The Performance of Difference Coding for Sets and Relational Tables

WEI BIAO WU

University of Chicago, Chicago, Illinois

AND

CHINYA V. RAVISHANKAR

University of California—Riverside, Riverside, California

Abstract. We characterize the performance of difference coding for compressing sets and database relations through an analysis of the problem of estimating the number of bits needed for storing the spacings between values in sets of integers. We provide analytical expressions for estimating the effectiveness of difference coding when the elements of the sets or the attribute fields in database tuples are drawn from the uniform and Zipf distributions. We also examine the case where a uniformly distributed domain is combined with a Zipf distribution, and with an arbitrary distribution. We present limit theorems for most cases, and probabilistic convergence results in other cases. We also examine the effects of attribute domain reordering on the compression ratio. Our simulations show excellent agreement with theory.

Categories and Subject Descriptors: E.4 [**Coding and Information Theory**]: *data compaction and compression*; G.3 [**Probability and Statistics**]: *distribution functions*; H.2.4 [**Database Management**]: Systems—*relational databases*

General Terms: Algorithms, Performance, Theory

Additional Key Words and Phrases: Statistical limit theorems, lossless compression, data warehousing, on-line analytical processing

1. *Introduction*

Difference coding, a widely used method [Netravali and Haskell 1988; Viterbi and Omura 1979], represents a series of values by the differences between them. In most applications, successive values in the original sequence are correlated, so that their differences are small, yielding compression. The coding is lossless if the differences are preserved exactly, but this technique is sometimes combined with quantization to convert a continuous range of difference values to a discrete range

Authors' addresses: W. B. Wu, Department of Statistics, University of Chicago, 5734 South University Avenue, Chicago, IL 60637, e-mail: wbwu@galton.uchicago.edu; C. V. Ravishankar, Computer Science & Engineering Department, University of California—Riverside, Riverside, CA 92521, e-mail: ravi@cs.ucr.edu.

Journal of the ACM, Vol. 50, No. 5, September 2003, pp. 665–693.

of output values. Quantization results in lossy compression, which is acceptable in applications like image or voice coding [Netravali and Haskell 1988].

Lossy compression is generally not appropriate for compressing relational databases, but lossless difference coding has been used in this domain [Ng and Ravishankar 1997]. Compression in databases has significant practical advantages. Commercial databases can be very large, and data warehouse sizes can easily reach 10^{12}–10^{13} bytes or more. Disk I/O tends to be the bottleneck to query performance, since queries of interest frequently request statistics formed over a large number of records from the database. Database compression reduces both storage requirements as well as the data transfer volumes between disk and main memory. Compression reduces I/O bandwidth requirements at the cost of higher processor loads. This is a desirable trade-off, since processor performance is improving much faster than disk performance.

Difference coding is particularly applicable to sets and databases, since the usual requirement that the ordering between elements be preserved no longer applies. We are at liberty to choose the ordering between the elements of a set that maximizes the compression. The Tuple-Difference Coding (TDC) method [Ng and Ravishankar 1997] for database compression is based on this insight. Each tuple in a database table is first treated as an integer, and the table sorted by rows. Successive tuples are then differenced, and the differences used to represent the table. The ith tuple in the table can be reconstructed from the first tuple and the first $(i - 1)$ difference values. In practice, to avoid such laborious reconstruction, the sorted table is partitioned into disk-block sized chunks and TDC applied separately to each chunk. This approach has been shown to perform well in practice when used for compressing relational databases.

In this article, we evaluate the performance of difference coding applied to the more general context of sets of integers. Our results will apply directly to relational databases, since a relational table is a set of tuples, and can be modeled as a sample of integers from a large but finite set. In the remainder of the article, we will therefore treat the terms "database" and "set" as equivalent.

2. *Tuple-Difference Coding*

The original work describing TDC [Ng and Ravishankar 1997] discusses practical details such as how to handle textual attributes in TDC, and provides 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. A number of factors affecting compression are listed and their effects discussed in qualitative terms in [Ng and Ravishankar 1997].

Our present purpose is to provide a sounder theoretical basis for the performance characteristics of difference coding applied to sets and relational tables.

2.1. A MODEL FOR PERFORMANCE ANALYSIS. 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[1]

[1] We will sometimes specify the domain as $\{1, 2, \ldots, |D_i|\}$ instead. However, this change of notation will not alter the semantics of compression, since TDC is based on the spacings between successive samples.

if $\mathcal{R} = D_1 \times D_2 \times \cdots \times D_r$. We call $\mathcal{D}$ a database with schema $\mathcal{R}$ if $\mathcal{D} \subseteq \mathcal{R}$. Each record in the database is an r-tuple $\langle d_1, d_2, \ldots, d_r \rangle$, with $d_i \in D_i$.

TDC works as follows: 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 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.

2.1.1. *Some Issues.* 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 between the domains. 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 a database table is a set of tuples, it will not contain duplicate tuples, so this sample must be taken without replacement from the joint distribution space. 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.

We can apply TDC as follows: By sorting on $\varphi(X_i)$ (see Eq. (1)), we can construct the sorted database $X_{(1)} <_\varphi \cdots <_\varphi X_{(n)}$, and by differencing them, the corresponding set of $n - 1$ tuple spacings $\{\delta_k = X_{(k+1)} - X_{(k)}\}$. Thus, estimating the size characteristics of the compressed database is equivalent to estimating $\sum_{k=1}^{n-1} \lceil \log_2(\delta_k + 1) \rceil$.

For convenience, we use natural logarithms rather than base-2 logarithms, and form the statistic $\Lambda_{\mathcal{F}} = \sum_{k=1}^{n-1} \ln(X_{(k+1)} - X_{(k)} + 1)$. In practice, $\Lambda_{\mathcal{F}}$ represents a lower bound on the size of the database, since additional real-world overhead is involved in coding and representing these values. However, such overheads are trivially estimated. Our chief challenge in the remainder of the article will be to estimate $\Lambda_{\mathcal{F}}$ for different attribute spaces $\mathcal{F}$.

While the ordering of attribute domains D_i is irrelevant to semantics, some ordering must be chosen for storing the tuples on disk. Lexicographic sorting is used to order the tuples, but 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.

2.2. SOME USEFUL RESULTS FROM PROBABILITY THEORY. In our development, we find the following well-known results [Chow and Teicher 1988; Shiryayev 1995] from probability theory useful. We state them without proof.

THEOREM 2.1 (SLUTSKY'S THEOREM). *If the sequences* $\{X_{1k}\}, \{X_{2k}\}, \ldots, \{X_{rk}\}, k = 1, 2, \ldots$ *of random variables (r is fixed) are stochastically convergent to the constants* $a_1, a_2, \ldots, a_r$*, then an arbitrary rational function* $R(X_{1k}, X_{2k}, \ldots, X_{rk})$ *converges to the constant* $R(a_1, a_2, \ldots, a_r)$*, provided this constant is finite.*

THEOREM 2.2 (LEBESGUE'S DOMINATED CONVERGENCE THEOREM). *Let* $\eta, \xi, \xi_1, \xi_2, \ldots$ *be random variables such that* $|\xi_n| \le \eta$*,* $E\eta < \infty$*, and* $\xi_n \to \xi$ $(a.s.)$*. Then* $E|\xi| < \infty$*,* $E\xi_n \to E\xi$*, and* $E|\xi_n - \xi| \to 0$ *as* $n \to \infty$*.*

THEOREM 2.3 (LÉVY'S CONTINUITY THEOREM). *Let* $f_n(t) = E \exp(\sqrt{-1}t\xi_n)$*,* $t \in \mathbb{R}$ *be the characteristic functions of* $\xi_1, \xi_2, \ldots,$ *If* $\xi_n \xrightarrow{\mathcal{D}} \xi$*, then* $f_n(t) \to f(t)$ *uniformly in* $|t| \le T$ *for all* $T > 0$*, where* $f(t)$ *is the characteristic function of* ξ*. Conversely, if* $f_n(t)$ *converges to a limit* $f(t)$ *on* $(-\infty, \infty)$ *which is continuous at* $t = 0$*, then* $f(t)$ *is a characteristic function of some random variable* ξ *and* $\xi_n \xrightarrow{\mathcal{D}} \xi$*.*

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 the database is a set. 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 $X_{(1)} < \cdots < X_{(n)}$ 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 Eq. (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

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

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

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

proving Eq. (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 4.7 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 4.7 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

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

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 [2000] for the case where Z has uniform distribution, using large deviation techniques.

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

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

LEMMA 3.1. *Let I_i, $1 \le i \le N$ be independent Bernoulli random variables with $q_i = \Pr[I_i = 1] = 1 - \exp(-n'p(i))$. Then $\Pr[|\sum_{i=1}^{N}(I_i - q_i)| > n'\varepsilon] \le 2\exp(-n'\varepsilon^2/3)$ holds for all $0 < \varepsilon \le 1/10$.*

PROOF. Let $t > 0$. Denote $S_N = \sum_{i=1}^{N}(I_i - q_i)$. Then, by Markov's inequality,

$$\begin{aligned}\log \Pr[S_N > n'\varepsilon] &\le \log[\exp(-n'\varepsilon t)E(e^{S_N})]\\ &= -n'\varepsilon t + \sum_{i=1}^{N} \log[\exp(-tq_i)(1-q_i) + \exp(t(1-q_i))\,q_i].\end{aligned}$$

Elementary manipulations show that $\log[\exp(-tq)(1-q) + \exp(t(1-q))\,q] \le t^2(1.1q - q^2)/2$ holds for all $0 < t < 1/10$ and $0 < q < 1$. Let sets $\mathcal{I} = \{i : 1 \le i \le N,\ n'p(i) \ge 1\}$ and $\mathcal{J} = \{1, \dots, N\} - \mathcal{I}$. Then $|\mathcal{I}| \le n'$ since $\sum_{i=1}^{N} p(i) = 1$. If $i \in \mathcal{J}$, then $n'p(i) < 1$ and hence $q_i \le n'p(i)$ since $1 - \exp(-t) \le t$ for $0 \le t \le 1$. Clearly, for all q, $1.1q - q^2 \le 0.55^2$. So

$$\begin{aligned}\sum_{i=1}^{N}\left(1.1q_i - q_i^2\right) &= \sum_{i\in\mathcal{I}} + \sum_{i\in\mathcal{J}} \le \sum_{i\in\mathcal{I}} 0.55^2 + \sum_{i\in\mathcal{J}} 1.1n'p(i)\\ &\le 0.55^2|\mathcal{I}| + 1.1n' \le 1.4025n'.\end{aligned}$$

Hence, $\log \Pr[S_N > n'\varepsilon] \le -n'\varepsilon t + t^2 1.4025n'/2 \le -n'\varepsilon^2/3$ by letting $t = \varepsilon/1.4025$. The other case $\Pr[S_N < -n'\varepsilon]$ can be handled similarly. □

Remark 3.2. The constants in the upper bound in Lemma 3.1 are not best possible, but it suffices for our application. The classical Hoeffding's inequality [Hoeffding 1963] can yield a bound of order $\exp[-2(n'\varepsilon)^2/N]$, which appears to be too rough in our context since N is typically much larger than n'.

Let $M = \sum_{i=1}^{N} \Pr[V_i > 0] = \sum_{i=1}^{N}[1 - \exp(-n'p(i))]$. Then, applying Lemma 3.1 to $I_i = 1(Y_i > 0)$, we have

$$\Pr\left[\left|\sum_{i=1}^{N} 1(Y_i > 0) - M\right| \ge n'\varepsilon\right]$$

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

which vanishes to 0 at a geometric rate. Hence

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

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 $Z_{(1)} \leq \cdots \leq Z_{(n')}$ 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 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 article, 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 $X_{(1)} < \cdots < X_{(n)}$, 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 $\Lambda_U = \sum_{k=1}^{n-1} \ln(X_{(k+1)} - X_{(k)} + 1)$ 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], and Shao and Marjorie [1995]. Pyke [1965] reviews the literature in this area. 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 4.1. *Let $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$ be the order statistics of i.i.d. uniform(0,1) random variables $U_1, \ldots, U_n$. Then, if $\gamma = 0.5771\cdots$ is Euler's constant,*

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

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 nonzero probability.

This difficulty can be overcomed by changing $\ln(X_{(k+1)} - X_{(k)})$ to $\ln(X_{(k+1)} - X_{(k)} + 1)$. We develop Theorems 4.3 and 4.4 for sampling without or with replacement, respectively.

At first sight, it seems feasible to apply Darling's Theorem to our situation by simply substituting the discrete random variables $X_i = \lfloor NU_i \rfloor$, $1 \le i \le n$ for the continuous random variables NU_i, $1 \le i \le n$. However, this straightforward substitution becomes problematic since the errors will be large if we replace the spacing term $\ln(X_{(k+1)} - X_{(k)} + 1)$ in Theorems 4.3 and 4.4 by the continuous version $\ln(NU_{(k+1)} - NU_{(k)} + 1)$ unless $NU_{(k+1)} - NU_{(k)}$, $1 \le k \le n-1$ 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 4.3 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 proof of Theorem 4.3, 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 4.1.

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 4.3 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 4.3, 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 $X'_{(1)} \leq \cdots \leq X'_{(n)}$, or equivalently, $\lfloor NU_{(1)} \rfloor \leq \lfloor NU_{(2)} \rfloor \leq \cdots \leq \lfloor NU_{(n)} \rfloor$ using the representation

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

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

We use this representation form throughout the article. 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 4.2. *Let* $\{Y, Y_i, i \geq 1\}$ *be i.i.d.* exp(1) *random variables. Then*

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

PROOF. Clearly $n \Pr[1/Y > n \ln n] = n[1 - \exp(-1/(n \ln n))] \to 0$ as $n \to \infty$. Then

$$\frac{1}{n \ln n} \left(\sum_{i=1}^{n} \frac{1}{Y_i} - nE[Y^{-1} \mathbf{1}_{Y^{-1} \leq n \ln n}] \right) \xrightarrow{\mathcal{P}} 0$$

by Klass and Teicher [1977]. Thus, the lemma follows since

$$\lim_{\epsilon \to 0} \frac{E[Y^{-1} \mathbf{1}_{Y \geq \epsilon}]}{-\ln \epsilon} = \lim_{\epsilon \to 0} \frac{\int_{\epsilon}^{\infty} y^{-1} \exp(-y)\, dy}{-\ln \epsilon} = \lim_{\epsilon \to 0} \frac{-\epsilon^{-1} \exp(\epsilon)}{-\epsilon^{-1}} = 1$$

by the L'Hospital rule. □

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 $n^2 = o(N_n)$.[1] In this case, we are able to obtain

[1] It is usual to write $f(n) = o(g(n))$ when f and g are functions of an independent variable t such that $\lim_{n \to \infty} f(n)/g(n) = 0$. So we will frequently write $n = o(N_n)$ to emphasize our view of N as a function of the independent variable n and $\lim_{n \to \infty} n/N_n = 0$.

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 4.3. *Let $(X_1, \dots, X_n)$ be n numbers sampled from the set $\{0, 1, \dots, N_n - 1\}$ equiprobably, and without replacement. Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics of $(X_1, \dots, X_n)$, and define the random variable $\Lambda_U = \sum_{k=1}^{n-1} \ln(X_{(k+1)} - X_{(k)} + 1)$. 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 $\gamma = -E(\ln Y) = 0.57721\dots$, or Euler's constant, and $\alpha^2 = Var(\ln Y - Y) = \pi^2/6 - 1 = 0.644934\dots$

PROOF. We write $N = N_n, \sigma = \sigma_U, \mu = \mu_U$ for simplicity. Let $X'_1, \dots, X'_n$ be i.i.d. random variables with common distribution $\Pr[X'_1 = k] = 1/N$ for $k = 0, 1, \dots, N-1$. First, we claim the following distributional equality, which will transform the dependent random variables $(X_1, \dots, X_n)$ to i.i.d. random variables $(X'_1, \dots, X'_n)$:

$$(X_1, \dots, X_n) \overset{\mathcal{D}}{=} (X'_1, \dots, X'_n | X'_1, \dots, X'_n \text{ are different}). \tag{6}$$

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

$$\begin{aligned}&\Pr[X_1 = x_1, \dots, X_n = x_n] = 0\\ &\quad = \Pr[X'_1 = x_1, \dots, X'_n = x_n | X'_1, \dots, X'_n \text{ are different}].\end{aligned}$$

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

$$\begin{aligned}&\Pr[X'_1 = x_1, \dots, X'_n = x_n | X'_1, \dots, X'_n \text{ are different}]\\ &\quad = \Pr[X'_1 = x_1, \dots, X'_n = x_n] / \Pr[X'_1, \dots, X'_n \text{ are different}]\\ &\quad = \left(\frac{1}{N}\right)^n \left\{\prod_{j=1}^{n-1}\left(1 - \frac{j}{N}\right)\right\}^{-1} = \binom{N}{n}^{-1}\\ &\quad = \Pr[X_1 = x_1, \dots, X_n = x_n].\end{aligned}$$

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

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

where $X'_{(1)} \leq \cdots \leq X'_{(n)}$ is the order statistics of $(X'_1, \dots, X'_n)$, 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

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

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

$$\left(X'_{(1)}, \ldots, X'_{(n)}\right) \overset{\mathcal{D}}{=} \left(\left\lfloor NU_{(1)} \right\rfloor, \ldots, \left\lfloor NU_{(n)} \right\rfloor\right). \tag{7}$$

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 Eq. (5), and let event

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

From Eq. (7), $\Pr[B_n] = \Pr[\hat{B}_n]$. 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:

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

Event D_n^c, the complement of event D_n, leads us to the approximation $S_{n+1} \approx n$. If event $C_n D_n^c$ occurs, then we can show by a straightforward approach that $\ln(\lfloor NS_{k+1}/S_{n+1} \rfloor - \lfloor NS_k/S_{n+1} \rfloor + 1)$ can be approximated by $\ln(NY_{k+1}/S_{n+1})$.

Hence, we really need to show that $\Pr[C_n D_n^c] = 1 + o(1)$, 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 $\Pr[C_n] \ge \Pr[C_n D_n^c] \ge \Pr[NY_2 > 3n, \ldots, NY_n > 3n, D_n^c] \ge (\exp(-3n/N))^{n-1} - \Pr[D_n] = 1 + o(1)$.

Let $\{x\}$ denote the fractional part of x (i.e., $\{x\} = x - \lfloor x \rfloor$), and let $\varepsilon_{nk} = \{NS_k/S_{n+1}\} - \{NS_{k+1}/S_{n+1}\} + 1 \in (0, 2)$. 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

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

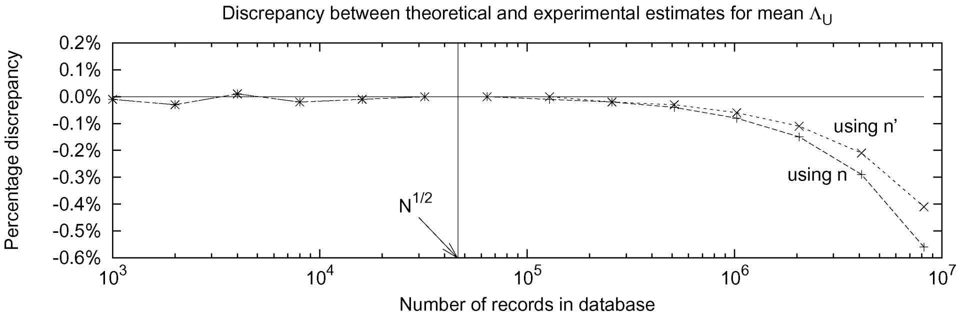

FIG. 1. Uniform distribution: Agreement between Theorem 4.3 and experiment.

$$\begin{aligned}
&\le \sum_{k=1}^{n-1} \frac{2S_{n+1}}{NY_{k+1}} + 2\left(\frac{\ln n}{n}\right)^{1/2} + \ln n \\
&\le \frac{3n}{N} \sum_{k=1}^{n-1} \frac{1}{Y_{k+1}} + 2\ln n
\end{aligned}$$

By Lemma 4.2, notice that $\Pr[C_n D_n^c] = 1 + o(1)$ and $n^2 = o(N_n)$, we obtain

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

which leads to Theorem 4.3 via Slutsky's Theorem and the classical central limit theorem $[(n-1)^{1/2}\alpha]^{-1} \sum_{k=1}^{n-1} (\ln Y_{k+1} + \gamma - Y_{k+1} + 1) \xrightarrow{\mathcal{D}} N(0,1)$. The exact value of the asymptotic variance α^2 is presented in Corollary 4.6 below. □

Figure 1 compares the estimates of Λ_U from Theorem 4.3 with the results of experiments on databases of different sizes containing integers drawn uniformly without replacement from $\{1, 2, \ldots, 2^{31} - 1\}$. Theory and experiment agree to within a fraction of one percent even for databases as large as $2 \cdot 10^6$ (showing the robustness of the theorem, since $\sqrt{N} \approx 46,000$ in this case).

The major idea in the proof of Theorem 4.3 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 4.4. *Assume that $n^2 = o(N_n)$. Let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ 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,*

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

where μ_U, σ_U *are the same as that in Theorem* 4.3.

Remark 4.5. Theorems 4.3 and 4.4 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 $\alpha^2 = \mathit{Var}(\ln Y - Y)$ directly, but we note that Theorems 4.1 and 4.4 jointly lead to the following interesting observation.

COROLLARY 4.6. *If* Y *is* $\exp(1)$ *distributed, then* $\alpha^2 = \mathit{Var}(\ln Y - Y) = \pi^2/6 - 1 = 0.644934\ldots$.

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 Eq. 3. Under the assumption $n < N/2$, we claim that

$$\left| E[J_n] - N \ln\left(1 - \frac{n}{N}\right) \right| < \frac{n+2}{N} \tag{8}$$

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

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

Therefore

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

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 8 means that, on average, we have to draw $n' = -N \ln(1-n/N) \approx E[J_n]$ i.i.d. samples uniformly from $\{1, 2, \ldots, N\}$ to get n distinct values. Therefore, in applying Theorem 4.4 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 4.3.

It is instructive to examine the applicability of Eq. (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 8.

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 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 4.7 below and Remark 4.2 give approximations of Λ_Z based on a sampling scheme *with* replacement; that is, 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 4.7 to obtain reasonable estimates for Λ_Z.

A problem is that $E[J_n]$ can be calculated directly from Eq. (3) only for very special cases; the only really tractable case may well be the uniform distribution. We must therefore solve for n' from Eq. (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 $t \in [1, N]$,

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

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

$$\frac{nH_N}{n'} - \ln\frac{NH_N}{n'} = 1 - \gamma. \tag{9}$$

Although an explicit formula for the root n' of Eq. (9) does not exist, we can use the fixed-point iteration scheme

$$f_{k+1} = f(f_k),\ f_1 = f(1),\ f(t) = \frac{nH_N}{1 - \gamma + \ln(NH_N) - \ln(t)},\ k \in \mathbb{N}. \tag{10}$$

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 4.7, 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 $X'_{(1)} \leq \cdots \leq X'_{(n)}$ being the corresponding order statistics. 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 $(N) \overset{\mathcal{D}}{=} Q_N(U)$. 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

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

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 4.7. *Let $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$ 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 $X_{(1)} < X_{(2)} < \cdots < X_{(k_0)}$ are very likely to be $1, 2, \ldots, k_0$. We take $k_0 = \lfloor n\rho_n \rfloor$ here. This observation suggests that $\Lambda'_U = \sum_{k=1}^{k_0-1} \ln(X_{(k+1)} - X_{(k)} + 1)$ 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 $\Delta'_U/(n \ln N) \xrightarrow{\mathcal{P}} 0$. Since the logarithm function is concave, we may apply Jensen's inequality to get

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

Now for any $\varepsilon > 0$,

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

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

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

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

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

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,

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

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 $0 \ge \ln(1 - N^{-Y_{k+1}/S_{n+1}} + N^{-S_{k+1}/S_{n+1}}) \ge \ln(1 - N^{-Y_{k+1}/S_{n+1}})$,

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

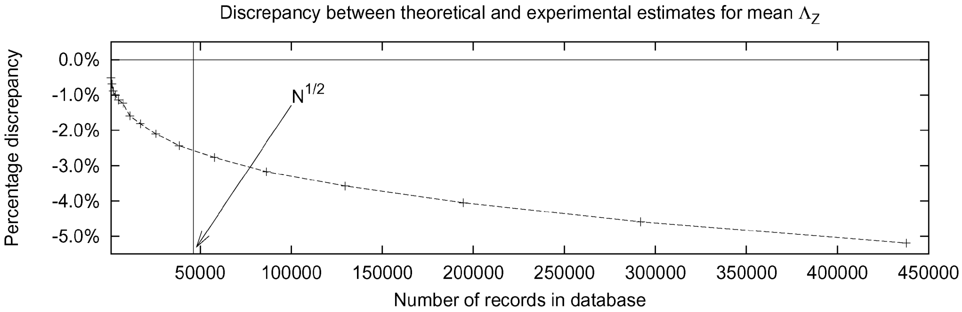

FIG. 2. Zipf distribution: Agreement between Theorem 4.7 and experiment.

Observe that if Y_{k+1} is exp(1), then $1 - \exp(-Y_{k+1})$ is uniform(0,1), $E \ln(1 - \exp(-Y_{k+1})) = -1$, hence by Markov's inequality, the first term $\leq -(\varepsilon \ln N)^{-1} E \ln(1 - \exp(-Y_{k+1})) = (\varepsilon \ln N)^{-1}$. Obviously, the second term goes 0 via the Weak Law of Large Numbers, which completes the proof of the first statement of Theorem 4.7. 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 4.8. Under the conditions of Theorem 4.7, 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 4.8 matches the results of experiments on databases of different sizes containing integers drawn without replacement from a Zipf distribution over $\{1, 2, \ldots, 2^{31} - 1\}$. 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.

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

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

and

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

PROOF. As in the proof in Theorem 4.7, we write

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

Using the same argument as in Theorem 4.7, we have $(n \ln N)^{-1} I_n - 1/2 \xrightarrow{\mathcal{P}} 0$. For any $\varepsilon > 0$, let $n > n_0$ be large enough such that $(\ln N)^{-1}(n+1) < 1/2$, then

$$\begin{aligned}&\Pr\biggl[\frac{1}{n \ln N}\bigl|J_n\bigr| > \varepsilon\biggr]\\ &\le \Pr\Biggl[\frac{1}{n \ln N}\sum_{k=1}^{n-1} \ln\bigl(1 - N^{-Y_{k+1}/S_{n+1}}\bigr) < -\varepsilon, \frac{S_{n+1}}{n+1} \le 2\Biggr] + \Pr\biggl[\frac{S_{n+1}}{n+1} \ge 2\biggr]\\ &\le \Pr\Biggl[\frac{1}{n \ln N}\sum_{k=1}^{n-1} \ln(1 - \exp(-Y_{k+1})) < -\varepsilon\Biggr] + \Pr\biggl[\frac{S_{n+1}}{n+1} \ge 2\biggr].\end{aligned}$$

Again by the same arguments as in Theorem 4.7, 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 $\{(n \ln N_n)^{-1}\Delta_U - 1/2, n \ge 2\}$. □

Remark 4.10. Under the conditions of Theorem 4.9, we have $\lim_{n\to\infty} \ln n/\ln N_n = 0$, thus $\rho_n \approx 0$. Then interestingly enough, both Theorem 4.9 and Theorem 4.7 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

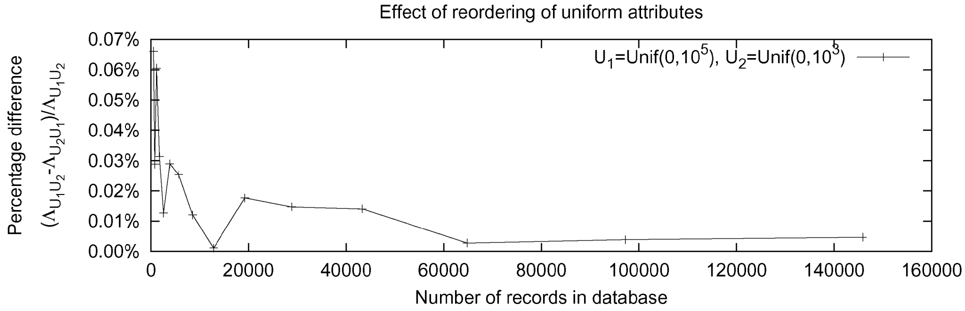

FIG. 3. Two uniform attributes: Effect of attribute reordering on $(\Lambda_{U_1U_2} - \Lambda_{U_2U_1})/\Lambda_{U_1U_2}$.

$Z_{(2)} - Z_{(1)}, \ldots, Z_{(n)} - Z_{(n-1)}$ 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 $X_1, X_2, \ldots, X_k$ 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. NONUNIFORM 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.

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 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 $Z_{(1)} \leq Z_{(2)}$. Now,

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

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

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

As before,

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

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

Now let us assume $n > m$. From the formulas 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 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.

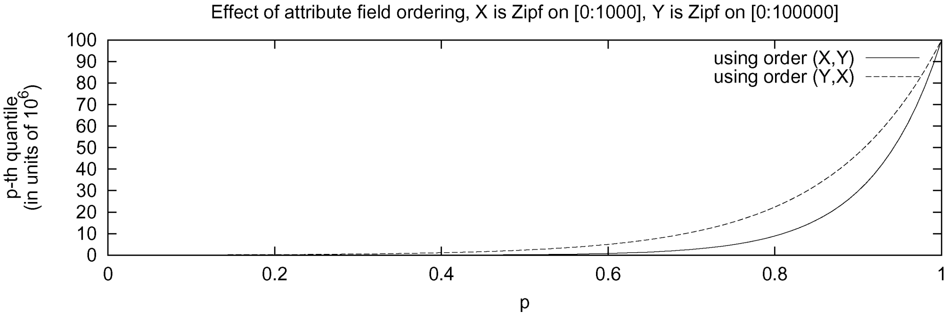

FIG. 4. Two Zipf attributes: Characterizing skew through the distribution of quantiles.

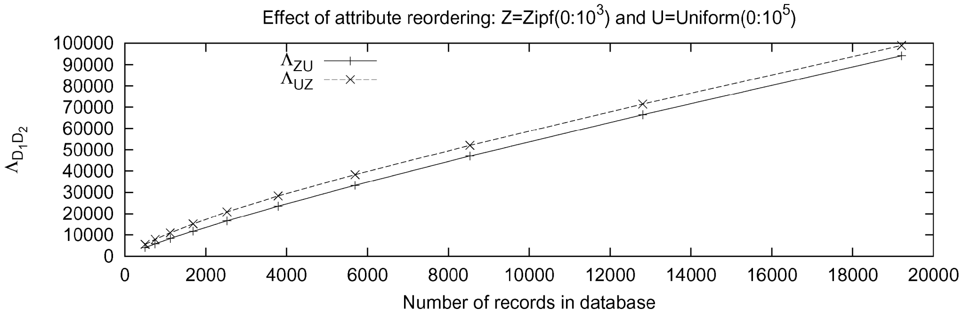

FIG. 5. Attribute reordering: One Zipf and one uniform attribute.

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 nonuniform 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 the optimal ordering. We note however, that for two Zipf random variables, the advantage of optimal ordering over an arbitrary ordering seems small since $E\Lambda_{xy} - E\Lambda_{xy} \approx 1/3 \ln(n/m)$, 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 $Z_{(1)} < Z_{(2)} < \cdots < Z_{(d)}$ 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

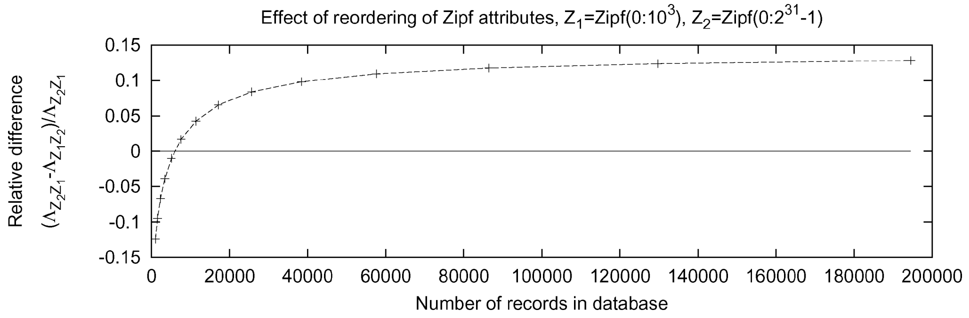

FIG. 6. Attribute reordering: two Zipf attributes.

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, nonuniformity within the set of spacings is really an issue. Such nonuniformity 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).

6. *Limit Theorems for the Multifield 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 nonuniform domains will be placed in some suitable order at the head of the tuple. We may view these nonuniformly distributed domains as jointly constituting a single composite domain with an arbitrary discrete distribution.

Let us therefore model the nonuniform domains $X_1, \ldots, X_{m-1}$ 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 $X_1, \ldots, X_k$ into just two fields. Let Z be an arbitrary discrete random variable assuming values from $\{1, 2, \ldots, d\}$, and U be uniform on $\{1, 2, \ldots, u_n\}$. 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 $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ 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 $\{1, 2, \ldots, d\}$, then (Z, U) can be viewed as single large uniformly distributed field. Theorem 4.3 can be directly applied to obtain the following result.

COROLLARY 6.1. *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,

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

In this formula, $\Lambda_j = \sum_{i=1}^{N_j - 1} \ln(U_{j,i+1} - U_{j,i} + 1)$, or given $N_j = l$, $\Lambda_j = \Lambda_j(l) = \sum_{i=1}^{l-1} \ln(U_{j,i+1} - U_{j,i} + 1)$, $U_{j,1} \leq U_{j,2} \leq \cdots U_{j,l}$ is the order statistics of $U_{j,1}, \ldots, U_{j,l}$, where $\{U_{j,i}, 1 \leq j \leq d, i \geq 0\}$, are i.i.d. random variables uniform on $\{1, \ldots, u_n\}$ 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 Shiryayev [1995], we may therefore write the inequality $\Pr[|N_j/n - p_j| \geq \varepsilon] \leq 2\exp(-2n\varepsilon^2)$ 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 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 6.2, we may assume without undue concern that $N_j \geq 2$.

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

THEOREM 6.2. *Let each record in the database comprise two discrete random fields (Z, U), where Z is an arbitrary distribution on $\{1, \dots, d\}$, and U is uniform on $\{1, \dots, u_n\}$. If $n^2 = o(u_n)$, then as $n \to \infty$,*

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

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

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

PROOF. We first motivate the result with heuristics before proceeding to the rigorous argument. 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 4.3, 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 let us proceed the rigid argument. For $k \ge 1$, set $\mu(k) = (k - 1)(-\gamma + \ln u_n - \ln k)$, and let $\mu(k) = 0$ if $k \le 1$. To obtain the limiting distribution, we apply the Lévy Continuity Theorem by analyzing characteristic functions. We first note that given $N_1 = n_1, \dots, N_d = n_d$, $\Lambda_1, \dots, \Lambda_d$ are independent. Hence, for $t \in \mathbb{R}$, we have via conditioning,

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

We next assert the three convergence results (11), (12), and (13) and proceed to prove them using the Lebesgue Dominated Convergence Theorem and Slutsky's

Theorem. These results will lead to Theorem 6.2 via the Lévy Continuity Theorem.

$$E\left[\exp\left(\sqrt{-1}t\frac{\Lambda_j-\mu(N_j)}{\sqrt{n}}\right)\middle|N_j\right]\longrightarrow\exp\left(\frac{-\alpha^2p_jt^2}{2}\right)\text{ a.s. },\qquad(11)$$

$$\sum_{j=1}^{d}\frac{\mu(N_j)-\mu(np_j)}{\sqrt{n}}\xrightarrow{\mathcal{D}}N\left(0,\sum_{j=1}^{d}p_j(\ln p_j)^2-\left(\sum_{j=1}^{d}p_j\ln p_j\right)^2\right),\qquad(12)$$

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

$$\frac{1}{n^{1/2}}\left[\ln(U_{j+1,1}-U_{j,N_j}+u_n+1)-\ln\left(\frac{u_n}{np_j}+\frac{u_n}{np_{j+1}}\right)\right]\xrightarrow{\mathcal{P}}0.\qquad(13)$$

To show (11), we proceed as follows: Since event $\{N_j=l\}$ and $\Lambda_j(l)$ are independent,

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

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

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

for (11), we only need to show $A=0$, $B=0$ a.s. Again, by inequality $\Pr[|N_j/n-p_j|\ge\varepsilon]\le2\exp(-2n\varepsilon^2)$, $\lim_{k\to\infty}\sum_{n=k}^{\infty}\Pr\left[N_j\notin\mathcal{I}_n\right]\le2\lim_{k\to\infty}\sum_{n=k}^{\infty}\exp[-2n(n^{-1/3})^2]=0$. Hence $B=0$ a.s. via the Borel–Cantelli lemma. If $A\neq0$, then there exists an $\varepsilon>0$, a subsequence $\{n'\}\subset\mathbb{N}$ and $l(n')\in\mathcal{I}_{n'}$ such that along this subsequence, $|g(t;n',l(n'))-\exp(-\alpha^2p_jt^2/2)|>\varepsilon$. However, by the Lévy Continuity Theorem, we do have $|g(t;n',l(n'))-\exp(-\alpha^2p_jt^2/2)|\to0$ following from $(n')^{-1/2}[\Lambda_j(l(n'))-\mu(l(n'))]\xrightarrow{\mathcal{D}}N(0,\alpha^2p_j)$, which is due to $l(n')/n'\to p_j$ and $[\alpha^2l(n')]^{-1/2}[\Lambda_j(l(n'))-\mu(l(n'))]\xrightarrow{\mathcal{D}}N(0,1)$ asserted by Theorem 4.3 since $l(n')\to\infty$.

To prove (12), 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 $\Sigma_{ii}=p_i(1-p_i)$, $\Sigma_{ij}=-p_ip_j$. Using the Delta method [van der Vaart 1998], which expands a function of random variables about its mean with a 1-step

Taylor expansion to compute its variance, we now have

$$n^{1/2}[\nu(\hat{p}_n) - \nu(\hat{p})] \xrightarrow{\mathcal{D}} N\left(0, \frac{\partial \nu}{\partial \hat{q}}\bigg|_{\hat{p}} \Sigma \left(\frac{\partial \nu}{\partial \hat{q}}\right)^{\tau}\bigg|_{\hat{p}}\right). \tag{14}$$

In view of $n^{1/2}(\hat{p}_n - \hat{p}) \xrightarrow{\mathcal{D}} N(0, \Sigma)$, Taylor's expansion $\nu(\hat{p}_n) - \nu(\hat{p}) \approx (\hat{p}_n - \hat{p})(\partial\nu/\partial\hat{q})^{\tau}|_{\hat{p}}$ gives some intuition of the application of the Delta method for (14). Now we can use (14) to prove (12) 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

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

For (13), we only need to show that

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

The notation $X_n = O_P(1)$, as in van der Vaart [1998], means that the random sequence X_n is stochastically bounded; that is, 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

$$\frac{np_{j+1}}{u_n}U_{j+1,1} \xrightarrow{\mathcal{D}} \exp(1), \; \frac{np_j}{u_n}(u_n - U_{j,N_j}) \xrightarrow{\mathcal{D}} \exp(1). \tag{15}$$

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

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

The last step follows from the Lebesgue Dominated Convergence Theorem. □

Figure 7 shows how closely Theorem 6.2 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 $(0, 10^5)$.

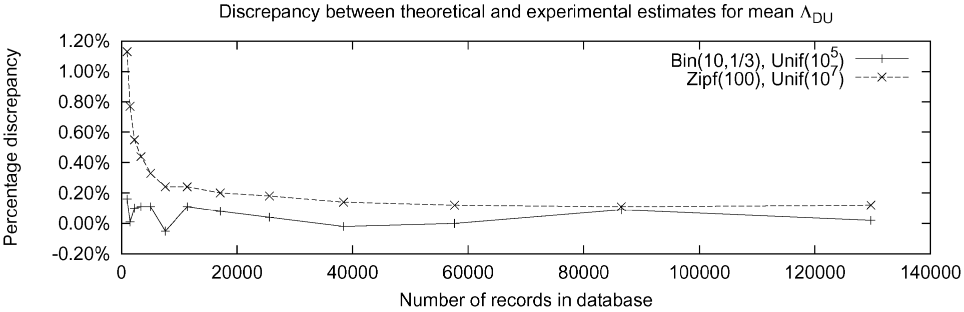

FIG. 7. Multiple fields: Agreement between Theorem 6.2 and experiment.

Agreement in both cases is to within a fraction of one percent over a large range, illustrating the power of Theorem 6.2.

Remark 6.3. 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 (15) 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

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

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,

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

Finally, an elementary but interesting computation leads to

$$\begin{aligned} E\ln\left(\frac{Y_1}{p_{j+1}} + \frac{Y_2}{p_j}\right) &= \int_0^\infty \int_0^\infty \exp(-(s+t))\ln\left(\frac{s}{p_{j+1}} + \frac{t}{p_j}\right) ds\, dt \\ &= \frac{p_j \ln p_{j+1} - p_{j+1}\ln p_j}{p_{j+1} - p_j} - \gamma \end{aligned}$$

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

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

Remark 6.4. 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 article 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 nontrivial 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 article 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 article assumes knowledge of data distributions, but in practice, this information is not always available. Much more likely is nonparametric 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.

ACKNOWLEDGMENTS. The authors are grateful to the reviewers for their reading of this article, and for the suggestions they have made for its improvement.

REFERENCES

BLUMENTHAL, S. 1968. Logarithms of sample spacings. *SIAM J. Appl. Math. 16*, 1184–1191.

CHOW, Y. S., AND TEICHER, H. 1988. *Probability Theory*. Springer-Verlag, New York.

CSÖRGŐ, S., AND WU, W. B. 2000. Random graphs and the strong convergence of bootstrap means. *Combin. Prob. Comput.*, 9, 315–347.

DARLING, D. A. 1953. On a class of problems relating to the random division of an interrval. *Ann. Math. Stat. 24*, 239–253.

HOEFFDING, W. 1963. Probability inequalities for sums of bounded random variables. *J. ASA 58*, 13–30.

KLASS, M., AND TEICHER, H. 1977. Iterated logarithm laws for asymmetric random variables barely with or without finite mean. *Ann. Prob. 5*, 861–874.

KOLCHIN, V. F., SEVAST'YANOV, B. A., AND CHISTYAKOV, V. P. 1978. *Random Allocations*. Wiley, New York.

LI, W. 1992. Random texts exhibit Zipf-law-like word frequency distribution. *IEEE Trans. Inf. Theory 38*, 1842–1845.

NETRAVALI, A. N., AND HASKELL, B. G. 1988. *Digital Pictures—Representation and Compression*. Plenum Press, New York and London.

NG, W.-K., AND RAVISHANKAR, C. V. 1997. Block-oriented compression techniques for large statistical databases. *IEEE Trans. Knowl. Data Eng. 9*, 314–328.

PYKE, R. 1965. Spacings. *J. Roy. Stat. Soc., Ser. B 27*, 395–449.

SHAO, Y., AND MARJORIE, G. 1995. Limit theorems for the logarithm of sample spacings. *Stat. Prob. Lett. 24*, 121–132.

SHIRYAYEV, A. N. 1995. *Probability*. Springer-Verlag, New York.

VAN DER VAART, A. W. 1998. *Asymptotic Statistics*. Cambridge University Press, Cambridge, Mass.

VITERBI, A. J., AND OMURA, J. K. 1979. *Principles of Digital Communication and Coding*. McGraw-Hill, New York.

RECEIVED JANUARY 1999; REVISED APRIL 2002; ACCEPTED MAY 2003