\mathbb{P}\mathrm{r}(U)$. We should mention at once that this assumption does not contradict our former assumption that urns have a finite capacity : The *conditional* probability that an urn U receives the next ball, knowing that it already holds j balls, is of course related to j!

We assume that there are i distinct kinds of blocks ($1 \leq i \leq j$); there are m_i blocks of the kind i, each of which holds b_i objects. We translate this into an urn model, where the urns of type i each have b_i cells, and where each cell can hold at most one ball.

The total number of blocks (urns) is $m = \sum_{1 \leq i \leq j} m_i$; the cumulated number of objects (balls) in blocks of type i is $m_i b_i$, and the total number of objects in the database is $p = \sum_{1 \leq i \leq j} m_i b_i$. We shall denote by π_i the probability that a selected object is stocked in a block of type i; of course $\sum_{1 \leq i \leq j} \pi_i = 1$. This can be illustrated using the example of Section 3, where we separate the persons owning a car from those who don't: π_1 would be the probability of having a car, knowing that the person earns more than 1000 \$: $\pi_1 = \mathbb{P}\mathrm{r}\,(Car/Sal > 1000\ \$)$ and $\pi_2 = \mathbb{P}\mathrm{r}\,(\rceil Car/Sal > 1000\ \$) = 1 - \pi_1$.

4.1 Theoretical results.

In this part, we characterize the probability distribution of the random variable *number X of selected blocks,* conditioned by the number n of selected objects; a brief sketch of the mathematical proofs can be found in the section 4.3.

When we consider two types of blocks ($j = 2$), we have exact expressions for the mean and variance of X :

Proposition 41 *Let*

$$f(m_1, m_2) := \sum_{i=0}^{n} \binom{b_1 m_1}{i} \binom{b_2 m_2}{n-i} \left(\frac{\pi_1 \, b_2 \, m_2}{\pi_2 \, b_1 \; m_1} \right)^i . \tag{10}$$

The average number of selected blocks, conditioned by the number n of selected objects, is

$$E[X] = m - \left(m_1 \frac{f\,(m_1 - 1, m_2)}{f\,(m_1, m_2))} + m_2 \frac{f\,(m_1, m_2 - 1)}{f\,(m_1, m_2)} \right). \tag{11}$$

This formula can be extended to give the variance, and to include general j [GN2], but the usefulness of such an extension may be discussed : We have seen that exact computation of Yao's formula, which involves a binomial coefficient, is already costly, and we have here sums of products of such terms! We shall rather concentrate our efforts on obtaining asymptotic expressions, which can be computed with a reasonable amount of effort, and which give good accuracy in many cases, as we shall see in Section 4.2.

We now turn to the asymptotic study, and notice at once that, in difference to the uniform case where we had to deal with two variables n and m, we now have to deal with $j+1$ variables : the number n of selected objects and the numbers m_i of blocks of type i, for the j types. These variables may grow at different rates; moreover we have $2j$ parameters π_i and b_i. We shall limit ourselves to the easiest generalization of the uniform case : *The number m_i of blocks of each type is proportional to n.* Under this assumption, we have the following result :

Theorem 41 *When the number of blocks of each type is proportional to the total number n of selected objects, the random variable X asymptotically follows a Gaussian limiting distribution, with mean and variance*

$$E[X] \sim \sum_{i=1}^{j} \frac{m_i}{(1 + \pi_i \rho / b_i m_i)^{b_i}}; \tag{12}$$

$$\sigma^2[X] \sim \sum_{i=1}^{j} \frac{m_i}{(1 + \pi_i \rho / b_i m_i)^{b_i}} \left(1 - \frac{1}{(1 + \pi_i \rho / b_i m_i)^{b_i}} \right) \tag{13}$$
$$-\rho \, \frac{\left(\sum_{i=1}^{j} \frac{\pi_i}{(1+\pi_i \rho / b_i m_i)^{b_i+1}} \right)^2}{\left(\sum_{i=1}^{j} \frac{\pi_i}{(1+\pi_i \rho / b_i m_i)^{2}} \right)^2}.$$

In these formulae, ρ is a function of n, and is the unique positive solution of the equation in y: $\sum_{i=1}^{j} \pi_i y/n(1+\pi_i y/b_i m_i) = 1$.

4.2 Numerical applications.

We have first considered two types of urns ($j = 2$), with $m_1 = 100$ and $b_1 = 10$ for the first type, $m_2 = 200$ and $b_2 = 15$ for the second type. We have also considered different values for the probabilities π_1 and π_2 that an object belongs to a block of the first or second type. Table 2 gives the average value of X in several cases : The first three columns give the exact values for different π_i, the next three columns the approximate values under the same assumptions, and the last column is obtained by the formula (9) of Gardarin et al (which does not allow us to consider different probabilities π_i). When the number n of selected objects is relatively low, the last formula gives results that may be considered as valid in the first case, but that differ sensibly from the actual ones in the other cases, where the probability distribution is more biased. (For n greater than 1500, almost all blocks are selected anyway, and the different formulae give comparable results.) Another noteworthy remark is that, here again, the asymptotic approximation is of quite good quality, and that the time for computing this approximation is far lower than the corresponding time for the exact case.The time necessary to compute $f(m_1, m_2)$, and then $E[X]$, is $O(n)$ for the exact formula, and $O(1)$ for the asymptotic formula approximation. For the numerical examples we have considered (see Table 2), the computation time of the asymptotic formulae is at most 0.017s, the formula of Gardarin at al. (which leads to somewhat imprecise results) requires a few seconds, and the time for the exact formula can be as much as a few minutes! As a consequence, we have used the approximate formulae for the curves of Figure 3.a, which presents the average number of selected blocks for probabilities $\pi_1 = 9/10$ and $\pi_2 = 1/10$; we have also plotted the estimation from Gardarin et al., which is always larger than the exact result.

n	Exact values			Approximate values			Uniformity
	$\pi_1 = 1/4$ $\pi_2 = 3/4$	$\pi_1 = 4/5$ $\pi_2 = 1/5$	$\pi_1 = 9/10$ $\pi_2 = 1/10$	$\pi_1 = 1/4$ $\pi_2 = 3/4$	$\pi_1 = 4/5$ $\pi_2 = 1/5$	$\pi_1 = 9/10$ $\pi_2 = 1/10$	
200	147.6	122.2	108.1	147.4	122.1	108.0	147.8
400	224.1	178.3	148.5	224.0	178.4	148.5	224.3
600	263.0	218.2	182.7	262.8	218.2	182.7	263.1
800	282.3	249.7	218.7	282.2	249.6	218.7	282.4
1000	291.7	272.5	251.8	291.7	272.4	251.8	291.8
1200	296.2	286.8	276.1	296.2	286.7	276.1	296.3
1500	298.9	296.6	294.0	298.9	296.6	293.9	298.9
2000	299.9	299.8	299.7	299.9	299.8	299.7	299.9

Table 2. Average number of selected blocks for two types of blocks with different probabilities

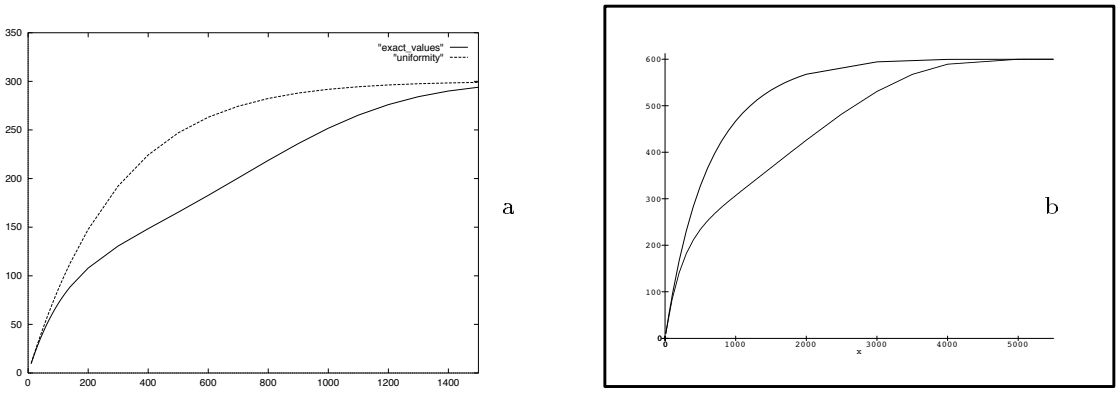

Fig. 3. a: Average number of selected blocks, for two types of blocks with probabilities $\pi_1 = 9/10$ and $\pi_2 = 1/10$, plotted against the same number under a uniform distribution. b: Three types of blocks

The average value of X is a simple function of the parameter ρ, which is obtained by solving a polynomial equation of degree j whose coefficients are expressed with the parameters n, m_i, b_i and π_i. If a comparatively simple expression of ρ exists when $j = 2$, such a formula either becomes quite complicated, or does not exists, for larger j, and we then use a numerical approximation of ρ. We present in Figure 3.b results for three types of blocks, which require a numerical approximation of the solution ρ. Here again, we have plotted the results against the result from Gardarin et al. in [GGT], and have checked that our formulae give a lower average number. We can check that the simplification that allowed them to compute their formula, by making strong assumptions of uniformity, gives a result that is always greater than the exact one; such a systematic overestimation, as noted long ago by Christodoulakis [Ch2], can lead to errors in the further process of optimization.

4.3 Sketch of proofs.

We give here the main ideas and the sketch of the proofs of our theorems; the interested reader can find the complete proofs and detailed computations in [GN2]. We start from the generating function $F(x, y)$, with x marking the number of non empty urns, i.e. of selected blocks, and y marking the total number of balls in the m urns :

$$F(x,y) = \prod_{i=1}^{j} \left(x + (1+\pi_i\, y)^{b_i} - 1\right)^{m_i}. \tag{14}$$

The right coefficients of the generating function and of its derivatives give us the desired information on the probability distribution, for example its value at some point, its average value and its variance. To obtain the limiting distribution, we use Levy's theorem on the convergence of characteristic functions. The

characteristic function of the distribution of X, conditioned by the number n of selected objects, is obtained from the quotient of the two coefficients $[y^n]F(x,y)$ and $[y^n]F(1,y)$. We approximate these coefficients by a saddle point method, treating x as a parameter; in a second step we choose $x = e^{-t/\sigma[X]}$ to obtain the characteristic function, and check that it converges for large n towards the function of a Gaussian distribution of suitable expectation and variance.

An alternative proof is obtained by applying recent results of Bender and Richmond [BR] to our generating function $F(x,y)$.

5 Discussion and extensions

We have shown that an adequate modelization allows us to compute much more than the average number of blocks to be retrieved, in a situation where all the blocks have the same probability to be selected. We have computed the variance of this number and its limiting distribution. In our opinion, the points that are relevant to practical applications are the following : The limiting distribution is Gaussian with mean and variance of the same order, the exact distribution is close to the limiting one for reasonable values of the parameters, and the approximate values of the expectation and variance are given by very simple formulae. This means that, when in a situation where one wants to use Yao's formula, one can use the average value with confidence that the actual number is not too far from it, and that we do not even need to use the exact average value (whose computation, involving binomial coefficients, i.e. factorials, can be costly), but can use a very efficient approximation.

We have also shown that extensions of Yao's formula to non-uniform probabilities are best done by the urn model. For piecewise distributions, we have given the exact and approximate mean and variance, and proved that the limiting distribution is again Gaussian : Once more, we are in a situation where using the expectation instead of the random variable is not likely to induce much error.

Former results on non-uniform distributions are few, and are much less precise than ours. To the best of our knowledge, they have been limited to the expectation of the number of retrieved blocks: For example, Christodoulakis gave in [Ch1] upper and lower bounds for this expectation. Our approch differs from the usual one in that we do not assume independency of the variables under study, although allowing dependencies may lead to greater generality, we still have to define and obtain conditional probabilities (the π_i of Section 4), which may limit the use of our results.

The fact that most of our results are asymptotic, i.e valid for large numbers of blocks and of selected objects, may at first glance seem a restriction. However, numerical computations have shown that the exact and approximate values are very close to each other for parameters as low as a few hundred, which is certainly relevant to most databases. If this is not the case, then one should use the exact formulae, which still give feasible computations for low values of the parameters. If one wishes for a mathematical justification of our approximations, then one

might try and quantify the speed of convergence towards the limiting distribution (reasonably fast in urn models).

From a mathematical point of view, there is still work to be done : If urn models with uniform probabilities (and unbounded urns) are quite well-known, to our knowledge the extensions relative to bounded urns and non-uniform probabilities are not so well characterized. An interesting problem is to define a class of probability distributions on the urns that still give rise to a Gaussian behaviour, and with a variance low enough that the bound on the error when using the average value instead of the random variable remains within acceptable bounds.

References

[BR] A.BENDER and L.B. RICHMOND. Multivariate asymptotique for product of large powers with applications to Lagrange inversion. Technical report, University of Waterloo, April 1998.

[Ca] A.F. CARDENAS. Analysis and performance of inverted data base structures. *Comm. ACM*, 19(5), 1975.

[Ch1] S. CHRISTODOULAKIS. Issues in query evaluation. *Database Engineering*, 5(3):220–223, September 1982.

[Ch2] S. CHRISTODOULAKIS. Estimating block selectivities. *Information Systems*, 9(1):69–79, 1983.

[FS] P. FLAJOLET and R. SEDGEWICK. *An introduction to the analysis of algorithms.* Addison-Wesley, 1996.

[GGT] G. GARDARIN, J.-R. GRUSER, and Z.H. TANG. A cost model for clustered object oriented databases. In *21st VLDB Conference, Zürich, Switzerland*, pages 323–334, 1995.

[Ga] D. GARDY. Normal limiting distributions for projection and semijoin sizes. *SIAM Journal on Discrete Mathematics*, 5(2):219–248, May 1992.

[GN1] D. GARDY and L. NEMIROVSKI. Formule de Yao et modèles d'urnes. In *14st BDA Conference, Hammamet, Tunisia*, (to appear) 1998.

[GN2] D. GARDY and L. NEMIROVSKI. Une application des modèles d'urnes aux bases de données : la formule de Yao et ses extensions. Technical report, Laboratoire PRISM, Université de Versailles, October 1998.

[JK] N.L. JOHNSON and S. KOTZ. *Urn models and their application.* Wiley & Sons, 1977.

[KSC] V. KOLCHIN, B. SEVAST'YANOV, and V. CHISTYAKOV. *Random Allocations.* Wiley & Sons, 1978.

[Ya] S.B. YAO. Approximating block accesses in data base organizations. *Comm. ACM*, 20(4), 1977.