A Better Tool for Query Optimization

Dina Bitton
Bradley Vander Zanden

TR 85-671
April 1985

Department of Computer Science
Cornell University
Ithaca, New York 14853

A Better Tool for Query Optimization

Dina Bitton
Bradley Vander Zanden

Computer Science Department
Cornell University, Ithaca NY 14853

Abstract: When evaluating the performance of a query strategy, one must often estimate the number of distinct values of an attribute in a randomly selected subset of a relation. Most query optimizers compute this estimate based on the assumption that prior to the selection, the attribute values are uniformly distributed in the relation. In this paper we depart from this assumption and instead consider Zipf distributions that are known to accurately model text and name distributions. Given a relation of cardinality n where a non-key atrribute A has a Zipf distribution, we derive both an exact formula and an approximate non-iterative formula for the expected number of distinct A-values contained in a sample of k randomly selected tuples. The approximation is accurate, and it is very easy to compute. Thus it provides a practical tool to deal with non-uniform distributions in query optimization.

1. Introduction

Estimating the number of distinct tuples produced at a certain stage in processing a relational query is an important problem in performance evaluation of database systems and in query optimization [Jarke and Koch 1984]. This estimate provides an indication of the size of a relation obtained after a selection and a projection [Christodoulakis 1983], it is useful in the evaluation of semi-join strategies [Kerschberg et al. 1980] and it also provides a basis to evaluate the cost of duplicate tuple elimination [Bitton and DeWitt].

In all of the above instances, an estimate of the relevant performance parameter has been obtained by reducing the problem to a *multiset sampling problem*. For the sake of clarity, we will deal with estimating the number of distinct tuples that satisfy the following relational query. Suppose that starting with a relation R of cardinality n, where A and B are two independent attributes we select k tuples based on their B value and then project on attribute A. Initially, in R, the values of A constitute a multiset of size n(that is a set of n elements among which only $m<n$ are distinct). The selection can be viewed as a random sample of k elements from this multiset. Then the size of the projection on A can be estimated as the expected number of distinct elements in the sample. We will use the notation X_{nk} for the random variable representing the number of distinct elements occurring in a random sample of size k from a multiset of size n. A number of previous studies have dealt with estimating the expected value of X_{nk} (denoted as $E[X_{nk}]$). Often, the problem has been equivalently

formulated as estimating the number of disk accesses required to retrieve k records, since records residing on the same block can be treated as identical elements of a multiset. The proposed estimates differ depending on whether sampling without replacement [Yao 1977] or with replacement [Cardenas 1975; Cheung 1982] is assumed, or on the kind of approximation used to compute the actual values of $E[X_{nk}]$ [Whang et al 1983]. Indeed the closed formulas obtained for these estimates often contain complex combinatorial terms, and would be of very limited use without approximations that have good computation characteristics [Whang et al 1983, Luk 1983].

Like many other performance parameters, the number of distinct tuples is usually estimated under the assumption that the attributes are uniformly distributed. However, a number of recent studies show that this assumption may lead to estimates that deviate substantially from actual values [Christodoulakis 1983, Piatetsky-Shapiro and Connell 1984]. These findings have been reinforced by theoretical results that prove that uniformity assumptions systematically lead to pessimistic estimates for the number of distinct tuples and a number of other related performance parameters [Christodoulakis 1984]. As an alternative to a theoretical uniform distribution, *Zipf distributions* have been shown to model well the data in certain large databases [Fedorowicz 1982 and 1984] when non-unique attribute values or text words are stored.

The results of these studies indicate that two areas of research warrant further work:

(1) For the most part, previous estimates only apply to the uniform cases. That is, results were derived based on the assumption that attribute values are uniformly distributed (or records are uniformly distributed among the blocks), that the sample retrieved is random and that all tuples are selected with the same probability.

(2) Even when closed formulas are obtained, computing the combinatorial terms involved in these formulas requires a high number of floating point operations. Thus approximations must be derived to increase the usability of these results by query optimizers. The work in [Whang et al 1983] constitutes a first step in this direction.

In this paper, we depart from the uniform distribution assumption and consider a family of *Zipf distributions*. We derive both an exact formula and an accurate approximation, that is amenable to fast computation, for the expected number of distinct tuples $E[X_{nk}]$ in the case where the n initial attribute values have a Zipf distribution. For the sake of clarity, we have chosen to restrict ourselves to the context of attribute values in a relation (although our results are equally applicable to the evaluation of disk accesses). Thus we are concerned with the distribution of the random variable X_{nk} that represents the number of distinct attribute values present in a random subset of k tuples sampled from a relation of size n, given that the attribute values inititially obey a Zipf distribution. To deal with this distribution, we propose an approximation of the form

$$E[X_{nk}] \approx A + \sqrt{B + \frac{k}{n}}$$

for the expected value of X_{nk}, and evaluate the error-rate associated with this approximation.

The remainder of this paper is organized as follows. In Section 2, we briefly describe Zipf distributions. In Section 3 we derive an exact formula for $E[X_{nk}]$. In Section 4, we describe a least-square curve-fitting experiment where we derive a good approximation for $E[X_{nk}]$ in the case corresponding to a relation with n = 1,000,000 tuples and 100,000 distinct attribute values. In Section 5, we generalize the result of this experiment to a broad range of relation sizes and Zipf distributions by establishing certain properties of these distributions. We briefly summarize our results in section 6.

2. Zipf Distributions

Assuming that the attribute values are uniformly distributed is often unrealistic. For example if an attribute represents salary or age brackets of employees in a company, it is clear that the majority of employee records will fall within certain brackets while other brackets will only represent a small number of employees. In this section, we deal with other theoretical distributions, the Zipf distributions, that have been shown to model closely the occurrence of words in text data and the record fields in certain large databases [Fedorowicz 1982 and 1984]. These distributions are based on the observation that when values are ranked according to their frequency of occurrence (the most frequent values first), there is a constant relationship between the rank and the frequency of occurrence. Zipf's first law simply states that *rank times frequency = constant*. Zipf's second law [Booth 1967] divides the distribution into frequency counts and handles better low frequency terms. In a multiset of n elements distributed according to Zipf's second law, the number of different attribute values that occur j times is defined by the formula:

$$I_j = (rn)^{1/\alpha}\left[\frac{1}{j^{1/\alpha}} - \frac{1}{(j+1)^{1/\alpha}}\right] \qquad (2.1)$$

If m is the number of distinct attribute values, the constants r and alpha are related by

$$m = (rn)^{1/\alpha}$$

r is called the "richness" constant, as it has been observed that it measures the richness of the vocabulary used in a text [Booth 1967]. Alpha has no intuitive meaning, but it is used to obtain a more accurate approximation to the distribution of the observed data. Several studies have empirically demonstrated that α=1 provides a good approximation in many applications [Booth 1967; Fedorowicz 1984]. In the remainder of this paper we will only consider the Zipf distribution that corresponds to α = 1, thus

$$I_j = (rn)\left[\frac{1}{j} - \frac{1}{(j+1)}\right] = \frac{m}{j(j+1)} \tag{2.2}$$

3. An exact formula for E[X$_{nk}$]

Consider a multiset of n elements, m of which are distinct. Assuming that random samples are drawn from the multiset without replacement (i.e., an element that has been drawn is not replaced in the multiset before other elements are drawn), the number of distinct values X in a sample of size k may be expressed as the sum of m identically distributed random variables:

$$X = X_1 + X_2 + \ldots + X_m, \; where\ X_i = \begin{cases} 1, & \text{if the i}^{\text{th}}\text{ attribute value is selected} \\ 0, & \text{if the i}^{\text{th}}\text{ attribute value is not selected} \end{cases}$$

If the ith value occurs f_i times then

$$P(X_i = 0) = \frac{\binom{n-f_i}{k}}{\binom{n}{k}} \tag{3.1}$$

Thus the expected value of X is

$$E[X_{nk}] = \sum_{i=1}^{m} P(X_i = 1) = \sum_{i=1}^{m}\left[1 - \frac{\binom{n-f_i}{k}}{\binom{n}{k}}\right] \tag{3.2}$$

If the values are uniformly distributed all the f_i are equal to n/m and the sum reduces to Yao's formula:

$$E[X_{nk}] = m\left[1 - \frac{\binom{n-f}{k}}{\binom{n}{k}}\right] \tag{3.3}$$

However, it is unlikely that a closed formula could also be derived for Zipf distributions. By reordering the terms in the sum (3.2) by frequency of occurrence we derive the following theorem.

Theorem 1: If a multiset of n elements obeys the Zipf distribution (2.2), the expected number of distinct values in a sample of size k is

$$E[X] = \sum_{j=1}^{c} \frac{m}{j(j+1)} \left[1 - \frac{\binom{n-j}{k}}{\binom{n}{k}} \right] \tag{3.4}$$

where c ≈ exp(1/r + .423).

Proof: Equation (3.4) can be obtained from (3.2) by grouping terms with the same frequency f_j, and noting that I_j represents the size of this group. Then the summation becomes

$$E[X] = \sum_{j=1}^{c} I_j \left[1 - \frac{\binom{n-j}{k}}{\binom{n}{k}} \right]$$

and (3.4) follows by substituting expression (2.2) for I_j. The constant c is derived by noting that:

$$\sum_{j=1}^{c} jI_j = n \Rightarrow \sum_{j=1}^{c} \frac{m}{j+1} = n$$

Using the approximation

$$\sum_{j=1}^{c} \frac{m}{j+1} \approx m(ln(c+1) + .577 - 1) \approx m(ln\, c - .423)$$

and some algebraic manipulation, we obtain the desired result.

4. Usability and approximations

The formula derived in Section 3 for the expected value of the variable X_{nk} includes combinatorial terms whose evaluation for given values of the parameters n and k generally requires a large number of floating point multiplications. On the other hand, the usability of this formula largely depends on the ease of computing it. In particular, when intended to support a query optimizer's decision, it must be repetititvely computed for alternative query strategies and for each relational operation implied by a strategy. The upper bound on the time that can be spent for this computation must be low enough to make the optimization process desirable. Otherwise, choosing a strategy at random might be more efficient than determining what the fastest strategy is. Thus, exact formulas such as (3.3) and (3.4) have little practical value in the context of query optimization unless they can be approximated by formulas that are more amenable to fast computation.

4.1. A least-square approximation

A number of previous papers have dealt with finding approximations for $E[X_{nk}]$ in the uniform distribution case [Yao 1977, Whang et al. 1983]. The motivation for finding a good approximation for this expected value in the case of a Zipf distribution is even stronger, given that the exact formula we derived (3.4) is not in closed form and involves more complex terms. This problem was addressed in a recent work by Luk [Luk 1983], where an attempt was made to approximate a Zipf distribution formula for $E[X_{nk}]$ using the formula for a uniformly distributed multiset with the same number of distinct values. However, for large files, the relative error was found to be unacceptably large (generally greater than 50% and in some cases close to 100%).

Our approach to this approximation problem is very different. Rather than trying to find another distribution that approximates the Zipf distribution but is more likely to lead to a simple exact formula for $E[X_{nk}]$, we investigated the possibility of directly approximating the formula that we derived for the Zipf distribution. We began by plotting some values of $E[X_{nk}]$ as a function of the sample size k, and realized that the shape of the graph resembled sqrt(k) (Figure 1). This led us to the conjecture that we could find coefficients a_1,a_2,a_3 such that:

$$k = a_1(E[X_{nk}])^2 + a_2E[X_{nk}] + a_3 \tag{4.1}$$

or equivalently

$$E[X_{nk}] = \frac{-a_2}{2a_1} + \sqrt{\left(\frac{a_2}{2a_1}\right)^2 - \frac{a_3}{a_1} + \frac{k}{a_1}} \tag{4.2}$$

In order to determine the approximation coefficients, we used a curve-fitting technique [Yule 1950].

4.2. Experiments with empirical data

We generated data for our initial approximation by choosing values corresponding to the Zipf distribution with m = 100,000 and r = .1 (thus n = 1,000,000). For this multiset, we varied the sample size from 500 to 1,000,000 by increments of 500 and derived the corresponding values of $E[X_{nk}]$ from the exact formula (3.4). We then ran a series of experiments in which the number of points fitted was varied between 50 and 200, and the first fitted point was varied between 1 and 6. Table 1 shows the results of these experiments; column 2 in the table shows the optimal number of points fitted, column 3 shows the size of

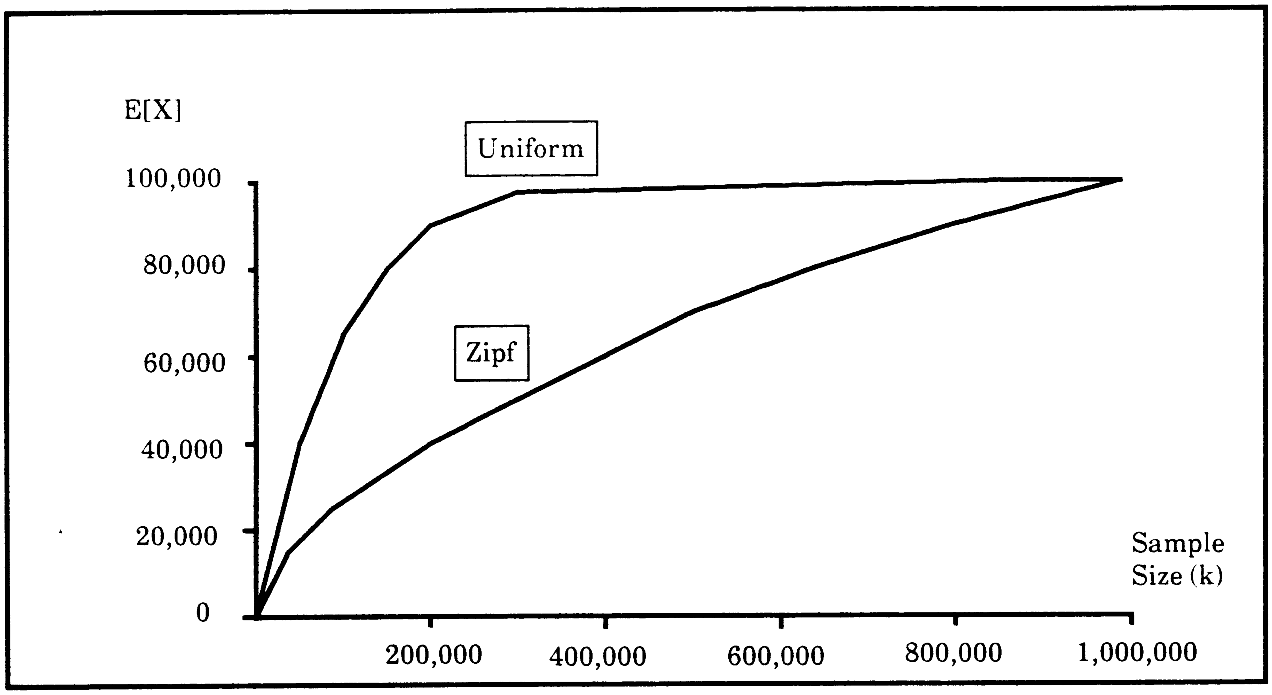

Figure 1: $E[X_{nk}]$ as a function of k for a Zipf and a Uniform distribution (r = .1).

the samples for which the error fell below 3%, and column 4 shows the maximum error that occurs after the error first falls below 3%.

Table 1
Relative Errors for Curves that Approximate E[X] Under Zipf Assumptions

Start Point	Points Fitted	Under 3%	Maximum Error
1	133	3000	2.99
2	140	3000	2.75
3	144	3000	2.75
4	148	3000	2.74
5	153	2500	2.85
6	163	2500	3.24

The approximation we chose utilizes a curve with a starting point of 4 and 148 fitted points. For this curve the values of the a's are $a_1 = 7.731e\text{-}07$, $a_2 = 1.99e\text{-}02$, and $a_3 = -7.125$.

5. A general approximation formula

The accuracy exhibited by our least-square approximation in the above experiments was so high that we felt it could not depend on the particular choice of values for the multiset size or the richness parameter. Indeed, our empirical results can be generalized to a range of multiset sizes and richness parameters. This generalization is based on the following two propositions that we derived for a Zipf distribution. In the following we refer to the Zipf distribution defined in Section 2 and assume sampling without replacement.

Proposition 1: If two multisets of size n_1 and n_2 have a Zipf distribution with the same richness parameter r, then for random samples of size k_1 and k_2 such that $k_1/n_1 = k_2/n_2$, the expression

$$\frac{m_2}{m_1} E[X_{n_1 k_1}] \qquad (5.1)$$

is a good approximation for $E[X_{n_2 k_2}]$.

Analytically validating this approximation would require finding absolute upper bounds for complex combinatorial expressions. Such bounds are usually difficult to find [King 1984]. An alternative approach is to conduct an exhaustive computer evaluation which examines the validity of the approximation over a wide range of parameter values. We have adopted the latter method and in Appendix A we present the detailed results of such an experiment. Here we briefly summarize the findings. Representative graphs of the results are plotted in Figures 2 and 3. As the graphs demonstrate, the approximation in (5.1) improves as the parameters n_1, n_2/n_1, and r increase. The analysis demonstrated that the approximation yields a maximum relative error of 8.12% over the range of parameter values that were tested. This error occurred at the values $n_1 = 10000$, $n_2/n_1 = 100$, $k_1/n_1 = .01$, $r = .25$. However, the values of $E[X_{n_2k_2}]$ and $(n_2/n_1)E[X_{n_1k_1}]$ still agreed in the first decimal digit (1.0 versus .92). In general the relative error of the approximation in expression (5.1) dropped below 2% for sample proportions greater than .01, 1% for sample proportions greater than .05, and approached 0 for sample proportions greater than .25. However, for larger values of n_1, n_2/n_1, and r, the relative error declined much more dramatically (as Figures 2 and 3 illustrate). In these cases, sample proportions as low as .001 produced relative errors less than .5%. Overall the results of the computer evaluation indicate that expression (5.1) holds over a wide range of values for the parameters n_1, n_2/n_1, k_1/n_1, and r.

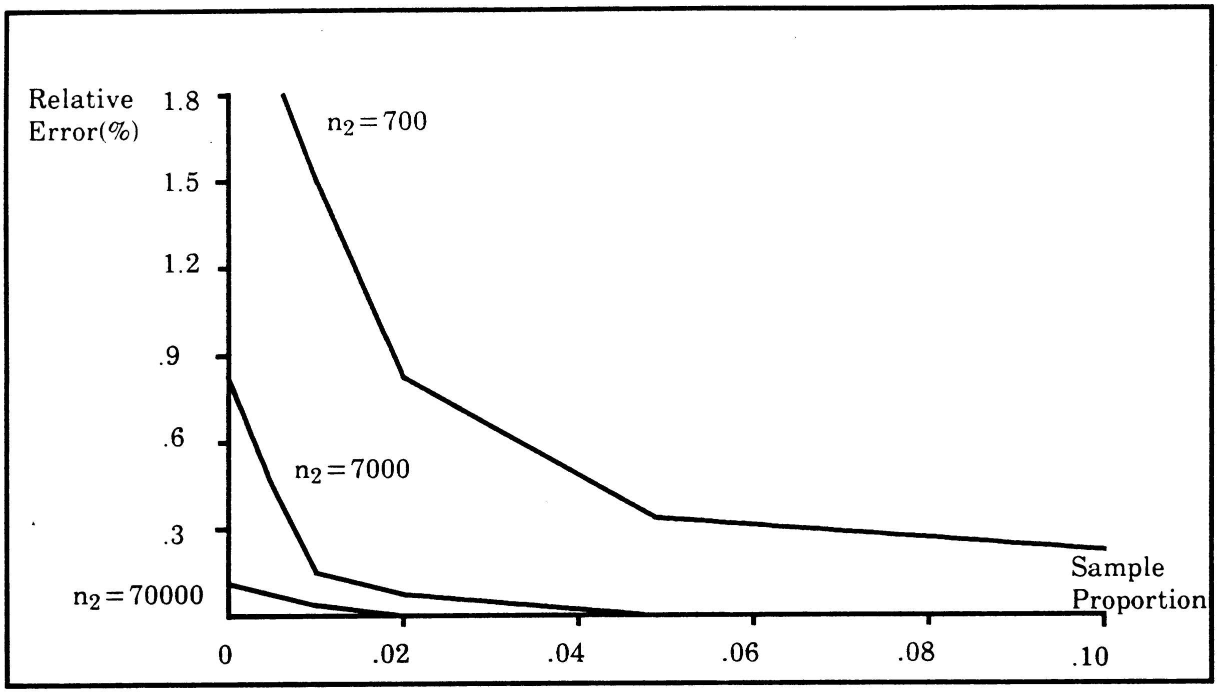

Figure 2: Relative error of the approximation in expression (5.1) when $n_1 = 700{,}000$ and $r = .15$

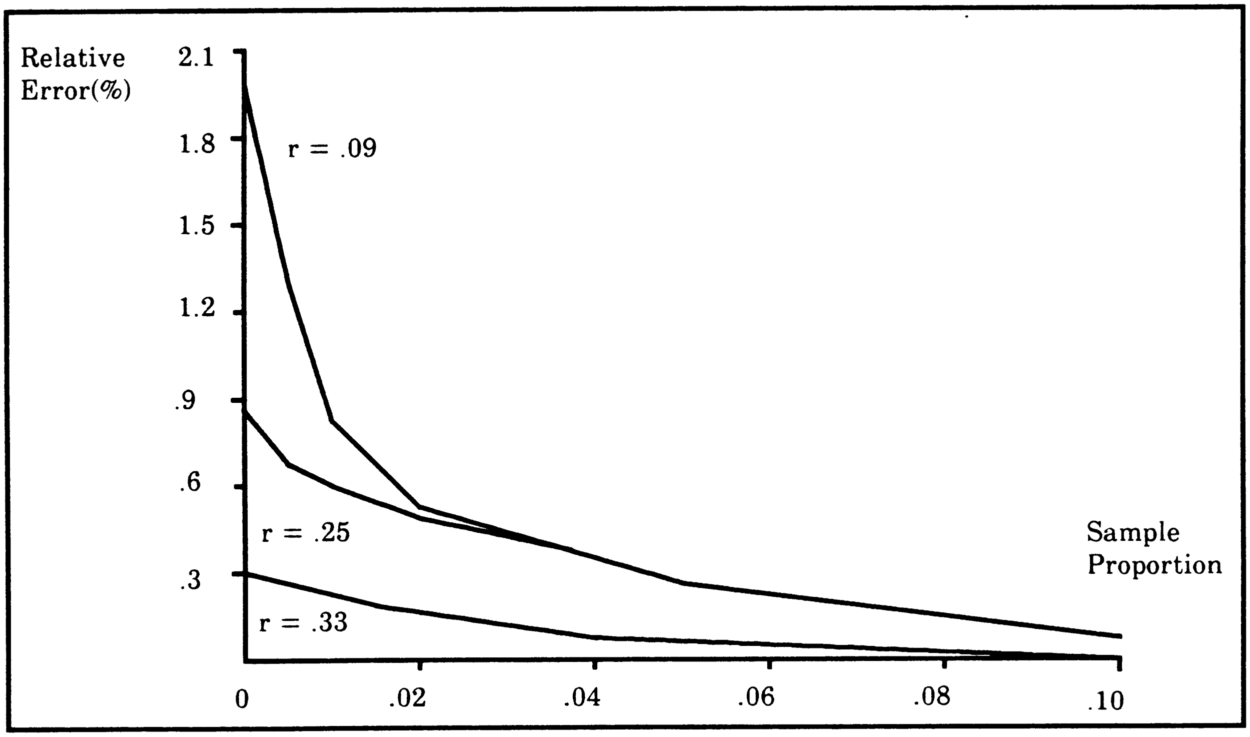

Figure 3: Relative error of the approximation in expression (5.1) when $n_1 = 10{,}000$ and $n_2 = 1{,}000$

Proposition 2: If two multisets of size n have a Zipf distribution with different richness parameters r_1 and r_2 then for large enough samples

$$E_{r_2}[X_{nk}] \approx \frac{m_2}{m_1} E_{r_1}[X_{nk}] + \frac{m_2}{c_1+1} - \frac{m_2}{c_2+1}$$

where $E_{r_i}[X_{nk}]$ represents the expected value of X_{nk} when the richness constant is r_i.

We also used a computer evaluation to demonstrate the validity of this approximation. For $r_1 < r_2$ the expression

$$E_{r_2}[X_{nk}] - \frac{m_2}{m_1} E_{r_1}[X_{nk}] + m_2\left(\frac{1}{c_2+1} - \frac{1}{c_1+1}\right) = \sum_{j=c_2+1}^{c_1} \frac{m_2}{j(j+1)}\left[\frac{\binom{n-j}{k}}{\binom{n}{k}}\right] \qquad (5.3)$$

depends on four parameters n, k, r_1, and r_2. The computer evaluation determined how large k must be, given values for n, r_1, and r_2, in order for the value of (5.3) to be negligible compared with $E_{r_2}[X_{nk}]$. For the purposes of the computer evaluation, we considered the approximation in (5.2) valid if the relative error was less than 3%.

The evaluation demonstrated that only r_2 had a noticeable impact on the required sample proportion. The size of the sample proportion versus r_2 is plotted in Figure 4.

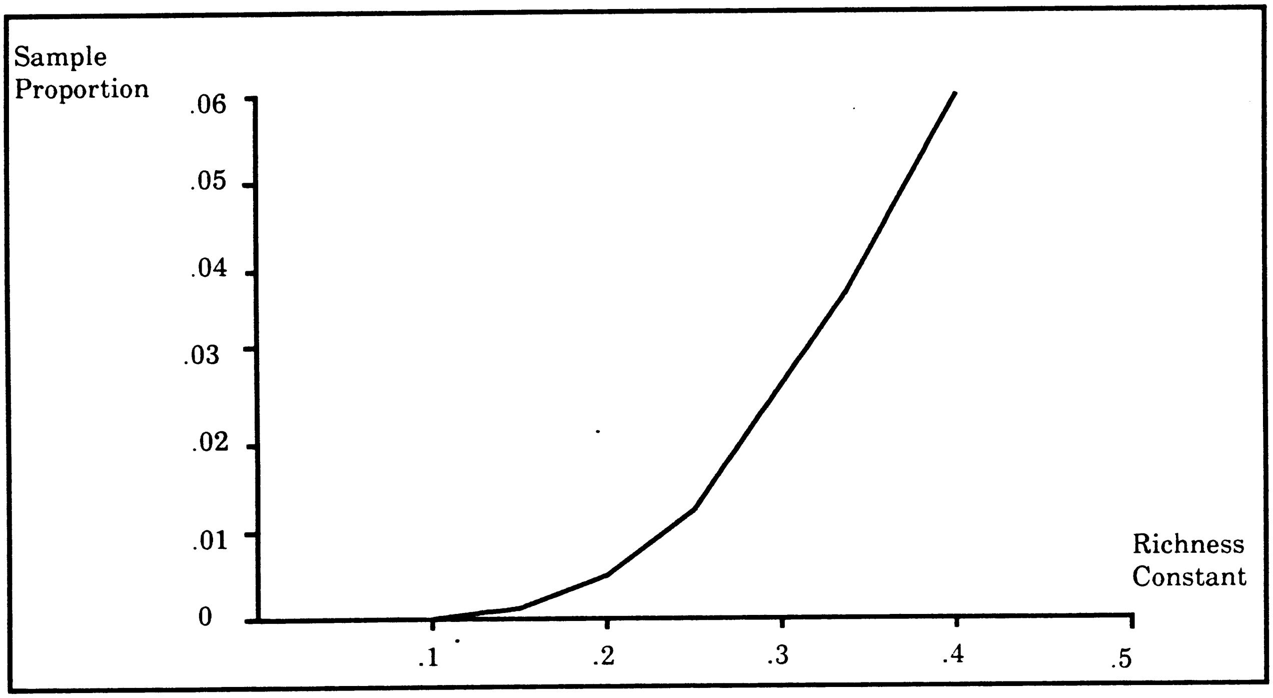

Figure 4: Sample proportion for which the relative error in expression (5.2) is less than 3% (.09 ≤ r ≤ .5). For r = .5, the required sample proportion is .106.

As the figure demonstrates, the sample proportion rises extremely slowly for small r_2 and then starts to rise rapidly for richness constants greater than .3. This result makes sense since small r_2 correspond to large c_2 and large c_2 lead to negligible values for (5.3), even for small sample proportions. On the other hand, large r_2 are associated with small values of c_2 and thus larger sample proportions are required to make the relative size of the sum (5.3) negligible. Appendix B presents a more detailed discussion of the results of the computer evaluation.

Proposition 1 relates two populations with different sizes while Proposition 2 relates two populations with different richness parameters. The combination of these two propositions allows us to generalize expression (4.2) to any population with an arbitrary size and arbitrary richness parameter. Figure 5 illustrates how this process works.

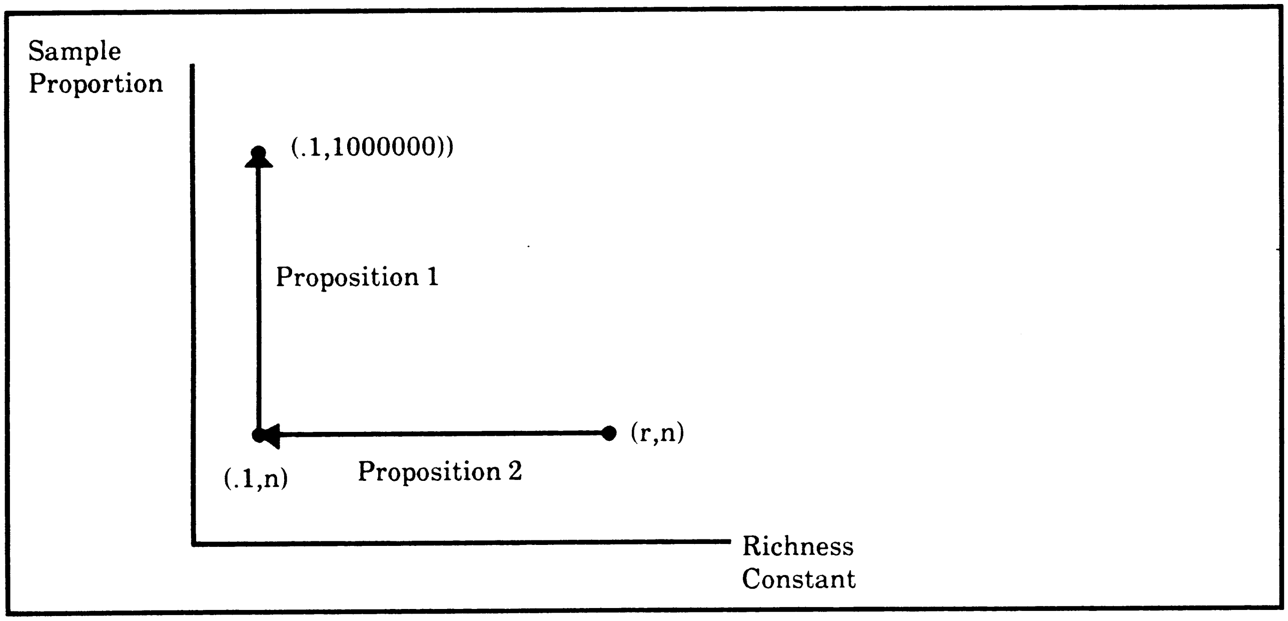

Figure 5: We can get an expression for $E_r[X_{nk}]$ by applying Proposition 2 and then Proposition 1

Suppose we are asked to find the expected number of distinct values in a sample of size k from a Zipf population with size n and richness parameter r. Initially we are located at the point (r,n) in Figure 5. Since we have a good approximation for $E_{.1}[X_{(1,000,000)(k)}]$, we would like to move to the point (.1,1000000). From Figure 5, we see that the first step involves applying Proposition 2 to move from the point (r,n) to the point (.1,n). At the end of this step we obtain

$$E_r[X_{nk}] = \frac{m}{.1n} E_{.1}[X_{nk}] + m\left(\frac{1}{\alpha} - \frac{1}{c}\right) \qquad (5.4)$$

$$where\ \alpha = e^{10.423}\ and\ c = e^{1/r\ +\ .423}$$

Then in the second and final step, we apply Proposition 1 to move from the point (.1,n) to the point (.1,1000000). We first write $E_{.1}[X_{nk}]$ as:

$$E_{.1}[X_{nk}] = \frac{.1n}{100,000} E_{.1}[X_{(1,000,000)(k/n*1,000,000)}] \qquad (5.5)$$

We then substitute expression (4.2) for $E_{.1}[X_{(1,000,000)(k/n*1,000,000)}]$ in (5.5) and substitute the resulting expression for $E_{.1}[X_{nk}]$ in (5.4) to obtain:

$$E_r[X_{nk}] \approx \frac{m}{100,000}\left[\frac{-a_2}{2a_1} + \sqrt{\left(\frac{a_2}{2a_1}\right)^2 - \frac{a_3}{a_1} + \left(\frac{1,000,000}{n}\right)\left(\frac{k}{a_1}\right)}\right] + m\left(\frac{1}{a} - \frac{1}{c}\right) \qquad (5.6)$$

We performed an exhaustive computer evaluation of this approximation using the following parameter values:

- n : 1000, 4000, 7000, 10000, 40000, 70000, 100000, 400000, 700000, 1000000, 4000000, 7000000, 10000000

  r : .09, .1, .15, .20, .25, 1/3, .5

Representative results of the evaluation are plotted in Figure 6.

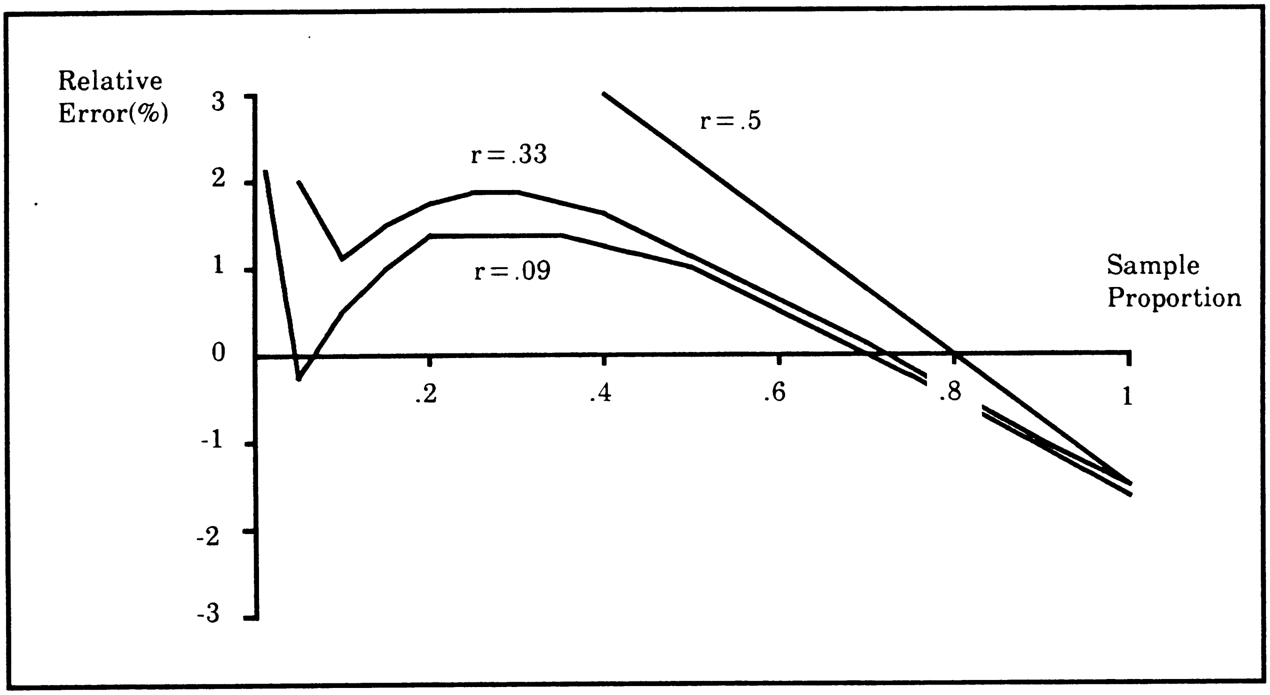

Figure 6: Relative error of expression (5.6) when n = 10,000,000. For r = .5, the relative error drops below 6% when the sample proportion exceeds .15.

Two conclusions can be drawn from the graph. First, the value of the richness constant is the primary determinant of the validity of the approximation. As the richness constant increases, the size of the relative error increases. This behavior is predicted by the proofs of both

Propositions 1 and 2. Second, as long as the sample proportion is greater than .2% of the population, Figure 6 shows the ranges in which it is safe to use (5.6) as an approximation for $E_r[X_{nk}]$. The results of the computer evaluation confirm these conclusions. Thus for a large range of n, r, and k, the approximation in (5.6) is so precise that it obviates the need for a closed form expression.

6. Contributions and future research

In this paper, we have investigated the problem of estimating certain performance parameters that play an important role in query optimization, in the presence of Zipf distributions. In particular, assuming that a non-key attribute in a relation obeys a Zipf distribution, we have obtained estimates for the expected number of distinct values of this attribute that remain after a selection query is applied to the relation. First, we have derived an exact formula for the expected value (Theorem 1). However, because this formula involves a complex sum of combinatorial terms we have proceeded to find an approximation that would be more amenable to computation. We derived this approximation in two steps. First, we applied a least-square technique to fit empirical data. For a multiset of 1,000,000 elements, 100,000 of which were distinct, we showed that $E[X_{nk}]$ behaved as a square root function of the sample size. Then, based on the result of this experiment and on two properties of the Zipf distribution (Propositions 1 and 2), we generalized our square root approximation to arbitrary values of n (the multiset size) and m (the number of distinct elements in the multiset). Furthermore, we proved that for practical values of n and m, and for large enough sample sizes, the relative error associated with this approximation fell below 3%.

These results are applicable to a range of problems that arise in query optimization. In particular, they may be applied to obtain accurate estimates for the number of disk accesses required to process a retrieval query.

7. References

Bitton, D. and DeWitt, D.J. "Duplicate Record Elimination in Large Data Files," *ACM Trans. Database Syst.*, 8, 2, June 1983, 255-265.

Booth, A.D. "A "Law" of Occurrences for Words of Low Frequency," *Information and Control*, 10, 1967, 386-393.

Cardenas, A.F. "Analysis and Performance of Inverted Database Structures," *Commun. ACM*, 18, 5, May 1975, 253-263.

Cheung, T. "Estimating Block Accesses and Number of Records in File Management," *Commun. ACM*, 25, 7, July 1982, 484-486.

Christodoulakis, S. "Estimating Block Transfers and Join Sizes," In *Proceedings SIGMOD 1983 Conference*, ACM, 40-50.

Christodoulakis, S. "Implications of Certain Assumptions in Database Performance Evaluation," *Trans. Database Syst.*, 9, 2, June 1984, 163-186.

Fedorowicz, J.E., "A Zipfian Model of an Automatic Bibliographic System: An Application to MEDLINE," *Journal of the American Society for Information Science*, 33, 4, July 1982, 223-232.

Fedorowicz, J. "Database Evaluation Using Multiple Regression Techniques," in *Proceedings SIGMOD 1984 Conference*, ACM, 70-77.

Jarke, M. and Koch, J. "Query Optimization in Database Systems," *ACM Comput. Surveys*, 16, 2, June 1984, 111-152.

Johnson, N.L. and Kotz, S. *Urn Models and Their Application*, New York, John Wiley and Sons, 1977.

King, A.C. "Some Inequalities for Factorials," *The Mathematical Gazette*, 68, 445, Oct. 1984, 206-208.

Luk, W .S. "On Estimating Block Accesses in Database Organizations," *Commun. ACM*, 26, 11, Nov. 1983, 945-947.

Piatetsky-Shapiro, G. and Connell, C. "Accurate Estimation of the Number of Tuples Satisfying a Condition," In *Proceedings SIGMOD 1984 Conference*, ACM, 256-276.

Whang , K., Wiederhold, G., and Sagalowicz, D. "Estimating Block Accesses in Database Organizations: A Closed Noniterative Formula," *Commun. ACM*, 26, 11, Nov. 1983, 940-944.

Yao, S. B. "Approximating Block Accesses in Database Organizations," *Commun. ACM*, 20, 4, April 1977, 260-261.

Yule, M.A. and Kendall, M.G. *An Introduction to the Theory of Statistics*, London, Charles Griffin and Company Limited, 1950.

APPENDIX A - EVALUATION OF EXPRESSION (5.1)

The validity of the approximation in expression (5.1) depends on four parameters: the size of the first population n_1, the size of the second population n_2, the size of the sample k_1, and j. A close inspection of (5.1) allows us to make several observations about the relationship between the size of the parameters and the size of the approximation error:

1. Small j: As the values of j decrease, the expression

$$\frac{\binom{n_1-j}{k_1}}{\binom{n_1}{k_1}} - \frac{\binom{n_2-j}{k_2}}{\binom{n_2}{k_2}} \tag{A.1}$$

approaches 0. This behavior suggests that the approximation in (5.1) is better for small c.

2. Large n_1/n_2: As the ratio n_1/n_2 approaches 1, expression (A.1) approaches 0 and thus the size of the approximation error in (5.1) decreases.

3. Large k_1/n_1: As the ratio k_1/n_1 approaches 1, expression (A.1) approaches 0 since both of its terms rapidly approach 0. Consequently large sample proportions should improve the approximation.

The validity of these observations was confirmed by the computer evaluation of the relative error in expression (5.1). The values of the parameters that were used are as follows:

n_1 : 1000, 4000, 7000, 10000, 4000, 70000, 100000, 400000, 700000, 1000000, 4000000, 7000000, 10000000

n_2/n_1 : 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, .3, .7, .9

k_1/n_1 : .001, .002, .005, .01, .015, .02, .05, .1, .25, .5, .75, .9

r : .09, .1, .15, .2, .25, .33, .5

In one series of experiments, we examined the validity of Observation 1 by holding the parameters n_1 and n_2/n_1 constant and steadily increasing the richness constant r and sample proportion k_1/n_1. The results indirectly verified Observation 1 by demonstrating that higher values of the richness constant r lead to better approximations in expression (5.1). These results are consistent with Observation 1 since higher values of r lead to lower values for c in the summation formula in (5.1). Thus $E[X_{n_1k_1}]$ and $E[X_{n_2k_2}]$ depend on the sum of a small number of terms with small j's. In another series of experiments, we investigated the validity of Observation 2 by holding n_1 and the richness constant r constant and steadily increasing the ratio n_2/n_1 and the sample proportion k_1/n_1. As predicted by Observation 2, we found that the accuracy of the approximation increased as the ratio n_2/n_1 increased. Finally both series of experiments confirmed Observation 3, since the relative error of the approximation in (5.1) decreased as the sample proportion k_1/n_1 increased.

APPENDIX B - EVALUATION OF EXPRESSION (5.2)

In this appendix we provide a more detailed discussion of the results of the computer evaluation of expression (5.2). The values of the parameters that we used are as follows:

n : 100, 400, 700, 1000, 4000, 7000, 10000, 40000, 70000, 100000, 400000, 700000, 1000000, 4000000, 700000, 10000000
r_1 : .09, .1, .15, .20, .25, 1/3
r_2 : .1, .15, .20, .25, 1/3, .5 subject to the restriction $r_1 < r_2$.

The simulation produced a mixed bag of results. As expected, the parameter n had no impact on the size of the sample proportion required to make the approximation in Proposition 2 valid. This outcome can be traced to two factors. First the emphasis on relative error rather than absolute error eliminated m_2 from equation (5.3). Second, Proposition 1 demonstrated that

$$\frac{\binom{n_1-j}{k}}{\binom{n_1}{k}} \approx \frac{\binom{n_2-j}{k}}{\binom{n_2}{k}}$$

for all n_1, n_2 of interest. Thus the size of n *should* be unimportant.

An unexpected result was that the parameter r_1 had no bearing on the required size of the sample proportion. This outcome was surprising since smaller values of r_1 should have resulted in larger values for c_1 and thus larger values of (5.3). However, a close analysis of the results indicated that after c_1 crossed a relatively low threshold value, the size of (5.3) relative to $E_{r_2}[X_{n_2k_2}]$ increased only negligibly. Although this threshold value varied, it tended to stay below 100.