The Performance of Tuple Difference Coding for Compressing Databases and Data Warehouses

Wei-Biao Wu and Chinya V. Ravishankar
The University of Michigan, Ann Arbor, MI. 48109
{wbwu,ravi}@umich.edu

Abstract

Databases and data warehouses have grown dramatically in size in recent years, and we are already seeing data warehouses that are tens of terabytes in size. Compression is important in such systems because it reduces space requirements as well as response times for queries, which are typically I/O-bound.

Conventional compression methods tend to compress and decompress objects in their entirety, and are generally unsuitable for databases which require selective access to points in the space of tuples they represent. The database compression method called Tuple-Difference Coding (TDC) has been shown to meet the requirements for database compression, and to achieve high compression ratios in practice. The factors affecting the performance of TDC are known, but their effects are not well understood.

This paper presents a theoretical analysis of the performance of TDC, and verifies these results through simulations. It presents analytical expressions for estimating the compression ratios when database tuples are composed of one or more attributes drawn from various distributions. It also studies the effects of attribute domain reordering on the compression ratio. Our simulations show excellent agreement with theory.

1 Introduction

The importance of database compression has grown in recent years because of dramatic increases in the volumes of data under management. Commercial enterprises have found it profitable to maintain extensive archives of customer transactions, and to seek competitive advantage by discovering patterns across and within transactions by interactive and automated queries. Systems to support interactive queries are generally called data warehouses [Kimball, 1996], while systems for automated querying are typically classified as data mining systems [Fayyad et al., 1996].

Such data archives typically contain records of transactions going back three to five years, and in the case of large-volume retailers and telephone companies, can easily reach 10^{12}–10^{13} bytes or more in size[1]. These data volumes are many orders of magnitude larger than database designers typically encountered only a few years ago. Despite significant differences in usage patterns, there are great structural similarities between conventional databases and such warehouse data archives. Since the differences in organization that do exist between them are not germane to the

[1] It has been reported, for example, that the mass retailer Wal-Mart has a 43-terabyte data warehouse at the time of this writing.

compression approach we discuss, we will use the term "database" to simply mean a collection of fixed-length records, so that the term will refer generically to both conventional databases and to data warehouses.

1.1 Compression techniques

The use of compression in databases can have significant practical advantages since queries of interest to us are typically aggregational, and tend to access substantial portions of the data archive [Ng and Ravishankar, 1997]. Database compression is useful for two related reasons: it reduces storage requirements, and it reduces the data transfer volumes between disk and main memory. Queries in this domain generally request elementary "additive" statistics like sums and variances formed over a large number of records selected from the archive according to user-specified criteria. Typical queries may request a cross-tabulation of profit by retail outlet, month, and product, or seek to discover correlations between sales of different products at various times. Since such queries tend to access large data volumes, disk I/O tends to be the bottleneck to query performance. The use of data compression reduces I/O loads by increasing processor loads, a good tradeoff, given that processor performance is increasing much faster than that of disks.

Data compression is related to the problem of *data modeling* and *source coding* [Williams, 1991, Rissanen and Langdon, 1981]. Conventional data compression techniques tend to use statistical approaches under a serial model, where the source coder (transmitter) accesses the data to be compressed serially, and compresses it according to a statistical model of the data that it constructs and updates on the fly. Similarly, the receiver decompresses the data serially, and maintains its own statistical model, which is updated and kept consistent with that of the transmitter as decompression proceeds.

However, the desirable characteristics for database compression methods differ significantly from those for conventional compression applications, making this approach unsuited for database compression. First, database operations may access, update, or delete individual tuples in the database. The serial model described above is inherently incompatible with this requirement, since an update performed at point t in the data stream would invalidate the statistical model past point t. A second problem is that compression efficiency clearly increases directly with the amount of statistical information available on the distribution of symbols in the source. Thus compression improves when the object to be compressed is large, and any fragmentation of the object reduces compression. This causes difficulty in databases, where there are advantages to applying compression at the level of disk blocks, rather than to the entire archive.

2 Tuple-Difference Coding

Tuple-Difference Coding (TDC) has been proposed as an alternative to conventional compression methods for databases [Ng and Ravishankar, 1997], and has been shown to meet the requirements of database compression well. TDC is quite practical since it can be integrated seamlessly into traditional database structures, and results in substantial compression efficiencies in practice.

We say $\mathcal{R}$ is a relational schema over $D_1, D_2, \ldots, D_r$ with $D_i = \{0, 1, \ldots, |D_i| - 1\}$ being the ith attribute domain[2] if $\mathcal{R} = D_1 \times D_2 \times \ldots \times D_r$. We call $\mathcal{D}$ a database with schema $\mathcal{R}$ if $\mathcal{D} \subseteq \mathcal{R}$.

[2] We will sometimes specify the domain as $\{1, 2, \cdots, |D_i|\}$ instead. However, this change of notation will not

Each record in the database is an r-tuple $\langle d_1, d_2, \ldots, d_r \rangle$, with $d_i \in D_i$.

The idea in TDC is very simple: given a database $\mathcal{D} = \{t_1, t_2, \ldots, t_n\}$ of tuples, a bijective mapping $\varphi : \mathcal{D} \to \mathbb{N}_{\mathcal{R}}$ is constructed, where $\mathbb{N}_{\mathcal{R}} = \{0, 1, \ldots, |\mathcal{R}| - 1\}$. Next, φ used to map each database record t_k to an integer, and the database is sorted on $\varphi(t_k)$ as key. Successive tuples are differenced, and the differences $\varphi(t_{k+1}) - \varphi(t_k)$ are stored instead of the original tuples t_k and t_{k+1} themselves. These differences tend to be significantly smaller than the original tuples, thus achieving compression. In practice, it is convenient to to use a φ that is equivalent to lexicographic sorting. Formally, given a tuple $t_k = \langle d_1, d_2, \ldots, d_r \rangle$, with $d_i \in D_i$,

$$\varphi(t_k) = \sum_{i=1}^{r} d_i \left(\prod_{j=i+1}^{r} |D_j| \right), \tag{1}$$

so that d_i is simply treated as a digit with base $|D_i|$, and t_k becomes a mixed-radix number. This mapping is invertible, so compression is lossless.

2.1 The performance of TDC

The original work describing TDC [Ng and Ravishankar, 1997] provides practical details such as how to handle textual attributes in TDC, and experimental results on query times and other performance parameters in practice. It demonstrates that TDC is superior to other database compression methods currently in use, and provides both better compression as well as faster query response times. The TDC method is also likely to be significant for practice. A patent has been issued for the TDC method, and there appears to be interest in using it in real data warehousing applications. A number of factors affecting compression are listed in [Ng and Ravishankar, 1997], their effects discussed in qualitative terms. It is the purpose of this paper to provide a sounder theoretical basis for the method and its performance characteristics.

2.1.1 A model for performance analysis

We have just outlined the conventional approach of modeling a database as a subset of the product space of the attribute domains in the schema, but two points are worthy of note. First, the distribution of values in different attribute domains D_i and D_j is typically different, though there may be correlations among these values. Second, the actual ordering of the D_i in the product space defining the schema $\mathcal{R}$ is typically immaterial to the database semantics.

It is reasonable, therefore, to view a database as a sample from the joint distribution of the domains D_i. Since it makes no sense for the database to contain duplicate tuples, this sample must be taken from the joint space without replacement. We proceed to form the database by choosing n samples $(X_1, \ldots, X_n)$ from the joint distribution space $\mathcal{F}(D_1, D_2, \ldots, D_r)$. Since this sampling must be performed without replacement, these are not i.i.d. random variables, a fact that causes significant technical difficulties for our analysis in the remainder of the paper. When no compression is performed, we must allocate enough space in the tuple to accommodate *any* value that may need to be stored in the database. Thus, if $N = |D_1| \cdot |D_2| \cdots |D_r|$, each tuple will need to be at least $\log_2 N$ bits in length, and the entire database will be $n \log_2 N$ bits in size.

alter the semantics of compression, since TDC is is based on the spacings between successive samples.

We can apply TDC as follows. By sorting on $\varphi(X_i)$ (see Equation 1), we can construct the sorted database $X_{(1)} <_\varphi \ldots <_\varphi X_{(n)}$, and by differencing them, the corresponding set of tuple spacings $\{X_{(k+1)} - X_{(k)}\}, k = 1, \ldots, n-1$. Since we store these spacings, estimating the size characteristics of the compressed database is equivalent to estimating the sum of the logarithms of these spacings. For convenience, we use natural logarithms rather than base-2 logarithms, and form the statistic $\Lambda_{\mathcal{F}} = \sum_{k=1}^{n-1} \ln(X_{(k+1)} - X_{(k)})$. One of our chief challenges in the remainder of the paper will be to estimate $\Lambda_{\mathcal{F}}$ for different attribute spaces $\mathcal{F}$. To estimate the compression efficiency, this value must be compared with the number $n \ln N$, where N is defined as above.

While attribute domain ordering is irrelevant to semantics, some ordering must be chosen for storing the tuples on disk. Since lexicographic sorting is used to order the tuples, we also consider the problem of optimal ordering of the attribute domains to minimize $\Lambda_{\mathcal{F}}$. Also, when correlations exist between the attribute domains in $\mathcal{R}$, the entropy of the database is lowered, so that compression methods tend to perform better in such cases. To obtain lower bounds on the performance of TDC, we therefore assume that attribute domains are uncorrelated. This model will form the basis for further analysis.

3 An equivalent sampling scheme and related topics

In Section 2.1.1, the n samples $X_1, \ldots, X_n$ drawn from the set $\mathcal{F}$ were assumed to be mutually different since it makes no sense for a database to store a tuple more than once. This scheme of sampling without replacement may be modeled as follows. Suppose $|\mathcal{F}| = N$ and the first sample X_1 has distribution $\Pr[X_1 = x_1] = p(x_1)$ for $x_1 \in \mathcal{F}$. Given the first outcome $X_1 = x_1$, X_2 can only be drawn from the set $\mathcal{F} - \{x_1\}$ with mass function $p(x_2)/[1 - p(x_1)]$. Inductively, given $X_1 = x_1, X_2 = x_2, \ldots, X_{k-1} = x_{k-1}$, we have the conditional probability

$$\Pr[X_k = x_k | X_1 = x_1, X_2 = x_2, \ldots, X_{k-1} = x_{k-1}] = \frac{p(x_k)}{1 - \sum_{i=1}^{k-1} p(x_i)}$$

for $x_k \in \mathcal{F} - \{x_1, \ldots, x_{k-1}\}, k = 2, \ldots, n$. Hence, the joint distribution of $X_1, \ldots, X_n$ will have the form

$$\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}, X_n = x_n] = \prod_{k=1}^{n} \frac{p(x_k)}{1 - \sum_{i=1}^{k-1} p(x_i)},$$

which is complicated enough to make a direct analysis of the corresponding order statistics $X_{(1)} < \ldots < X_{(n)}$ intractable. The random variables $X_1, \ldots, X_n$ are related in an extremely complicated fashion due to the dependence of X_k on all the previous outcomes $X_1, \ldots, X_{k-1}$. We must resort to some suitable transformations and approximations here.

We therefore introduce the following alternative scheme based on sampling with replacement, and show that it is equivalent to the one just described. We also use this alternative scheme to build databases when testing the validity of our theoretical analysis through simulations. Let $\{X_1, Z, Z_i, i \geq 1\}$ be i.i.d. random variables. Define stopping times

$$J_1 \equiv 1, J_{k+1} = \inf\{i > J_k : Z_i \notin \{Z_{J_1}, \ldots, Z_{J_k}\}\}, k \geq 1.$$

For any n, J_n can be shown to be proper in the sense that $\Pr[J_n < \infty] = 1$, which follows immediately from the following Equation 3. Based on $\{J_n, n \geq 1\}$, we can naturally obtain n mutually different random variables $Z_{J_1}, \ldots, Z_{J_n}$. We show that

$$(X_1, \ldots, X_n) \stackrel{\mathcal{D}}{=} (Z_{J_1}, \ldots, Z_{J_n}) \tag{2}$$

Set conditional probability $\hat{\Pr}[\cdot] = \Pr[\cdot | Z_{J_1} = x_1, \ldots, Z_{J_k} = x_k]$. Using the Markov property, we have

$$\begin{aligned}
&\hat{\Pr}[Z_{J_{k+1}} = x_{k+1}] \\
&= \sum_{l=1}^{\infty} \hat{\Pr}[Z_{J_k+l} = x_{k+1}, Z_{J_k+j} \in \{x_1, \ldots, x_k\}, \text{ for } j = 1, \ldots, l-1] \\
&= \sum_{l=1}^{\infty} \Pr[Z = x_{k+1}] \Pr[Z \in \{x_1, \ldots, x_k\}]^{l-1} \\
&= \frac{p(x_{k+1})}{1 - \sum_{i=1}^{k} p(x_i)}
\end{aligned}$$

proving Equation 2.

Let $T_k = J_{k+1} - J_k, k \geq 1$. It is easy to see that given the outcomes $(x_1, \ldots, x_n)$, the values T_k are geometrically distributed with mean $(1 - \sum_{i=1}^{k} p(x_i))^{-1}$. On average, we need $E[J_n]$ i.i.d. random variables to obtain n different values. This connection between the two sampling schemes enables us to work with an i.i.d. sampling scheme instead of the more complex scheme of sampling without replacement.

To illustrate the use of this idea, consider Theorem 4 below, which analyzes the performance of TDC on databases which may be modeled by sampling a Zipf variable *without* replacement. The theorem, in fact, gives approximations of $E\Lambda_Z$ based on a sampling scheme *with* replacement; that is, the proof of the theorem is developed in terms of n i.i.d. Zipf(N) random variables. Therefore, in applying Theorem 4 to obtain a reasonable estimate for $E\Lambda_Z$ when replacement is *not* allowed, we would use $E[J_n]$ in place of n.

However, it is still extremely difficult to use

$$E[J_n] = 1 + \sum_{k=1}^{n-1} E[T_k] = 1 + \sum_{k=1}^{n-1} E\left[1 - \sum_{i=1}^{k} p(X_i)\right]^{-1} \tag{3}$$

directly, given the very complicated nature of the joint distribution of $(X_1, \ldots, X_n)$. We finesse this problem by considering the converse issue:

Question. *Given n' i.i.d. random variables $Z_1, \ldots, Z_{n'}$, what is the number of different elements in this sample? Equivalently, what is the cardinality of $\{Z_1, \ldots, Z_{n'}\}$?*

In this set up, n' and $|\{Z_1, \ldots, Z_{n'}\}|$ assume the roles of $E[J_n]$ and n respectively in the original problem. This converse can be interpreted as random allocation problem, which is extensively studied in the literature, especially for weak convergence in terms of the Central Limit Theorem and Poisson approximations (c.f. [Kolchin et al., 1978]). This converse question has also been

studied in [Csörgő and Wu, 1998] for the case where Z has uniform distribution, using large deviation techniques.

Suppose we have N cells labeled by $1, \ldots, N$, and we view the random variables $Z_1, \ldots, Z_{n'}$ as n' balls with ball j being allocated to cell Z_j. Define random variables $Y_i = \sum_{j=1}^{n'} 1(Z_j = i), i = 1, \ldots, N$, representing the number of balls in the ith cell, where $1(A)$ is the indicator function. Hence the number of occupied cells $|\{Z_1, \ldots, Z_{n'}\}| = \sum_{j=1}^{n'} 1(Y_i > 0)$. Now $(Y_1, \ldots, Y_N)$ follows the multinomial distribution $Multi(n'; p(1), \ldots, p(N))$. Let V_i have Poisson distribution with mean $n'p(i)$ and suppose that $\{V_i, 1 \leq i \leq N\}$ are independent. Then we have the following Poisson representation of the multinomial distribution

$$(Y_1, \ldots, Y_N) \stackrel{\mathcal{D}}{=} (V_1, \ldots, V_N | V_1 + \ldots + V_N = n').$$

Denote $M = \sum_{i=1}^{N} \Pr[V_i > 0] = \sum_{i=1}^{N} [1 - \exp(-n'p(i))]$. Then by Markov's inequality, we have for any $\varepsilon > 0$,

$$\begin{aligned}
&\Pr\left[\left|\sum_{i=1}^{N} 1(Y_i > 0) - M\right| \geq n'\varepsilon\right] \\
&= \Pr\left[\left|\sum_{i=1}^{N} 1(V_i > 0) - M\right| \geq n'\varepsilon \middle| V_1 + \ldots + V_N = n'\right] \\
&\leq \frac{\Pr\left[|\sum_{i=1}^{N} 1(V_i > 0) - M| \geq n'\varepsilon\right]}{\Pr[V_1 + \ldots + V_N = n']} \\
&\leq \frac{e^{n'} n'^{n'}}{n'!} \frac{1}{(n'\varepsilon)^2} E\left[\sum_{i=1}^{N} 1(V_i > 0) - M\right]^2 \\
&= \frac{O(1)}{(n')^{3/2}} \left(\sum_{i:n'p(i)>1} + \sum_{i:n'p(i)\leq 1}\right) e^{-n'p(i)}(1 - e^{-n'p(i)}) \\
&= \frac{O(1)}{(n')^{3/2}} (e^{-1}n' + n') \\
&= \frac{O(1)}{(n')^{1/2}},
\end{aligned}$$

yielding

$$\frac{1}{n'}\left[\sum_{i=1}^{N} 1(Y_i > 0) - M\right] \xrightarrow{\mathcal{P}} 0.$$

Therefore, it is reasonable to take M as the expected number of occupied cells, and the expected number of i.i.d. copies needed, n', can be approximated via the equation

$$n = M = \sum_{i=1}^{N} [1 - \exp(-n'p(i))] \tag{4}$$

This scheme causes a complication if the definition of Λ given in Section 2.1.1 is used directly. Let $Z_{(1)} \leq \ldots \leq Z_{(n')}$ be the order statistics of $Z_1, \ldots, Z_{n'}$. Since we cannot guarantee that the Z_i

are mutually different, some of these spacings may be zero. The definition of Λ in in Section 2.1.1 is unusable since it involves the logarithms for these spacings. Instead, we should use one of the forms $\Lambda_Z = \sum_{k=1}^{n'-1} \ln \max(Z_{(k+1)} - Z_{(k)}, 1)$ or $\Lambda_Z = \sum_{k=1}^{n'-1} \ln(Z_{(k+1)} - Z_{(k)} + 1)$. In this paper, we suggest the second form, which appears conservative, and is mathematically convenient. In addition, it captures an aspect of the real world: whenever $Z_{(k+1)} - Z_{(k)} = 1$ in practice, we need one bit to store the difference. However, the corresponding term in the first form goes to zero and makes no contribution to the sum. Numerical simulation indicates that the difference between the two forms is negligible.

4 Limit theorems for the single-field case

We begin our analysis of the TDC technique with the simplest case. We assume that the database consists of n tuples, each tuple comprising a single attribute field A. This is a reasonable starting point for two reasons. First, in some cases, we are able to reduce the general case of r attribute domains to the case of a single attribute domain. Second, we use the single-attribute results to construct an analysis for the multiple-domain case.

4.1 Single attribute, uniform distribution

We first consider the case when the attribute values are drawn uniformly from a single attribute domain of size N. The uniform distribution is interesting for several reasons. First, many attributes domains that appear in practice are uniform. Second, the uniform distribution is known to yield the largest value for the sum of sample spacings of all distributions defined over a given range [Shao and Marjorie, 1995]. In this sense, the behavior for the uniform distribution form a lower bound for the compression efficiency of TDC. Also, the uniform distribution is a "least informative" distribution over a given range, and is useful as a model when little is known about a distribution. Finally, as we show in Section 5, a set of k uniformly distributed attribute domains can be modeled as a single uniformly distributed domain.

We proceed to form the database by choosing n integers $(X_1, \ldots, X_n)$ from $\{0, 1, \ldots, N-1\}$, so that each such n-tuple representing the database has the same probability $1/\binom{N}{n}$ of being selected. By sorting this database, we can construct the order statistics $X_{(1)} < \ldots < X_{(n)}$, and the corresponding set of spacings $\{X_{(k+1)} - X_{(k)}\}, k = 1, \ldots, n-1$. As in Section 2.1.1, we form the statistic $\Lambda_U = \sum_{k=1}^{n-1} \ln(X_{(k+1)} - X_{(k)})$ to estimate the size of the compressed database.

4.1.1 Prior work

We need to work with a discrete uniform distribution, and so we call the problem of estimating Λ_U the *discrete spacing* problem. To the best of our knowledge, prior work in this area has dealt exclusively with continuous distributions. See, for example, [Darling, 1953, Blumenthal, 1968, Pyke, 1965, Shao and Marjorie, 1995]. Pyke [Pyke, 1965] reviews the literature in this area. Darling [Darling, 1953] uses characteristic function techniques to obtain the following limit theorem for the continuous spacings of independent random variables uniform on $(0, 1)$.

Theorem 1. *Let $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ be the order statistics of i.i.d. uniform(0,1) random variables $U_1, \ldots, U_n$. Then,*

$$\frac{\sum_{i=1}^{n-1} \ln\left(U_{(i+1)} - U_{(i)}\right) + (n+1)(\ln n + \gamma)}{\sqrt{n(\pi^2/6 - 1)}} \xrightarrow{\mathcal{D}} N(0,1).$$

However, we can not directly extend these results to the discrete case. In particular, although sampling with and without replacement are equivalent for the continuous case, they are not so for discrete distributions. Sampling with replacement causes a singularity in the logarithmic term since $X_{(k+1)} - X_{(k)} = 0$ with non-zero probability.

There are two ways to overcome this difficulty, and we explore both possibilities in our work. The first way is to require sampling without replacement, as the nature of our problem dictates, and as we assumed in Section 3. The second approach is to change $\ln(X_{(k+1)} - X_{(k)})$ to $\ln(X_{(k+1)} - X_{(k)} + 1)$. We develop central-limit theorems (theorems 2 and 3 below) to address each of these choices, respectively.

At first sight, it seems feasible to apply Darling's Theorem to our situation by simply substituting the discrete random variables $X_i = \lfloor NU_i \rfloor, 1 \leq i \leq n$ for the continuous random variables $NU_i, 1 \leq i \leq n$. However, this straightforward substitution becomes problematic since the errors will be large if we replace the spacing term $\ln\left(X_{(k+1)} - X_{(k)}\right)$ in Theorem 2, or $\ln\left(X_{(k+1)} - X_{(k)} + 1\right)$ in Theorem 3 by the continuous version $\ln\left(NU_{(k+1)} - NU_{(k)}\right)$ unless $NU_{(k+1)} - NU_{(k)}, 1 \leq k \leq n-1$ are stochastically large. The reason for this difficulty is obvious: the error $\ln(t + dt) - \ln t \approx dt/t$ will be small for large t. We show that this difficulty may be circumvented when N is large enough, and specifically, when $n^2 = o(N)$. A significant aspect of our approach to the proof of Theorem 2 is that we estimate the possible errors caused by the continuous approximation we use, and show them to be negligible. We then proceed to obtain the limiting distribution. Although Darling's Theorem is not helpful in the the proof of Theorem 2, it does provide an incidental benefit. The asymptotic variance we obtain for our limit theorems is hard to estimate analytically, but we can infer it by comparison with Theorem 1.

4.1.2 A Central-Limit Theorem for discrete uniform spacings

Since Λ_U is the sum of a large number of random variables, we would expect the statistic to be distributed normally. However, in order to characterize the performance of TDC, we are especially interested in the mean and variance of this distribution.

In Theorem 2 below, we prove a version of the Central Limit Theorem for this case, and show that the expected value of Λ_U is approximately $n[-\gamma + \ln(N/n)]$, where $\gamma = 0.57721\ldots$ is Euler's constant. In contrast, the number of bits to store $X_1, \ldots, X_n$ without compression is $n \ln N$.

In showing Theorem 2, we first approximate the sample $(X_1, \ldots, X_n)$ by the i.i.d. random variables $X'_1, \ldots, X'_n$ distributed as $\lfloor NU \rfloor$, U is uniformly distributed over $(0,1)$. Obviously, because we are sampling without replacement, $(X_1, \ldots, X_n)$ are not independent, but if $n = o(N^{1/2})$, then we expect then to be asymptotically independent, since the probability of $X_i = X_j$ for some $1 \leq i < j \leq n$ is very small. In the proof, we deal with the order statistics $X'_{(1)} \leq \ldots \leq X'_{(n)}$, or equivalently, $\lfloor NU_{(1)} \rfloor \leq \lfloor NU_{(2)} \rfloor \leq \ldots \leq \lfloor NU_{(n)} \rfloor$ using the representation

$$(U_{(1)}, \ldots, U_{(n)}) \stackrel{\mathcal{D}}{=} \left(\frac{S_1}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}}\right), \tag{5}$$

where $Y_1, Y_2, \ldots$ are i.i.d. exp(1) random variables, and $S_j = \sum_{i=1}^{j} Y_j$.

We will use this representation form throughout the paper. Therefore, Λ_U can be approximated by $\sum_{k=1}^{n-1} \ln(NY_{k+1}/S_{n+1})$, which can be analyzed using the Strong Law of Large Numbers. In the process, however, we encounter sets with small probabilities, with which we must deal with care. We first prove the following lemma.

Lemma 1. Let $\{Y, Y_i, i \geq 1\}$ be i.i.d. exp(1) random variables. Suppose that $n^2 = o(N_n)$. Then[3]

$$\frac{n}{N_n \ln n} \sum_{i=1}^{n} \frac{1}{Y_i} \xrightarrow{\mathcal{P}} 0$$

Proof. Set $b_n = n^{1/2} N_n^{1/4}$. It suffices to show

$$\frac{n}{N_n} \sum_{i=1}^{n} \left[\frac{1}{Y_i} - E\frac{1}{Y} 1(Y > b_n^{-1}) \right] \xrightarrow{\mathcal{P}} 0 \tag{6}$$

and

$$\frac{n^2}{N_n \ln n} E\frac{1}{Y} 1(Y > b_n^{-1}) \longrightarrow 0. \tag{7}$$

Via the Markov inequality, (6) follows from

$$\begin{aligned} &\Pr\left[\frac{n}{N_n} \sum_{i=1}^{n} \frac{1}{Y_i} \neq \frac{n}{N_n} \sum_{i=1}^{n} \frac{1}{Y_i} 1(Y_i > b_n^{-1}) \right] \\ &\leq \Pr\left[\cup_{i=1}^{n} \frac{1}{Y_i} \neq \frac{1}{Y_i} 1(Y_i > b_n^{-1}) \right] \leq n \Pr[Y \geq b_n^{-1}] = n[1 - \exp(b_n^{-1})] \to 0 \end{aligned}$$

and

$$\begin{aligned} &\frac{n^2}{N_n^2} E\left[\sum_{i=1}^{n} Y_i^{-1} 1(Y_i > b_n^{-1}) - EY^{-1} 1(Y > b_n^{-1}) \right]^2 \\ &\leq \frac{n^3}{N_n^2} EY^{-2} 1(Y > b_n^{-1}) \leq \frac{n^3}{N_n^2} \int_{b_n^{-1}}^{\infty} y^{-2} e^{-y} dy \leq \frac{n^3 b_n}{N_n^2} \to 0. \end{aligned}$$

For (7), we will have

$$\frac{n^2}{N_n \ln n} E\frac{1}{Y}(Y > b_n^{-1}) = \frac{n^2}{N_n \ln n} \Big(\int_1^{\infty} + \int_{b_n^{-1}}^{1} \Big) \frac{e^{-y}}{y} dy \leq \frac{n^2 (\ln b_n + 1)}{N_n \ln n} \to 0$$

if we can show that $(N_n \ln n^2)^{-1} n^2 \ln N_n \to 0$. Define the sets $\mathcal{I} = \{n \in \mathbb{N} : n \geq 2, N_n \geq n^4\}$ and $\mathcal{J} = \{n \in \mathbb{N} : n \geq 2, N_n < n^4\}$. Without loss of generality, we assume that $|\mathcal{I}| = \infty, |\mathcal{J}| = \infty$. Hence

$$\limsup_{n \in \mathcal{I}, n \to \infty} \frac{n^2 \ln N_n}{N_n \ln n^2} \leq \limsup_{n \in \mathcal{I}, n \to \infty} \frac{n^2}{N_n^{1/2} \ln n^2} = 0$$

[3] This is a Feller-type theorem (c.f. Theorem 5.2.4 of [Chow and Teicher, 1988]). That result can not be applied directly, but our proof proceeds along the same lines. This lemma is interesting because convergence still holds although the mean $EY^{-1} = \infty$.

and

$$\limsup_{n\in\mathcal{J},n\to\infty} \frac{n^2 \ln N_n}{N_n \ln n^2} \leq 2 \limsup_{n\in\mathcal{J},n\to\infty} \frac{n^2}{N_n} = 0,$$

which completes the proof. □

We now present a theorem dealing with the performance of TDC when the values in the database are uniformly distributed, and when the database size n and the attribute domain size N_n obey $n^2 = o(N_n)$. In this case, we are able to obtain accurate results without recourse to the equivalent sampling scheme described in Section 3. In Section 4.1.3, we show how to extend these results to the case when $n^2 = o(N_n)$ fails to hold.

Theorem 2. *Let $(X_1, \ldots, X_n)$ be n numbers sampled from the set $\{0, 1, \ldots, N_n - 1\}$ equiprobably, and without replacement. Let $X_{(1)} < \ldots < X_{(n)}$ be the order statistics of $(X_1, \ldots, X_n)$, and define the random variable $\Lambda_U = \sum_{k=1}^{n-1} \ln(X_{(k+1)} - X_{(k)})$. Then, if $n^2 = o(N_n)$,*

$$\frac{\Lambda_U - \mu_U}{\sigma_U/\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1),$$

where $\mu_U = (n-1)[\ln N_n - \ln(n+1) - \gamma]$, $\sigma_U = \alpha\sqrt{n(n-1)}$, and γ, α are defined in terms of a standard exponential random variable Y as $\gamma = -E(\ln Y) = 0.57721\ldots$, or Euler's constant, and $\alpha^2 = Var(\ln Y - Y) = \pi^2/6 - 1 = 0.644934\ldots$

Proof. We write $N = N_n, \sigma = \sigma_U, \mu = \mu_U$ for simplicity. Let $X_1', \ldots, X_n'$ be i.i.d. random variables with common distribution $\Pr[X_1' = k] = 1/N$ for $k = 0, 1, \ldots, N-1$. First, we claim the following distributional equality, which will transform the dependent random variables $(X_1, \ldots, X_n)$ to i.i.d. random variables $(X_1', \ldots, X_n')$:

$$(X_1, \ldots, X_n) \stackrel{\mathcal{D}}{=} (X_1', \ldots, X_n' | X_1', \ldots, X_n' \text{ are different}). \tag{8}$$

For $x_1, \ldots, x_n \in \{0, 1, \ldots, N-1\}$, if $x_i = x_j$ for some $i \neq j$, then

$$\begin{aligned} &\Pr[X_1 = x_1, \ldots, X_n = x_n] = 0 \\ &= \Pr[X_1' = x_1, \ldots, X_n' = x_n | X_1', \ldots, X_n' \text{ are different}]. \end{aligned}$$

If $x_1, x_2, \ldots, x_n$ are mutually different, then

$$\begin{aligned} &\Pr[X_1' = x_1, \ldots, X_n' = x_n | X_1', \ldots, X_n' \text{ are different}] \\ &= \Pr[X_1' = x_1, \ldots, X_n' = x_n] / \Pr[X_1', \ldots, X_n' \text{ are different}] \\ &= \left(\frac{1}{N}\right)^n \left\{ \prod_{j=1}^{n-1} \left(1 - \frac{j}{N}\right) \right\}^{-1} = \binom{N}{n}^{-1} \\ &= \Pr[X_1 = x_1, \ldots, X_n = x_n]. \end{aligned}$$

Hence for fixed $\lambda \in \mathbb{R}$,

$$\begin{aligned}
&\Pr\Big[\frac{1}{(n-1)^{1/2}\alpha}(\Lambda_U-\mu)<\lambda\Big]\\
&=\Pr\Big[\sum_{k=1}^{n-1}\ln^+(X'_{(k+1)}-X'_{(k)})-\mu<\lambda(n-1)^{1/2}\alpha\Big|X'_1,\ldots,X'_n \text{ are different}\Big]\\
&=\Pr[B_n|A_n] \text{ (say)},
\end{aligned}$$

where $X'_{(1)}\le\ldots\le X'_{(n)}$ is the order statistics of $(X'_1,\ldots,X'_n)$, and $\ln^+ x=\ln(\max(1,x))$. Observing that $\Pr[A_n]=\prod_{j=1}^{n-1}(1-j/N)=1+O(n^2/N)=1+o(1)$, and that

$$\frac{\Pr[B_n]}{\Pr[A_n]}\ge\frac{\Pr[A_nB_n]}{\Pr[A_n]}=\Pr[B_n|A_n]\ge\frac{\Pr[B_n]}{\Pr[A_n]}+1-\frac{1}{\Pr[A_n]},$$

we find that $\Pr[B_n|A_n]$ is close to $\Pr[B_n]$. Hence it suffices to show $\lim_{n\to\infty}\Pr[B_n]=\Phi(\lambda)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\lambda}e^{-t^2/2}dt$ as $n\to\infty$. Since $X'_k\overset{\mathcal{D}}{=}\lfloor NU_k\rfloor$, where $U_1,\ldots,U_n$ are i.i.d. random variables uniformly distributed over $(0,1)$, we know

$$(X'_1,\ldots,X'_n)\overset{\mathcal{D}}{=}(\lfloor NU_1\rfloor,\ldots,\lfloor NU_n\rfloor),$$

which yields

$$(X'_{(1)},\ldots,X'_{(n)})\overset{\mathcal{D}}{=}(\lfloor NU_{(1)}\rfloor,\ldots,\lfloor NU_{(n)}\rfloor). \tag{9}$$

Let $Y,Y_1,\ldots$ be i.i.d. $exp(1)$ random variables, and let $S_m=\sum_{i=1}^m Y_i$. We now make use of Equation 5, and let event

$$\hat{B}_n=\Big\{\sum_{k=1}^{n-1}\ln^+(\lfloor NS_{k+1}/S_{n+1}\rfloor-\lfloor NS_k/S_{n+1}\rfloor)-\mu<\lambda(n-1)^{1/2}\alpha\Big\}.$$

From Equation 9, $\Pr[B_n]=\Pr[\hat{B}_n]$. Now we can estimate $\Pr[B_n]$ by approximating the integer parts by the values themselves. Roughly speaking, the summand $\lfloor NS_{k+1}/S_{n+1}\rfloor-\lfloor NS_k/S_{n+1}\rfloor$ in the logarithmic terms will be close to NY_{k+1}/n, which is stochastically large since we have $N/n\to\infty$ and $S_{n+1}/n\to EY=1$, by the usual Strong Law of Large Numbers.

To be more precise, we introduce the events C_n,D_n, as follows:

$$C_n=\Big\{\frac{NY_2}{S_{n+1}}>2,\ldots,\frac{NY_n}{S_{n+1}}>2\Big\},D_n=\Big\{\Big|\frac{S_{n+1}-(n+1)}{\sqrt{n\ln n}}\Big|>1\Big\}.$$

Under event C_n, the summands in the logarithmic terms are larger than 1, so it makes sense to take logarithms. Event D_n^c, the complement of event D_n, leads us to the approximation $S_{n+1}\approx n$. If event $C_nD_n^c$ occurs, then we can show by a straightforward approach that $\ln^+(\lfloor NS_{k+1}/S_{n+1}\rfloor-\lfloor NS_k/S_{n+1}\rfloor)$ can be approximated by $\ln(NY_{k+1}/n)$.

Hence, we really need to show that $\Pr[C_nD_n^c]=1+o(1)$, or, that $\Pr[C_n^c]+\Pr[D_n]=o(1)$. To prove this, first $\Pr[D_n]\le(n\ln n)^{-1}E[S_{n+1}-(n+1)]^2=o(1)$. Next, for large n, we have $\Pr[C_n]\ge\Pr[C_nD_n^c]\ge\Pr[NY_2>3n,\ldots,NY_n>3n,D_n^c]\ge(e^{-3n/N})^{n-1}-\Pr[D_n]=1+o(1)$.

Let $\{x\}$ denote the fractional part of x (i.e., $\{x\} = x - \lfloor x \rfloor$), and let $\varepsilon_{nk} = \{NS_k/S_{n+1}\} - \{NS_{k+1}/S_{n+1}\} \in (-1, 1)$. If $\omega \in C_n D_n^c$, and n is sufficiently large, we can obtain the following estimates: $|S_{n+1}/(n+1) - 1| < (\ln n/n)^{1/2}$, $|S_{n+1}/(n+1) - 1 - \ln(S_{n+1}/(n+1))| \le (S_{n+1}/(n+1) - 1)^2 < \ln n/n$. From these estimates, we now have

$$\begin{aligned}
&\Big| \sum_{k=1}^{n-1} \ln^+ \Big(\Big\lfloor \frac{NS_{k+1}}{S_{n+1}} \Big\rfloor - \Big\lfloor \frac{NS_k}{S_{n+1}} \Big\rfloor \Big) - \mu - \Big(\sum_{k=1}^{n-1} (\ln Y_{k+1} + \gamma) - S_{n+1} + n + 1 \Big) \Big| \\
&= \Big| \sum_{k=1}^{n-1} \ln \Big(1 + \frac{S_{n+1}\varepsilon_{nk}}{NY_{k+1}} \Big) + 2\Big(\frac{S_{n+1}}{n+1} - 1 \Big) + (n-1)\Big(\frac{S_{n+1}}{n+1} - 1 - \ln \frac{S_{n+1}}{n+1} \Big) \Big| \\
&\le \sum_{k=1}^{n-1} \frac{2S_{n+1}}{NY_{k+1}} + 2\Big(\frac{\ln n}{n} \Big)^{1/2} + \ln n \\
&\le \frac{3n}{N} \sum_{k=1}^{n-1} \frac{1}{Y_{k+1}} + 2 \ln n
\end{aligned}$$

By Lemma 1, notice that $\Pr[C_n D_n^c] = 1 + o(1)$, we obtain

$$\lim_{n\to\infty} \frac{1}{n^{1/2}} \Big| \sum_{k=1}^{n-1} \ln^+ \Big(\Big\lfloor \frac{NS_{k+1}}{S_{n+1}} \Big\rfloor - \Big\lfloor \frac{NS_k}{S_{n+1}} \Big\rfloor \Big) - \mu - \Big(\sum_{k=1}^{n-1} (\ln Y_{k+1} + \gamma) - S_{n+1} + n + 1 \Big) \Big| \stackrel{\mathcal{P}}{=} 0,$$

which leads to Theorem 2 via Slutsky's Theorem and the classical central limit theorem $[(n-1)^{1/2}\alpha]^{-1} \sum_{k=1}^{n-1} (\ln Y_{k+1} + \gamma - Y_{k+1} + 1) \stackrel{\mathcal{D}}{\longrightarrow} N(0,1)$. The exact value of the asymptotic variance α^2 is presented in Corollary 1 below. □

Figure 1 compares the estimates of Λ_U from Theorem 2 with the results of experiments on databases of different sizes containing integers drawn uniformly without replacement from $\{1, 2, \cdots, 2^{31} - 1\}$. The two curves show the relative error observed when the database size n is used directly in Theorem 2, and the lower error observed when the value n' from Equation 4 is used instead. Theory and experiment agree to within a fraction of one percent even for databases as large as $2 \cdot 10^6$, showing that the theorem is robust, since $\sqrt{N} \approx 46,000$ in this case.

The major idea in the proof of Theorem 2 was to first show the asymptotic equivalence of the sample $(X_1, \ldots, X_n)$ without replacement and n i.i.d. uniform(N_n) random variables under the constraint $n^2 = o(N_n)$ and hence reduce to the classical central limit theorem based on the i.i.d. case. After some minor modifications, the proof in the second part also implies the following theorem for the n i.i.d. uniform(N_n) random variables.

Theorem 3. *Assume that $n^2 = o(N_n)$. Let $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ be the order statistics of i.i.d. uniform(N_n) random variables $X_1, \ldots, X_n$. Define $\Lambda_U = \sum_{k=1}^{n-1} \ln(X_{(k+1)} - X_{(k)} + 1)$. Then,*

$$\frac{\Lambda_U - \mu_U}{\sigma_U/\sqrt{n}} \stackrel{\mathcal{D}}{\longrightarrow} N(0,1),$$

where μ_U, σ_U are the same as that in Theorem 2.

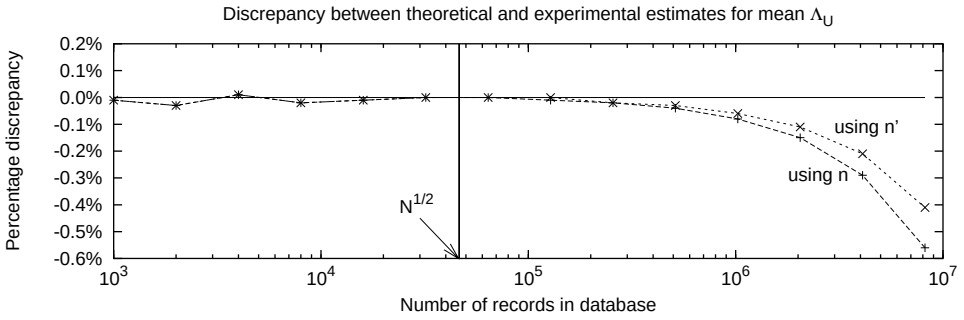

Figure 1: Uniform distribution: agreement between Theorem 2 and experiment

Remark 1. Theorems 2 and 3 can be used to construct confidence intervals based on the limiting distributions.

Obtaining the exact form of the variance term is an interesting exercise. It is somewhat challenging to obtain $\alpha^2 = Var(\ln Y - Y)$ directly, but we note that Theorems 1 and 3 jointly lead to the following interesting observation.

Corollary 1. If Y is $\exp(1)$ distributed, then $\alpha^2 = Var(\ln Y - Y) = \pi^2/6 - 1 = 0.644934\ldots$.

4.1.3 The case of large databases

When $n^2 = o(N_n)$ is not satisfied, we must fall back on the equivalent sampling scheme described in Section 3. Since for the uniform distribution, $p(x) = N^{-1}$ for all $x \in \mathcal{F}$, we have $E[J_n] = 1 + \sum_{k=1}^{n-1} (1 - k/n)^{-1}$ by Equation 3. Under the assumption $n < N/2$, we claim that

$$\left| E[J_n] - N \ln\left(1 - \frac{n}{N}\right) \right| < \frac{n+2}{N} \tag{10}$$

Define function $g(t) = (1 - t/N)^{-1}$. For integer $k \in [1, n-1]$, if $t \in [k - 1/2, k + 1/2]$, the Taylor expansion yields

$$g(t) = g(k) + (t-k)g'(k) + \frac{(t-k)^2}{2} g''(\xi)$$

for some $\xi \in [k - 1/2, k + 1/2]$. If $t \in (0, N/2)$, then

$$|g''(t)| = \left| \frac{2}{N^2} \left(1 - \frac{t}{N}\right)^{-3} \right| \leq \frac{16}{N^2}.$$

Therefore

$$\begin{aligned} \left| \int_{1/2}^{n-1/2} g(t)dt - \sum_{k=1}^{n-1} g(k) \right| &\leq \sum_{k=1}^{n-1} \left| \int_{k-1/2}^{k+1/2} [g(t) - g(k)]dt \right| \\ &\leq \sum_{k=1}^{n-1} \frac{16}{N^2} \int_{k-1/2}^{k+1/2} \frac{(t-k)^2}{2} dt \leq \frac{2n}{3N^2} < \frac{1}{3N}. \end{aligned}$$

Next,

$$\begin{aligned}\left|E[J_n] - N\ln\left(1-\frac{n}{N}\right)\right| &< \left|1 + \int_{1/2}^{n-1/2} g(t)dt - N\ln\left(1-\frac{n}{N}\right)\right| + \frac{1}{3N} \\ &< \left|\frac{1}{2} + N\ln\left(1-\frac{1}{2N}\right)\right| + \left|\frac{1}{2} - N\ln\left(1-\frac{2n-1}{2N}\right) + N\ln\left(1-\frac{n}{N}\right)\right| + \frac{1}{3N} \\ &< \frac{n+2}{N}.\end{aligned}$$

Inequality 10 means that, on average, we have to draw $n' = -N\ln(1-n/N) \approx E[J_n]$ i.i.d. samples uniformly from $\{1, 2, \cdots, N\}$ to get n distinct values. Therefore, in applying Theorem 3 for a sample of size n obtained without replacement, we must use the adjusted sample size n' in place of n to get a reasonable result. We also observe that under the assumption $n^2 = o(N_n)$ we get $n' = -N\ln(1-n/N) = n + O(n^2/N) \approx n$, supporting our treatment in the proof of Theorem 2.

It is instructive to examine the applicability of Equation 4 here. From this equation, the adjusted sample size n' satisfies $n = \sum_{i=1}^{N}[1 - \exp(-n'/N)] = N[1 - \exp(-n'/N)]$, so that we have $n' = -N\ln(1-n/N)$, which is in excellent agreement with Inequality 10.

4.2 Single attribute, Zipf distribution

We say that random variable X has the Zipf distribution with parameter N if $\Pr[X = k] = k^{-1}/H_N$, $k = 1, 2, \ldots, N$, where $H_k = \sum_{i=1}^{k} i^{-1}$. The Zipf distribution is of practical interest because many attribute domains appear to follow this distribution in practice. It was first studied in the context of the distributions of word frequencies in documents, but it was soon found to arise in a wide range of other applications. It is now known [Li, 1992] that the Zipf distribution arises naturally in many contexts. For example, when strings are formed from letters chosen randomly from an alphabet with fixed probabilities, the distribution of words is Zipf.

The Zipf distribution can pose considerable analytical difficulties, particularly in the context of the problem we are addressing. When we take a sample $X_1, \ldots, X_n$ without replacement from the the set $S = \{1, \ldots, N\}$, whose elements are distributed as Zipf(N), the joint distribution of the X_i is very complicated. We find the sampling equivalence results of Section 3 especially useful for this case. Theorem 4 below and Remark 2 give approximations of Λ_Z based on a sampling scheme *with* replacement; i.e., the sample analyzed is of n i.i.d. Zipf(N) random variables. Since repetition is not allowed, we may apply the arguments in Section 3, and use $E[J_n]$ in place of n in Theorem 4 to obtain reasonable estimates for Λ_Z.

A problem is that $E[J_n]$ can be calculated directly from Equation 3 only for very special cases; the only really tractable case may well be the uniform distribution. We must therefore solve for n' from Equation 4, and proceed as follows. Define $M = \sum_{i=1}^{N}[1 - \exp(\lambda/i)]$, $\lambda = n'/H_N$. By the monotonicity of the function $g(t) = 1 - \exp(-\lambda/t) \in (0,1)$ when $t \in [1, N]$,

$$2 \geq \left|M - \int_1^N (1 - \exp(-\lambda/t)dt\right|$$

$$
\begin{aligned}
&= \Big|M - \lambda \int_{\lambda/N}^{\lambda} \frac{1-\exp(-x)}{x^2} dx\Big| \\
&\geq \lambda\Big|\frac{M}{\lambda} - \int_1^{\infty} \frac{1-\exp(-x)}{x^2} dx - \int_0^1 \frac{1-\exp(-x)-x}{x^2} dx - \ln\frac{N}{\lambda}\Big| \\
&\quad -1 - \lambda\Big|\int_0^{\lambda/N} \frac{1-\exp(-x)-x}{x^2} dx\Big| \\
&\geq \lambda\Big|\frac{M}{\lambda} - \ln\frac{N}{\lambda} + 1 - \gamma\Big| - 1 - O\Big(\frac{\lambda^2}{N}\Big).
\end{aligned}
$$

Therefore, instead of solving for n' from Equation 4 with $M = n$, we can solve for n' from the approximated equation

$$\frac{nH_N}{n'} - \ln\frac{NH_N}{n'} = 1 - \gamma. \tag{11}$$

Although an explicit formula for the root n' of Equation 11 does not exist, we can use the fixed-point iteration scheme

$$f_{k+1} = f(f_k),\ f_1 = f(1),\ f(t) = \frac{nH_N}{1-\gamma+\ln(NH_N)-\ln(t)},\ k \in \mathbb{N}. \tag{12}$$

Since $f(t)$ is monotone and grows very slowly, the scheme converges to a fixed point within just a few iterations.

Before proceeding to Theorem 4, which deals with the estimation of Λ_Z, we first adopt the following adjustments. Suppose $X'_1, \ldots, X'_n$ are i.i.d. Zipf(N) random variables, with $X'_{(1)} \leq \ldots \leq X'_{(n)}$ being the corresponding order statistics. Since we have not required the X_i in the sample to be mutually different, the resulting order statistics cause problems since we need to take logarithms of the differences in getting Λ_Z. We therefore adjust the order statistics to $X'_{(1)} < X'_{(2)} + 1 < \ldots < X'_{(n)} + n - 1$. Let the quantile function Q_N be defined such that $Q_N(t) = k$ if $H_{k-1}/H_N \leq t < H_k/H_N$, for $k = 1, 2, \ldots, N$.

Now, for a random variable U uniform on $(0,1)$, the quantile function $Q_N(U)$ as defined above satisfies Zipf$(N) \stackrel{\mathcal{D}}{=} Q_N(U)$. For mathematical convenience, we may take $f_N(t) = N^t, t \in [0,1]$ to approximate $Q_N(t)$, since we have the estimate for the total variation distance

$$
\begin{aligned}
&d_{TV}(Q_N(U), \lfloor f_N(U)\rfloor) := \sup\{|\Pr[Q_N(U) \in A] - \Pr[\lfloor f_N(U)\rfloor \in A]|, A \subset \mathbb{Z}^+\} \\
&\leq \sum_{k=1}^{N} \Big|\frac{k^{-1}}{H_N} - \frac{\ln(k+1)-\ln k}{\ln N}\Big| + \frac{\ln(N+1)-\ln N}{\ln N} = O\Big(\frac{1}{\ln N}\Big).
\end{aligned}
$$

Therefore, we can use $\Delta_U = \sum_{k=1}^{n-1} \ln(N^{U_{(k+1)}} - N^{U_{(k)}} + 1)$ to approximate $\Lambda_Z = \sum_{k=1}^{n-1} \ln(X'_{(k+1)} - X'_{(k)} + 1)$. As to Δ_U, we have the following limit theorem, which asserts that under suitable conditions, the expected value of Δ_U is $\frac{1}{2}(1-\rho_n)^2 n \ln N$, where $\rho_n = \ln n / \ln N$.

Theorem 4. *Let $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ be the order statistics of i.i.d. uniform(0,1) random variables $U_1, \ldots, U_n$. If $\lim_{n\to\infty} n/N_n^{1/2} = 0$, and $\sup_{n>1} \ln N_n / \ln n = C < \infty$, then we have*

$$\frac{\Delta_U}{n \ln N_n} - \frac{1}{2}(1-\rho_n)^2 \xrightarrow{\mathcal{P}} 0,$$

and

$$\frac{E\Delta_U}{n\ln N_n} - \frac{1}{2}(1-\rho_n)^2 \longrightarrow 0,$$

as $n \to \infty$, *where* $\rho_n = \ln n / \ln N$.

Proof. We write $N = N_n$ for simplicity. The Zipf distribution is very skewed towards the high-probability elements, so for any integer $k_0 \in \mathbb{N}$, the first k_0 values in the order statistics $X_{(1)} < X_{(2)} < \ldots < X_{(k_0)}$ are very likely to be $1, 2, \ldots, k_0$. We take $k_0 = \lfloor n\rho_n \rfloor$ here. This observation suggests that $\Lambda'_U = \sum_{k=1}^{k_0-1} \ln(X_{(k+1)} - X_{(k)})$ should be stochastically small. In terms of our approximation, for the corresponding sum $\Delta'_U = \sum_{k=1}^{k_0-1} \ln(N^{U_{(k+1)}} - N^{U_{(k)}} + 1)$, we will prove $\Delta'_U/(n\ln N) \stackrel{\mathcal{P}}{\longrightarrow} 0$. Since the logarithm function is concave, we may apply Jensen's inequality to get

$$\frac{1}{k_0-1}\Delta'_U \le \ln\Big(\frac{1}{k_0-1}\sum_{k=1}^{k_0-1}(N^{U_{(k+1)}} - N^{U_{(k)}} + 1)\Big) \le \ln\Big(\frac{1}{k_0-1}N^{U_{(k_0)}} + 1\Big).$$

Now for any $\varepsilon > 0$,

$$\begin{aligned}
&\Pr\Big[\frac{k_0-1}{n\ln N}\ln\Big(\frac{1}{k_0-1}N^{U_{(k_0)}} + 1\Big) > \varepsilon\Big] \\
&\le \Pr\Big[U_{(k_0)} > \frac{\ln(k_0-1) + \ln(N^{\varepsilon/\rho_n} - 1)}{\ln N}\Big] \\
&= \Pr\Big[\frac{S_{k_0}/k_0}{S_{n+1}/(n+1)} > \frac{n+1}{k_0}\frac{\ln(k_0-1) + \ln(N^{\varepsilon/\rho_n} - 1)}{\ln N}\Big]
\end{aligned}$$

Since $\frac{S_{k_0}/k_0}{S_{n+1}/(n+1)} \stackrel{\mathcal{P}}{\longrightarrow} 1$ by the Weak Law of Large Numbers and

$$\liminf_{n\to\infty} \frac{n+1}{k_0}\frac{\ln(k_0-1) + \ln(N^{\varepsilon/\rho_n} - 1)}{\ln N} \ge \liminf_{n\to\infty}\Big(1 + \frac{\varepsilon}{\rho_n^2}\Big) \ge 1 + \frac{\varepsilon}{C^2},$$

we have $\lim_{n\to\infty} \Delta'_U/(n\ln N) \stackrel{\mathcal{P}}{=} 0$. Next define

$$\begin{aligned}
\Delta''_U &= \sum_{k=k_0}^{n-1} \ln(N^{U_{(k+1)}} - N^{U_{(k)}} + 1) \\
&\stackrel{\mathcal{D}}{=} \frac{\ln N}{S_{n+1}}\sum_{k=k_0}^{n-1} S_{k+1} + \sum_{k=k_0}^{n-1} \ln(1 - N^{-Y_{k+1}/S_{n+1}} + N^{-S_{k+1}/S_{n+1}}) \\
&= I_n + J_n(\text{say}).
\end{aligned}$$

Elementary calculations show that $E[\sum_{k=k_0}^{n-1}(S_{k+1} - k - 1)]^2 \le n^3$, since for exp(1) random variable Y, we know $E(Y-1) = 0, E(Y-1)^2 = 1$. Hence,

$$\frac{S_{n+1}}{n}\Big(\frac{I_n}{n\ln N} - \frac{\sum_{k=k_0}^{n-1}(k+1)}{nS_{n+1}}\Big) = \frac{1}{n^2}\sum_{k=k_0}^{n-1}(S_{k+1} - k - 1) \stackrel{\mathcal{P}}{\longrightarrow} 0,$$

or, $(n \ln N)^{-1} I_n - (1-\rho_n^2)/2 \xrightarrow{\mathcal{P}} 0$. Now we consider J_n. Given any $\varepsilon > 0$, since $0 \geq \ln(1 - N^{-Y_{k+1}/S_{n+1}} + N^{-S_{k+1}/S_{n+1}}) \geq \ln(1 - N^{-Y_{k+1}/S_{n+1}})$,

$$\begin{aligned}
&\Pr\left[\frac{1}{n \ln N}|J_n| > \varepsilon\right] \\
&\leq \Pr\left[\frac{1}{n \ln N} J_n < -\varepsilon, \frac{\ln N}{S_{n+1}} < 1\right] + \Pr\left[\frac{\ln N}{S_{n+1}} \geq 1\right] \\
&\leq \Pr\left[\frac{1}{n \ln N} \sum_{k=k_0}^{n-1} \ln(1 - e^{-Y_{k+1}}) < -\varepsilon\right] + \Pr\left[\frac{\ln N}{S_{n+1}} \geq 1\right]
\end{aligned}$$

Observe that if Y_{k+1} is $\exp(1)$, then $1 - e^{-Y_{k+1}}$ is uniform(0,1), $E \ln(1 - e^{-Y_{k+1}}) = -1$, hence by Markov's inequality, the first term $\leq -(\varepsilon \ln N)^{-1} E \ln(1 - e^{-Y_{k+1}}) = (\varepsilon \ln N)^{-1}$. Obviously, the second term goes 0 via the Weak Law of Large Numbers, which completes the proof of the first statement of Theorem 4. By Jensen's inequality,

$$0 < \frac{\Delta_U}{n \ln N_n} \leq \frac{1}{\ln N_n} \ln\left[\frac{1}{n-1} \sum_{k=1}^{n-1} \left(N^{U_{(k+1)}} - N^{U_{(k)}} + 1\right)\right] < \frac{\ln(N_n + n)}{\ln N_n} < 2,$$

hence random variables $\{\Delta_U/(n \ln N_n) - (1-\rho_n)^2/2, n \geq 2\}$ are uniformly integrable. Then the second convergence result stated in the theorem follows easily from the first one and the Mean Convergence Criterion [Chow and Teicher, 1988].

□

Remark 2. Under the conditions of Theorem 4, a more careful analysis leads to the stronger result

$$\lim_{n \to \infty} \frac{\Delta_U - \mu_n'}{n} \overset{\mathcal{P}}{=} 0,$$

where $\mu_n' = (1/2)(1 - \rho_n^2) n \ln N + n(1 - \rho_n)(\ln \ln N - \gamma - \ln n)$, $\gamma = 0.5772\ldots$ is Euler's constant. Since the details of the proof are complicated, we omit the proof and only provide an outline here. First, to obtain $\Delta_U'/n \xrightarrow{\mathcal{P}} 0$, we use the Law of the Iterated Logarithm [Chow and Teicher, 1988] $\limsup_{n \to \infty} |(S_n - n)/\sqrt{n \ln \ln n}| = \sqrt{2}$, a much finer estimate than we can obtain from WLLN. For I_n, the estimate used in the proof of Theorem 2 can yield $n^{-1} I_n - (1/2)(1 - \rho_n^2) \ln N \xrightarrow{\mathcal{P}} 0$. Since $\ln N^{-Y_{k+1}/S_{n+1}} \xrightarrow{\mathcal{P}} 0$, we can use Taylor's expansion $1 - N^{-Y_{k+1}/S_{n+1}} \approx (\ln N) Y_{k+1}/S_{n+1}$. Hence J_n can be further approximated by $-(n - k_0)(\gamma + \ln n)$ by the usual SLLN $(n - k_0)^{-1} \sum_{k=k_0}^{n} \ln Y_{k+1} \to E \ln Y = -\gamma$ and $S_{n+1}/n \to 1$ a.s.. Together, these facts imply the refined limit theorem.

Figure 2 evaluates how well Remark 2 matches the results of experiments on databases of different sizes containing integers drawn without replacement from a Zipf distribution over $\{1, 2, \cdots, 2^{31} - 1\}$. To validate both the analysis and the approximations driving it, we used the actual value of Λ_Z obtained from experiments in place of Δ_U. Figure 2 shows the percentage difference between $\Lambda_Z/(n \ln N)$ and $\mu_n'/(n \ln N)$. Agreement is to within a few percent even for databases that are quite large, suggesting that our formula is an excellent predictor of experimental results.

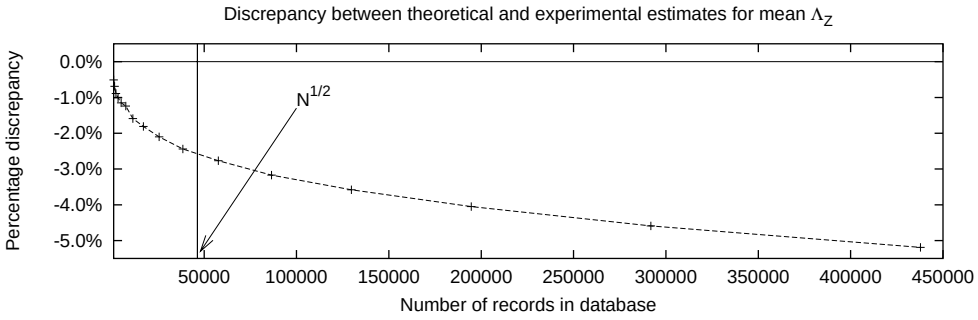

Figure 2: Zipf distribution: agreement between Remark 2 and experiment

We appear to have satisfactorily addressed the problem of estimating Λ_Z for relatively large databases. However, the situation for small databases is somewhat different, since the small number of samples means that the spacings between them are likely to be larger. We now address the case where the database is small, and present the following result.

Theorem 5. *Let $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ be the order statistics of i.i.d. uniform(0,1) random variables $U_1, \ldots, U_n$. If $\lim_{n\to\infty} n/\ln N_n = 0$, then we have*

$$\lim_{n\to\infty} \frac{\Delta_U}{n \ln N_n} - \frac{1}{2} \stackrel{\mathcal{P}}{=} 0.$$

and

$$\lim_{n\to\infty} \frac{E\Delta_U}{n \ln N_n} - \frac{1}{2} = 0.$$

Proof. As in the proof in Theorem 4, we write

$$\begin{aligned}
\Delta_U &= \sum_{k=1}^{n-1} \ln(N^{U_{(k+1)}} - N^{U_{(k)}} + 1) \\
&\stackrel{\mathcal{D}}{=} \frac{\ln N}{S_{n+1}} \sum_{k=1}^{n-1} S_{k+1} + \sum_{k=1}^{n-1} \ln(1 - N^{-Y_{k+1}/S_{n+1}} + N^{-S_{k+1}/S_{n+1}}) \\
&= I_n + J_n(\text{say}).
\end{aligned}$$

Using the same argument as in Theorem 4, we have $(n \ln N)^{-1} I_n - 1/2 \stackrel{\mathcal{P}}{\longrightarrow} 0$. For any $\varepsilon > 0$, let $n > n_0$ be large enough such that $(\ln N)^{-1}(n+1) < 1/2$, then

$$\begin{aligned}
&\Pr\left[\frac{1}{n \ln N}|J_n| > \varepsilon\right] \\
&\leq \Pr\left[\frac{1}{n \ln N} \sum_{k=1}^{n-1} \ln(1 - N^{-Y_{k+1}/S_{n+1}}) < -\varepsilon, \frac{S_{n+1}}{n+1} \leq 2\right] + \Pr\left[\frac{S_{n+1}}{n+1} \geq 2\right] \\
&\leq \Pr\left[\frac{1}{n \ln N} \sum_{k=1}^{n-1} \ln(1 - e^{-Y_{k+1}}) < -\varepsilon\right] + \Pr\left[\frac{S_{n+1}}{n+1} \geq 2\right].
\end{aligned}$$

Again by the same arguments as in Theorem 4, we know $(n \ln N)^{-1} J_N \xrightarrow{\mathcal{P}} 0$. Thus the second convergence result stated in the theorem follows from the first one via uniform integrability, which is an immediate consequence of the uniform boundedness of the random sequence $\{(n \ln N_n)^{-1} \Delta_U - 1/2, n \geq 2\}$. □

Remark 3. Under the conditions of Theorem 5, we have $\lim_{n\to\infty} \ln n / \ln N_n = 0$, thus $\rho_n \approx 0$. Then interestingly enough, both Theorem 5 and Theorem 4 are consistent, and give the result $E\Delta_U \approx (1/2) n \ln N_n$.

4.3 Spacings for distributions with high concentration

A nonnegative integer-valued random variable Z is said to be highly concentrated if Z takes values in a set of few elements with high probability. Thus, the Binomial, Poisson, Geometric, or general Zipf distribution are highly concentrated. (The general Zipf distribution is defined by $\Pr[Z = k] \sim ck^{-\alpha}$, as $k \to \infty$, $\alpha > 1$.) When highly concentrated distributions are sampled without replacement, the spacings $Z_{(2)} - Z_{(1)}, \ldots, Z_{(n)} - Z_{(n-1)}$ are very likely to be 1, where $Z_{(1)}, \ldots, Z_{(n)}$ are the order statistics of n samples $Z_1, \ldots, Z_n$. Thus, the total number of bits required in this case is likely to be close to $O(n)$. Defining Λ_Z simply as the sum of the logarithms of the difference will lead to a smaller estimate since the logarithmic terms corresponding to differences of 1 will be zero. Fortunately, adopting the conservative form $\Lambda_Z = \sum_{k=1}^{n-1} \ln(Z_{(k+1)} - Z_{(k)} + 1)$ suggested in Section 3 leads to $\Lambda_Z \approx n \ln 2$, in perfect agreement with practice.

5 Optimal ordering of attribute domains

When multiple attribute fields $X_1, X_2, \cdots, X_k$ are present in a database tuple, it is clear that the ordering of the attribute fields in will influence the value resulting from the application of φ (see Section 2) to the tuples. In this section, we consider the question of how to order the attribute fields so that $E\Lambda$ reaches its minimum.

5.1 Uniform attribute domains

Consider first the case when the k fields are all uniformly distributed, so that X_i is uniform over $(1, |D_i|)$. Somewhat contrary to intuition, $E\Lambda$ will remain unaffected in this case by attribute domain reordering, since the random integer $X_1 \cdot |D_2| \cdot |D_3| \cdots |D_k| + X_2 \cdot |D_3| \cdot |D_4| \cdots |D_k| + \cdots + X_k$ is, regardless of field ordering, always distributed uniformly over the set $\{a+1, a+2, \ldots, a+b\}$, if we define $a = |D_2| \cdot |D_3| \cdots |D_k| + |D_3| \cdot |D_4| \cdots |D_k| + \cdots + |D_k|$, and $b = |D_1| \cdot |D_2| \cdots |D_k|$. This somewhat paradoxical result is confirmed by our simulations, which are shown in Figure 3.

5.2 Non-uniform attribute domains

The case of non-uniform attributes is more complex. In fact, the optimal attribute ordering actually depends on the database size. A full analysis is elusive, but we provide general characterizations of behaviors for different cases.

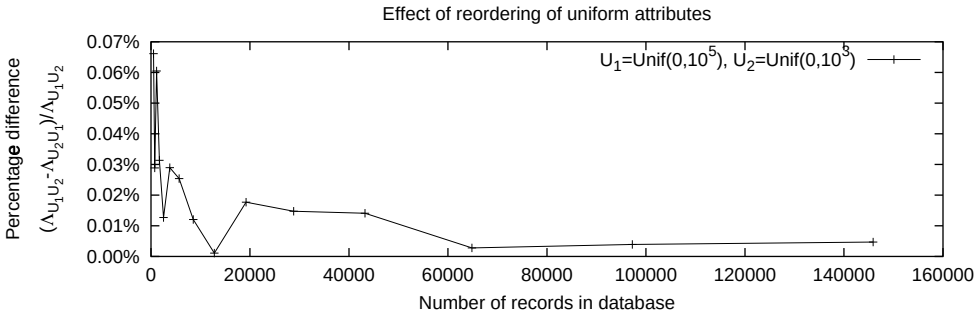

Figure 3: Two uniform attributes: effect of attribute reordering on $(\Lambda_{U_1U_2} - \Lambda_{U_2U_1})/\Lambda_{U_1U_2}$

5.2.1 Small databases

Let us first consider the simplest case, where there are only two fields, and the database contains just two records. We will use the analysis for this case to provide insights into more general situations. Suppose that X, Y are independent random variables distributed as Zipf(m) and Zipf(n) respectively. Therefore, $Z = (X, Y) = nX + Y$ has distribution function

$$F_Z(z) = \Pr[Z \le z] = \frac{H_{x-1}}{H_m} + \frac{H_y}{xH_n} \approx \frac{\ln x + \gamma}{\ln m + \gamma} \approx \frac{\ln(z/n) + \gamma}{\ln m + \gamma} := \tilde{F}_Z(z)$$

for $z = y + nx, x = 1, \ldots, m, y = 1, \ldots, n$. Hence Z can be approximated by random variable $\tilde{F}_Z{}^{-1}(U) = n \exp[U(\ln m + \gamma) - \gamma]$, where U is uniform on $(0, 1)$. Take Z_1, Z_2 to be i.i.d. copies of Z with order statistics $Z_{(1)} \le Z_{(2)}$. Now,

$$\begin{aligned} E\Lambda_{xy} = E \ln (Z_{(2)} - Z_{(1)} + 1) &\approx E \ln \left(n e^{U_{(2)}(\ln m + \gamma) - \gamma} - n e^{U_{(1)}(\ln m + \gamma) - \gamma} \right) \\ = \ln n - \gamma + (\ln m + \gamma)[EU_{(1)} + E(U_{(2)} - U_{(1)})] &= \ln n + \frac{2}{3} \ln m - \frac{\gamma}{3} \end{aligned}$$

We may, but do not derive this asymptotic formula from the original distribution function $F(z)$ since that route involves elementary but tedious calculations. We observe that the random variable $\tilde{F}_Z{}^{-1}(U)$ does not take the distribution of Y into account, which appears reasonable as the first field will dominate Λ when dealing with a sample size of two. Hence this approach also works for any discrete random variables Y taking possibly n values.

The same idea works when X is uniform on $(1, m)$.

For integer valued random variable Y taking at most n values, We use $\tilde{F}_U{}^{-1}(U) = mnU$ to replace $Z = (X, Y)$ since

$$F_U(z) = \Pr[nX + Y \le nx + y] \approx \frac{x}{m} \approx \frac{z}{mn}.$$

As before,

$$E\Lambda_{xy} = E \ln (Z_{(2)} - Z_{(1)} + 1) \approx E \ln (mnU_{(2)} - mnU_{(1)}) = \ln m + \ln n - \frac{11}{6}.$$

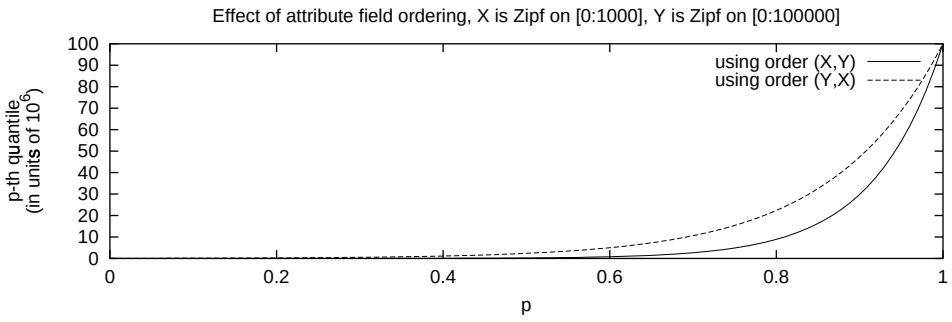

Figure 4: Two Zipf attributes: characterizing skew through the distribution of quantiles

where $Z_{(1)} \leq Z_{(2)}$ is the order statistics of Z_1, Z_2.

Now let us assume $n > m$. From the formulae above, if X is Zipf(m) and Y is Zipf(n), then

$$E\Lambda_{xy} \approx \ln n + \frac{2}{3}\ln m - \frac{\gamma}{3} > \ln m + \frac{2}{3}\ln n - \frac{\gamma}{3} \approx E\Lambda_{yx}$$

suggests that we need to put field Y first to minimize $E\Lambda$.

This result also appears paradoxical, since the domain of X is smaller than that of Y. Intuition might have suggested that placing X before Y would result both in smaller values of φ, as well as longer runs of leading zeroes in the sequence of of differences, leading to a lower value of Λ. This apparent contradiction can be resolved by considering the skew and concentration effects of the distributions involved. For any $p \in (0,1)$, the order (X,Y) gives the p-percentile $P_1(p) = \tilde{F_Z}^{-1}(p) = n\exp[p(\ln m + \gamma) - \gamma]$, by $\Pr[(X,Y) \leq P_1(p)] = p$, while the order (Y,X) gives the p-percentile $P_2(p) = m\exp[p(\ln n + \gamma) - \gamma] < P_1(p)$. Hence the latter is more skewed than the former, and consequently, the sample data is more likely to be concentrated on the left extreme, reducing $E\Lambda$. Figure 4 convincingly suggests this relationship by displaying the quantiles.

We may also interpret this phenomenon in terms of the distribution functions. Clearly, for integers $1 \leq x \leq m, 1 \leq y \leq n$, we have $\Pr[(X,Y) \leq (x,y)] = \Pr[X < x] + \Pr[X = x, Y \leq y] = H_{x-1}/H_m + H_y/(xH_mH_n)$ and $\Pr[(Y,X) \leq (y,x)] = \Pr[Y < y] + \Pr[Y = y, X \leq x] = H_{y-1}/H_n + H_x/(yH_mH_n)$. It can be shown that $\Pr[(X,Y) \leq (x,y)] \leq \Pr[(Y,X) \leq (y',x')]$, if $1 \leq x, x' \leq m, 1 \leq y, y' \leq n$ and $xn + y = y'm + x'$ through a rather complicated calculation.

Another extreme case is when all the fields are Zipf distributed so that X_i is Zipf$(|D_i|)$. In this situation, the analysis above suggests that to minimize $E\Lambda$, we order fields so that the first Zipf field corresponds to the largest value of $|D_i|$, the second field has the second largest value, and so on. If there exist both Zipf and uniform distributions among those fields, one should put those field with Zipf distributions first, then those with uniform distributions. Similarly, when there are fields with arbitrary non-uniform distributions, we place the uniformly distributed fields last and the field with the highest concentration first, and then the field with the second highest concentration, and so on. Figure 5 illustrates this effect by showing the values of Λ_{ZU} and Λ_{UZ} obtained through experiment.

Our analysis began by assuming a database size of 2, but can clearly be extended to databases of size small relative to $N = \prod_i |D_i|$. The concentration effect is again the key to determining

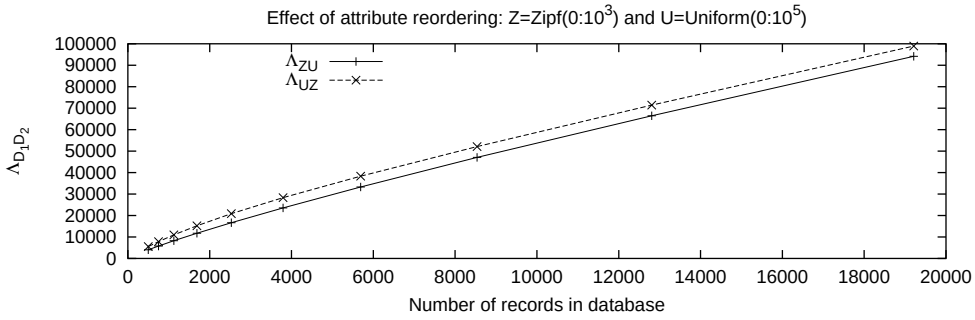

Figure 5: Attribute reordering: One Zipf and one uniform attribute

the optimal ordering. We note however, that for two Zipf random variables, the advantage of optimal ordering over an arbitrary ordering seems small since $E\Lambda_{xy} - E\Lambda_{yx} \approx 1/3 \ln(n/m)$, which is significant only when the ratio n/m is extremely large. Even for $n = 10^{16}$ and $m = 10$, the difference is merely $5 \ln 10$, which is not very significant.

5.2.2 Large databases

Consider now the case when database size n is large, but we still have two Zipf attributes X and Y. The situation is now quite different, since the concentration effect will no longer be crucial in determining Λ. Whether we order the attributes as (X, Y) or (Y, X), it is very likely that the initial segment of the order statistics $Z_{(1)} < Z_{(2)} < \ldots < Z_{(d)}$ will be the first consecutive d integers for some $d \in \mathbb{N}$. Thus, the lower values in the Zipf range are very likely to be quickly exhausted; so that their contribution to Λ are small since n is relatively large. The main contributions to Λ will come from the elements at the right extreme of the these order statistics.

From Jensen's inequality, we have $\Lambda_Z \leq (k-1) \ln[(Z_{(k)} - Z_{(1)})/(k-1)]$. It is intuitively clear that the two sides of this inequality will be closer together if spacings $Z_{(2)} - Z_{(1)}, \ldots, Z_{(k)} - Z_{(k-1)}$ are close to each other. Since k is large, $Z_{(k)}$ is close to mn for both orderings. Therefore, non-uniformity within the set of spacings is really an issue. Such non-uniformity is most significant at the right extreme, and more uniformity will lead to higher Λ. The quantile plot shown in Figure 4 of (X, Y) and (Y, X) shows that the former displays less uniformity at the right extreme, so that we expect the corresponding Λ to be smaller. Figure 6 illustrates this effect through experiments on two Zipf domains with $Z_1 < Z_2$. For smaller database sizes, the order $Z_2 Z_1$ yields lower Λ values (in agreement with our results in Section 5.2.1), while the order $Z_1 Z_2$ is better for larger database sizes.

Although this discussion provides adequate intuition for understanding the difference between the two orderings for small and large database sizes, it appears to be quite difficult to quantify and compare the effects caused by concentration and non-uniformity. It also appears difficult to determine the borderline represented by the value of k. We suggest that if $k < m = \min(m, n)$, then we use ordering (Y, X) and otherwise we use (X, Y).

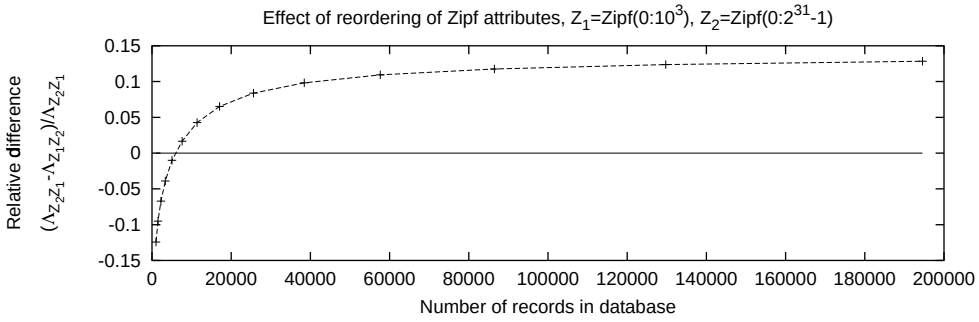

Figure 6: Attribute reordering: two Zipf attributes

6 Limit Theorems for the multi-field Case

We now turn to the problem of estimating values of Λ when the database has several fields. Suppose the database has k fields drawn from independent domains $D_1, \ldots, D_k$, respectively. Consider the corresponding random vector $\vec{X} = (X_1, \ldots, X_k)$ with X_i taking values in $\{1, \ldots, |D_i|\}, i = 1, \ldots, k$. As in Section 2, this random vector can be represented by the corresponding random integer $X_1 \cdot |D_2| \cdots |D_k| + \cdots + X_{k-1} \cdot |D_k| + X_k$.

Our analysis in Section 5 showed that lower values of Λ result when the uniformly distributed domains are placed at the least-significant end of the tuple. In this case, the remaining non-uniform domains will be placed in some suitable order at the head of the tuple. We may view these non-uniformly distributed domains as jointly constituting a single composite domain with an arbitrary discrete distribution.

Let us therefore model the non-uniform domains $X_1, \cdots, X_{m-1}$ as a single random variable Z with and arbitrary distribution, and assuming values in the set $\{1, \ldots, d\}$ with probabilities $\Pr[Z = k] = p_k > 0, k = 1, \ldots, d$ for some fixed $d \in \mathbb{N}$. The remaining fields $X_m, \ldots, X_k$ have uniform distributions, whence $X_m \cdot |D_3| \cdots |D_k| + \cdots + X_{k-1} \cdot |D_k| + X_k$ may be collapsed into a single random variable U uniformly distributed over the set $\{a + 1, a + 2, \ldots, a + b\}$, where $a = |D_{m+1}| \cdots |D_k| + \cdots + |D_{k-1}| \cdot |D_k| + |D_k|$ and $b = |D_m| \cdots |D_k|$.

Thus, in estimating Λ, we can collapse the fields $X_1, \cdots, X_k$ into just two fields. Let Z be an arbitrary discrete random variable assuming values from $\{1, 2, \cdots, d\}$, and U be uniform on $\{1, 2, \cdots, u_n\}$. Let Z and U be independent, and form the random vector (Z, U). Let $Y_i = (Z_i, U_i), i = 1 \ldots, n$, be n i.i.d. copies of this vector. Take $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$ to be the order statistics of the Y_i. Now form the statistic $\Lambda_{DU} = \sum_{i=1}^{n-1} \ln(Y_{(i+1)} - Y_{(i)} + 1)$.

If Z is uniform on $\{1, 2, \cdots, d\}$, then (Z, U) can be viewed as single large uniformly distributed field. Theorem 2 can be directly applied to obtain the following result.

Corollary 2. If $n^2 = o(u_n)$, then $(\Lambda_{DU} - \mu_n)/\sigma \xrightarrow{\mathcal{D}} N(0, 1)$, where $\sigma = \alpha n^{1/2}, \mu_n = (n-1)(\ln d + \ln u_n - \gamma - \ln(n+1))$.

If the distribution of Z is not uniform, we may proceed as follows. Since Z can take at most d values, we would expect to see groups of tuples in the database sharing the same value in their first fields. When the database is sorted, tuples in each such cluster will appear together, and their differences will show a zero value in the first field. We call each such cluster of tuples in the

sorted database a *run*. Therefore, we may split the original database into d smaller databases, each defined by a run corresponding to a value of Z, with the jth run having $N_j = \sum_{i=1}^{n} 1(Z_i = j)$ records. Since we may have $N_j = 0$, we allow runs to be empty. Consequently, Λ_{DU}, the overall statistic to estimate database size, can be decomposed into two components: one to model the spacings within the runs, and one to model the spacings across the runs. That is,

$$\Lambda_{DU} \stackrel{\mathcal{D}}{=} \sum_{j=1}^{d} \Lambda_j + \sum_{j=1}^{d-1} \ln(U_{j+1,1} - U_{j,N_j} + u_n + 1) := \Lambda_w + \Lambda_b.$$

In this formula, $\Lambda_j = \sum_{i=1}^{N_j - 1} \ln(U_{j,i+1} - U_{j,i} + 1)$, or given $N_j = l$, $\Lambda_j = \Lambda_j(l) = \sum_{i=1}^{l-1} \ln(U_{j,i+1} - U_{j,i} + 1)$, $U_{j,1} \le U_{j,2} \le \ldots U_{j,l}$ is the order statistics of $\bar{U}_{j,1}, \ldots, \bar{U}_{j,l}$, where $\{\bar{U}_{j,i}, 1 \le j \le d, i \ge 0\}$, are i.i.d. random variables uniform on $\{1, \cdots, u_n\}$ and independent of $Z_1, \ldots, Z_n$. Λ_w is the contribution to Λ_{DU} from within runs, and Λ_b can be regarded as the spacings between consecutive runs.

Obviously, $(N_1, \ldots, N_d)$ follows the multinomial distribution $Multi(n; p_1, \ldots, p_d)$, so that N_j has distribution $Bin(n; p_j)$. As in [Shiryaev, 1995], we may therefore write the inequality $\Pr[|N_j/n - p_j| \ge \varepsilon] \le 2\exp(-2n\varepsilon^2)$ for every $\varepsilon > 0$. When $N_j = 0$ or 1, we use the convention $\Lambda_j = 0$, and define the corresponding summand in Λ_b to be 0. However, given the the large-deviation style inequality above, we are assured that $N_j = 0$ or 1 with exponentially small probabilities. Therefore, in pursuing the limiting distribution of Λ_{DU} in Theorem 6, we may assume without undue concern that $N_j \ge 2$.

We now state the main theorem that allows us to estimate the size of a compressed database with multiple attributes.

Theorem 6. *Let each record in the database comprise two discrete random fields (Z, U), where Z is an arbitrary distribution on $\{1, \cdots, d\}$, and U is uniform on $\{1, \cdots, u_n\}$. If $n^2 = o(u_n)$, then as $n \to \infty$,*

$$\frac{\Lambda_{DU} - \mu_{DU}}{\beta n^{1/2}} \stackrel{\mathcal{D}}{\longrightarrow} N(0, 1),$$

where

$$\mu_{DU} = \sum_{j=1}^{d} (np_j - 1)(\ln u_n - \ln(np_j) - \gamma) + \sum_{j=1}^{d-1} \ln\left(\frac{u_n}{np_j} + \frac{u_n}{np_{j+1}}\right) := \mu_w + \mu_b$$

and

$$\beta^2 = \alpha^2 + \sum_{j=1}^{d} p_j (\ln p_j)^2 - \Big(\sum_{j=1}^{d} p_j \ln p_j\Big)^2, \alpha^2 = \pi^2/6 - 1 = 0.644934\ldots.$$

Proof. We first introduce some heuristics. Since N_j has distribution $Bin(n; p_j)$, we can replace N_j with the mean np_j. Then $EU_{j+1,1} \approx u_n/(np_{j+1}), E(u_n - U_{j,N_j}) \approx u_n/(np_j)$, so we approximate $E\Lambda_b$ by μ_b. By Theorem 2, the mean $E\Lambda_w \approx \mu_w$ and the variance is $\sum_{j=1}^{d} \alpha^2 np_j = \alpha^2 n$. The part $n[\sum_{j=1}^{d} p_j (\ln p_j)^2 - (\sum_{j=1}^{d} p_j \ln p_j)^2]$ in the overall variance $n\beta^2$ can be interpreted as the uncertainty in choosing differences across runs, which corresponds to Λ_b.

Now for $k \geq 1$, set $\mu(k) = (k-1)(-\gamma + \ln u_n - \ln k)$, and let $\mu(k) = 0$ if $k \leq 1$. To obtain the limiting distribution, we apply the Lévy Continuity Theorem by analyzing characteristic functions. We first note that given $N_1 = n_1, \ldots, N_d = n_d$, $\Lambda_1, \ldots, \Lambda_d$ are independent. Hence for $t \in \mathbb{R}$, we have via conditioning,

$$\begin{aligned}
&\exp\left[\sqrt{-1}t\sum_{j=1}^{d}\frac{\Lambda_j - \mu(np_j)}{\sqrt{n}}\right] \\
&= E\left\{E\left[\exp\left(\sqrt{-1}t\sum_{j=1}^{d}\frac{\mu(N_j)-\mu(np_j)}{\sqrt{n}} + \sqrt{-1}t\sum_{j=1}^{d}\frac{\Lambda_j-\mu(N_j)}{\sqrt{n}}\right)\middle| N_1,\ldots,N_d\right]\right\} \\
&= E\left\{\exp\left(\sqrt{-1}t\sum_{j=1}^{d}\frac{\mu(N_j)-\mu(np_j)}{\sqrt{n}}\right)\prod_{j=1}^{d}E\left[\exp\left(\sqrt{-1}t\frac{\Lambda_j-\mu(N_j)}{\sqrt{n}}\right)\middle| N_j\right]\right\}
\end{aligned}$$

We next assert the three convergence results (13), (14), and (15) and proceed to prove them using the Lebesgue Dominated Convergence Theorem and Slutsky's Theorem. These results will lead to Theorem 6 via the Lévy Continuity Theorem.

$$E\left[\exp\left(\sqrt{-1}t\frac{\Lambda_j-\mu(N_j)}{\sqrt{n}}\right)\middle| N_j\right] \longrightarrow e^{-\alpha^2 p_j t^2/2} \text{ a.s. }, \tag{13}$$

$$\sum_{j=1}^{d}\frac{\mu(N_j)-\mu(np_j)}{\sqrt{n}} \xrightarrow{\mathcal{D}} N\left(0, \sum_{j=1}^{d}p_j(\ln p_j)^2 - \left(\sum_{j=1}^{d}p_j\ln p_j\right)^2\right), \tag{14}$$

and for $j = 1, \ldots, d$,

$$\frac{1}{n^{1/2}}\left[\ln(U_{j+1,1} - U_{j,N_j} + u_n + 1) - \ln\left(\frac{u_n}{np_j} + \frac{u_n}{np_{j+1}}\right)\right] \xrightarrow{\mathcal{P}} 0. \tag{15}$$

To show (13), we proceed as follows. Since event $\{N_j = l\}$ and $\Lambda_j(l)$ are independent,

$$E\left[\exp\left(\sqrt{-1}t\frac{\Lambda_j-\mu(N_j)}{\sqrt{n}}\right)\middle| N_j = l\right] = E\left[\exp\left(\sqrt{-1}t\frac{\Lambda_j(l)-\mu(l)}{\sqrt{n}}\right)\right] := g(t;n,l).$$

Define set $\mathcal{I}_n = \{l \in \mathbb{N}, l \in (np_j - n^{2/3}, np_j + n^{2/3})\}$. Observing that

$$\begin{aligned}
&\limsup_{n\to\infty}\left|E\left[\exp\left(\sqrt{-1}t\frac{\Lambda_j-\mu(N_j)}{\sqrt{n}}\right)\middle| N_j\right] - e^{-\alpha^2 p_j t^2/2}\right| \\
&= \limsup_{n\to\infty}\left(\sum_{l\in\mathcal{I}_n} + \sum_{l\notin\mathcal{I}_n}\right)\left|g(t;n,l) - e^{-\alpha^2 p_j t^2/2}\right|1(N_j = l) \\
&\leq \limsup_{n\to\infty}\sup_{l\in\mathcal{I}_n}\left|g(t;n,l) - e^{-\alpha^2 p_j t^2/2}\right| + 2\limsup_{n\to\infty}1(N_j \notin \mathcal{I}_n) := A + B,
\end{aligned}$$

for (13), we only need to show $A = 0, B = 0$ a.s.. Again by inequality $\Pr[|N_j/n - p_j| \geq \varepsilon] \leq 2\exp(-2n\varepsilon^2)$, $\lim_{k\to\infty}\sum_{n=k}^{\infty}\Pr[N_j \notin \mathcal{I}_n] \leq 2\lim_{k\to\infty}\sum_{n=k}^{\infty}\exp[-2n(n^{-1/3})^2] = 0$. Hence $B = 0$

a.s. via the Borel–Cantelli lemma. If $A \neq 0$, then there exists an $\varepsilon > 0$, a subsequence $\{n'\} \subset \mathbb{N}$ and $l(n') \in \mathcal{I}_{n'}$ such that along this subsequence, $|g(t; n', l(n')) - e^{-\alpha^2 p_j t^2/2}| > \varepsilon$. However, by Lévy Continuity Theorem, we do have $|g(t; n', l(n')) - e^{-\alpha^2 p_j t^2/2}| \to 0$ following from $(n')^{-1/2}[\Lambda_j(l(n')) - \mu(l(n'))] \xrightarrow{\mathcal{D}} N(0, \alpha^2 p_j)$, which is due to $l(n')/n' \to p_j$ and $[\alpha^2 l(n')]^{-1/2}[\Lambda_j(l(n')) - \mu(l(n'))] \xrightarrow{\mathcal{D}} N(0,1)$ asserted by Theorem 2 since $l(n') \to \infty$.

To prove (14), define $\hat{p}_n = (N_1/n, \ldots, N_d/n), \hat{p} = (p_1, \ldots, p_d)$, and the entropy function $\nu(\hat{q}) = \sum_{j=1}^d q_j \ln q_j$ for a d-dimensional probability vector $\hat{q} = (q_1, \ldots, q_d)$. By the classical Central Limit Theorem for vectors, we have $n^{1/2}(\hat{p}_n - \hat{p}) \xrightarrow{\mathcal{D}} N(0, \Sigma)$, where Σ is a $d \times d$ positive definite matrix with $\Sigma_{ii} = p_i(1 - p_i), \Sigma_{ij} = -p_i p_j$. Using the Delta method described in [van der Vaart., 1998], we have

$$n^{1/2}[\nu(\hat{p}_n) - \nu(\hat{p})] \xrightarrow{\mathcal{D}} N\left(0, \frac{\partial \nu}{\partial \hat{q}}\bigg|_{\hat{p}} \Sigma \Big(\frac{\partial \nu}{\partial \hat{q}}\Big)^\tau \bigg|_{\hat{p}}\right),$$

which we can use to prove (14) by writing

$$\sum_{j=1}^d [\mu(N_j) - \mu(np_j)] = \sum_{j=1}^d \ln \frac{\hat{p}_{n,j}}{p_j} + n[\nu(\hat{p}) - \nu(\hat{p}_n)],$$

since

$$\hat{p}_{n,j} \to p_j \text{ a.s. and } \frac{\partial \nu}{\partial q}\bigg|_{\hat{p}} \Sigma \Big(\frac{\partial \nu}{\partial q}\Big)^\tau \bigg|_{\hat{p}} = \sum_{j=1}^d p_i (\ln p_i)^2 - \Big(\sum_{j=1}^d p_i \ln p_i\Big)^2.$$

For (15), we only need to show that

$$\frac{np_{j+1}}{u_n} U_{j+1,1} = O_P(1), \frac{np_j}{u_n}\Big(u_n - U_{j,N_j}\Big) = O_P(1).$$

The notation $X_n = O_P(1)$, as in [van der Vaart., 1998], means that the random sequence X_n is stochastically bounded; i.e., for each $\varepsilon > 0$, there exists a $K = K(\varepsilon) > 0$ such that $\sup_{n \geq 1} \Pr[|X_n| > K] < \varepsilon$. In fact, it is possible to obtain the stronger result

$$\frac{np_{j+1}}{u_n} U_{j+1,1} \xrightarrow{\mathcal{D}} \exp(1), \frac{np_j}{u_n}\Big(u_n - U_{j,N_j}\Big) \xrightarrow{\mathcal{D}} \exp(1). \tag{16}$$

Here we only prove the second case since the first one can be derived similarly. Actually, for $x \geq 0$,

$$\begin{aligned}
\lim_{n\to\infty} \Pr\Big[\frac{np_j}{u_n}\Big(u_n - U_{j,N_j}\Big) > x\Big] &= \lim_{n\to\infty} \sum_{l=1}^\infty \Pr\Big[\frac{np_j}{u_n}\Big(u_n - U_{j,l}\Big) > x, N_j = l\Big] \\
&= \lim_{n\to\infty} \sum_{l=1}^\infty \Big(\frac{1}{u_n}\Big[u_n - \frac{u_n x}{np_j}\Big]\Big)^l \Pr[N_j = l] = \lim_{n\to\infty} E\Big[\Big(\frac{1}{u_n}\Big[u_n - \frac{u_n x}{np_j}\Big]\Big)^{np_j}\Big]^{N_j/(np_j)} = e^{-x}
\end{aligned}$$

The last step follows from the Lebesgue Dominated Convergence Theorem.

□

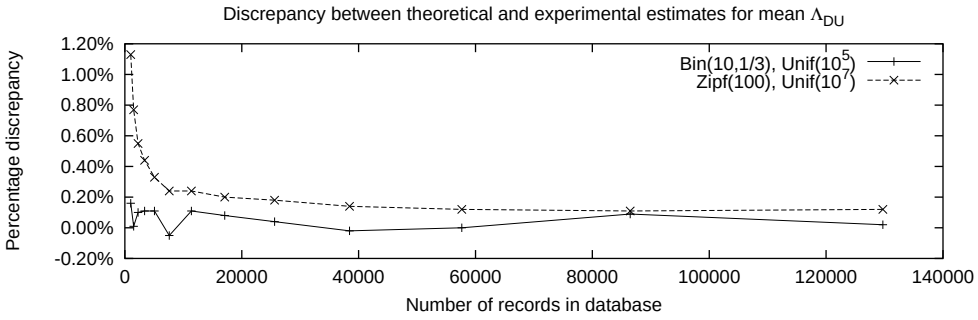

Figure 7: Multiple fields: agreement between Theorem 6 and experiment

Figure 7 shows how closely Theorem 6 agrees with values of Λ observed in practice. We generated two datasets, each with two distributions, one skewed and one uniform. The skewed distribution in the first dataset was Zipf(100), and its second field being uniform over $(0, 10^7)$. The other dataset had its first field distributed as $Binomial(10, 1/3)$, with its second field being uniform over $(0, 10^5)$. Agreement in both cases is to within a fraction of one percent over a large range, illustrating the power of Theorem 6.

Remark 4. The reader may observe that, in order to obtain a more accurate estimate of $E\Lambda_b = \sum_{j=1}^{d-1} \ln(U_{j+1,1} - U_{j,N_j} + u_n + 1)$, one must take advantage of the limiting distributions of $U_{j+1,1}$ and $u_n - U_{j,N_j}$ specified by (16) since bias will be caused if we directly replace $U_{j+1,1}, u_n - U_{j,N_j}$ by their asymptotic means $u_n/(np_{j+1}), u_n/(np_j)$. This goal can be achieved by the following steps. (We again omit the details because of the overwhelming complexity.) First, given N_j and N_{j+1}, $U_{j+1,1}, u_n - U_{j,N_j}$ are independent, since N_j, N_{j+1} are asymptotically independent. So we have

$$\frac{np_{j+1}}{u_n} U_{j+1,1} + \frac{np_j}{u_n}\Big(u_n - U_{j,N_j}\Big) \xrightarrow{\mathcal{D}} \frac{Y_1}{p_{j+1}} + \frac{Y_2}{p_j},$$

where Y_1, Y_2 are two i.i.d. exp(1) random variables. Next, following a careful estimation, the random sequence in the proceeding display can be shown to be uniform integrable. Hence,

$$\lim_{n\to\infty} E\Lambda_b - (d-1)\ln\frac{u_n}{n} = \sum_{j=1}^{d-1} E\ln\Big(\frac{Y_1}{p_{j+1}} + \frac{Y_2}{p_j}\Big).$$

Finally, an elementary but interesting computation leads to

$$E\ln\Big(\frac{Y_1}{p_{j+1}} + \frac{Y_2}{p_j}\Big) = \int_0^\infty \int_0^\infty e^{-(s+t)} \ln\Big(\frac{s}{p_{j+1}} + \frac{t}{p_j}\Big) dsdt = \frac{p_j \ln p_{j+1} - p_{j+1}\ln p_j}{p_{j+1} - p_j} - \gamma$$

through the parameter transformation $x = s/p_{j+1} + t/p_j, y = s + t$. To summarize, we outlined a better estimate

$$E\Lambda_b = (d-1)\Big[\ln\frac{u_n}{n} - \gamma\Big] + \sum_{j=1}^{d-1} \frac{p_j \ln p_{j+1} - p_{j+1}\ln p_j}{p_{j+1} - p_j}.$$

Remark 5. If the condition $n^2 = o(u_n)$ does not hold, we would use the techniques in Section 3. An equivalent sampling scheme must be adopted with n replaced by the adjusted size n'.

7 Conclusions and future Work

This paper provides the theoretical foundations for the Tuple Difference Coding method for compressing Large databases and data warehouses. As already noted, practical interest is growing in the TDC method, and the results given in this paper will help in the task of organizing the data in the warehouse so as to maximize the effects of compression.

The problem of estimating the effectiveness of compression using TDC reduces to the problem of estimating the sum of the logarithms of the spacings between elements of samples taken without replacement. This is a non-trivial problem, but for the purpose of estimating compression efficiency, we may consider the problem effectively solved using the techniques we have developed. In particular, the approach we develop in Section 3 to sampling without replacement in terms of sampling with replacement is likely to be useful beyond its applications in this paper.

This paper provides methods for estimating the compression for cases where the population from which database records are sampled is either uniform, Zipf, or the product of a uniform distribution and an arbitrary distribution. We have verified our theoretical results by conducting experiments, and agreement between theory and practice is always within a few percent, and to within a fraction of a percent in most cases.

The issue most in need of additional work is that of optimal ordering of attribute domains for achieving optimal compression. We have made significant progress on the issue in this paper, but do not yet have strong analytical results. This is material for further work. Also, our analysis in this paper assumes knowledge of data distributions, but in practice, this information is not always available. Much more likely is non-parametric knowledge of data characteristics, such as variance, skew, or information such as "80% of data is formed from 20% of the values". The estimation of compression efficiency from such non-parametric information is an important area of future work.

From the probability theory and statistics viewpoint, it appears quite important to derive the asymptotic distributions for discrete spacings under proper scaling. The results available to date require the strong assumption that the distribution functions are absolute continuous.

References

[Blumenthal, 1968] Blumenthal, S. (1968). Logarithms of sample spacings. *SIAM Journal on Applied Mathematics*, 16(6):1184–1191.

[Chow and Teicher, 1988] Chow, Y. S. and Teicher, H. (1988). *Probability Theory.* Springer Verlag, New York.

[Csörgő and Wu, 1998] Csörgő, S. and Wu, W. B. (1998). Random graphs and the strong convergence of bootstrap means. *Combinatorics, Probability and Computing (submitted).*

[Darling, 1953] Darling, D. A. (1953). On a class of problems relating to the random division of an interrval. *Annals of Mathematical Statistics*, 24:239–253.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthuruswamy, R. (1996). *Advances in Knowldege Discovery and Data Mining.* MIT Press.

[Kimball, 1996] Kimball, R. (1996). *The Data Warehouse Toolkit: Practical Techniques for Building Dimesional Data Warehouses*. John Wiley & Sons, New York.

[Kolchin et al., 1978] Kolchin, V. F., Sevast'yanov, B. A., and Chistyakov, V. P. (1978). *Random Allocations*. Wiley, New York.

[Li, 1992] Li, W. (1992). Random texts exhibit Zipf-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842–1845.

[Ng and Ravishankar, 1997] Ng, W.-K. and Ravishankar, C. V. (1997). Block-oriented compression techniques for large statistical databases. *IEEE Transactions on Knowledge and Data Engineering*, 9(2):314–328.

[Pyke, 1965] Pyke, R. (1965). Spacings. *Journal of the Royal Statistical Society, Series B*, 27:395–449.

[Rissanen and Langdon, 1981] Rissanen, J. and Langdon, G. G. (1981). Universal Modeling and Coding. *IEEE Transactions on Information Theory*, 27(1):12–23.

[Shao and Marjorie, 1995] Shao, Y. and Marjorie, G. (1995). Limit theorems for the logarithm of sample spacings. *Statistics & probability Letters*, 24(2):121–132.

[Shiryaev, 1995] Shiryaev, A. N. (1995). *Probability*. Springer Verlag, New York.

[van der Vaart., 1998] van der Vaart., A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[Williams, 1991] Williams, R. N. (1991). *Adaptive Data Compression*. Kluwer Academic, Boston.