RJ 10040 (90536) August 22, 1996 (Revised 7/16/99)
Computer Science

Research Report

HOEFFDING INEQUALITIES FOR JOIN-SELECTIVITY ESTIMATION AND ONLINE AGGREGATION

Peter J. Haas

IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

Research Division
Yorktown Heights, New York • San Jose, California • Zurich, Switzerland

HOEFFDING INEQUALITIES FOR JOIN-SELECTIVITY ESTIMATION AND ONLINE AGGREGATION

Peter J. Haas

IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099
e-mail: peterh@almaden.ibm.com

ABSTRACT: We extend Hoeffding's inequalities for simple averages of random variables to the case of cross-product averages. We also survey some new and existing Hoeffding inequalities for estimators of the mean, variance, and standard deviation of a subpopulation. These results are applicable to two problems in object-relational database management systems: fixed-precision estimation of the selectivity of a join and online processing of aggregation queries. For the first problem, the new results can be used to modify the asymptotically efficient sampling-based procedures of Haas, Naughton, Seshadri, and Swami so that there is a guaranteed upper bound on the number of sampling steps. For the second problem, the inequalities can be used to develop conservative confidence intervals for online aggregation; such intervals avoid the large intermediate storage requirements and undercoverage problems of intervals based on large-sample theory.

Keywords: Hoeffding inequality, join-selectivity estimation, query optimization, database sampling, online aggregation

1. Introduction and Summary

In many applications, it is necessary to estimate a *population mean*

$$\mu = \frac{1}{m}\sum_{i=1}^{m} v(i)$$

using random sampling; here $m > 1$ is a fixed integer and v is a real-valued function defined on the set $\{1, 2, \ldots, m\}$. Denote by $L_1, L_2, \ldots, L_n$ a random sample drawn uniformly with replacement from the set $\{1, 2, \ldots, m\}$ and by $L_1', L_2', \ldots, L_n'$ a random sample drawn uniformly without replacement. For $n \geq 1$, the estimators

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} v(L_i) \tag{1.1}$$

and

$$\overline{Y}_n' = \frac{1}{n \wedge m}\sum_{i=1}^{n \wedge m} v(L_i') \tag{1.2}$$

are each *unbiased* for μ in that $E[\overline{Y}_n] = E[\overline{Y}_n'] = \mu$. (Here $n \wedge m$ denotes the minimum of n and m.) In a famous paper, Hoeffding [11, Theorem 4] shows that

$$E[f(\overline{Y}_n')] \leq E[f(\overline{Y}_n)] \tag{1.3}$$

for $n \geq 1$ and any convex function f. In particular, it follows by taking $f(x) = x^2 - \mu^2$ in (1.3) that $\mathrm{Var}[\overline{Y}_n'] \leq \mathrm{Var}\,[\overline{Y}_n]$.

It is frequently useful to bound the probability that $\overline{Y}_n$ (resp., $\overline{Y}_n'$) deviates from μ by more than a specified amount. Suppose that the only information available prior to sampling consists of lower and upper bounds a and b, respectively, on the function v:

$$a \leq v(i) \leq b \tag{1.4}$$

for $1 \leq i \leq m$. In [11], Hoeffding not only establishes (1.3), but also shows that

$$P\left\{\, |\overline{Y}_n - \mu| \geq t \,\right\} \leq 2e^{-2nt^2/(b-a)^2} \tag{1.5}$$

and

$$P\left\{\, |\overline{Y}_n' - \mu| \geq t \,\right\} \leq 2e^{-2n't^2/(b-a)^2}. \tag{1.6}$$

for $t > 0$ and $n \geq 1$, where

$$n' = \begin{cases} n & \text{if } n < m; \\ +\infty & \text{if } n \geq m. \end{cases}$$

In the following, we extend Hoeffding's results by allowing v to be a function of more than one variable and establishing analogues of inequalities (1.3), (1.5), and (1.6) for "cross-product averages." We also survey some new and existing Hoeffding inequalities for the case in which we wish to estimate the mean, variance, or standard deviation of the numbers $\{ v(i)\colon\ i \in S \}$, where $S \subseteq \{ 1, 2, \ldots, m \}$. We refer to S as a "subpopulation" and assume that the membership or non-membership in S of an element i is discovered only when i is sampled.

Our results, which are stated formally in the remainder of this section, have a number of applications in object-relational database management systems (ORDBMS's). In Section 2 we consider the problem of using sampling to estimate the selectivity of a join to within a prespecified precision. Haas, Naughton, Seshadri, and Swami [7] have previously provided an asymptotically efficient procedure called *f_p_cross* for fixed-precision estimation of selectivities. We show how the new inequalities can be used to modify *f_p_cross* so that there is a guaranteed upper bound on the number of sampling steps executed by the procedure. In Section 3, we show that our results are pertinent to online aggregation processing as described in Hellerstein, Haas, and Wang [9]. Section 4 contains the proofs of our results.

1.1. Cross-Product Averages

Let $m_1, m_2, \ldots, m_K$ $(K \geq 1)$ be finite positive integers and v be a real-valued function defined on the set $\Lambda = \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_K$, where $\Lambda_k = \{ 1, 2, \ldots, m_k \}$ for $1 \leq k \leq K$. To avoid trivialities we assume throughout that $m_k > 1$ for $1 \leq k \leq K$. Suppose we wish to estimate the population mean

$$\mu = \frac{1}{m_1 m_2 \cdots m_K} \sum_{l_1=1}^{m_1} \sum_{l_2=1}^{m_2} \cdots \sum_{l_K=1}^{m_K} v(l_1, l_2, \ldots, l_K) \tag{1.7}$$

using random sampling. For $1 \leq k \leq K$, denote by $L_{k,1}, L_{k,2}, \ldots, L_{k,n}$ a random sample of size n drawn uniformly with replacement from the set Λ_k. Similarly, denote by $L'_{k,1}, L'_{k,2}, \ldots, L'_{k,n}$ a random sample drawn uniformly without replacement. We assume throughout that the sampling mechanisms for sets $\Lambda_1, \Lambda_2, \ldots, \Lambda_K$ are mutually independent. Set $\mathbf{N} = \{ 1, 2, \ldots \}^K$ and, for $\mathbf{n} = (n_1, n_2, \ldots, n_K) \in \mathbf{N}$, define the *cross-product averages* $\tilde{Y}_{\mathbf{n}}$ and $\tilde{Y}'_{\mathbf{n}}$ by

$$\tilde{Y}_{\mathbf{n}} = \frac{1}{n_1 n_2 \cdots n_K} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_K=1}^{n_K} v(L_{1,i_1}, L_{2,i_2}, \ldots, L_{K,i_K}) \tag{1.8}$$

and

$$\tilde{Y}'_{\mathbf{n}} = \left(\prod_{k=1}^{K} (n_k \wedge m_k) \right)^{-1} \sum_{i_1=1}^{n_1 \wedge m_1} \sum_{i_2=1}^{n_2 \wedge m_2} \cdots \sum_{i_K=1}^{n_K \wedge m_K} v(L'_{1,i_1}, L'_{2,i_2}, \ldots, L'_{K,i_K}), \tag{1.9}$$

respectively. It is straightforward to show that $E[\tilde{Y}_{\mathbf{n}}] = E[\tilde{Y}'_{\mathbf{n}}] = \mu$ for $\mathbf{n} \in N$. Moreover, $\tilde{Y}'_{\mathbf{n}} \equiv \mu$ for all $\mathbf{n} = (n_1, n_2, \ldots, n_K) \in \mathbf{N}$ such that $n_k \geq m_k$ for $1 \leq k \leq K$.

Although $\tilde{Y}_{\mathbf{n}}$ can be viewed as a simple average of $n_1 n_2 \cdots n_K$ random variables (and similarly for $\tilde{Y}'_{\mathbf{n}}$), the inequalities (1.3), (1.5), and (1.6) are not directly applicable. The problem is that the random variables that make up the average are not mutually independent: in general, $v(L_{1,i_1}, L_{2,i_2}, \ldots, L_{K,i_K})$ and $v(L_{1,j_1}, L_{2,j_2}, \ldots, L_{K,j_K})$ are dependent unless $i_k \neq j_k$ for $1 \leq k \leq K$. An analogous remark applies to $\tilde{Y}'_{\mathbf{n}}$.

Our first result extends the inequality in (1.3).

Theorem 1. *Let $\tilde{Y}_{\mathbf{n}}$ and $\tilde{Y}'_{\mathbf{n}}$ be defined as in* (1.8) *and* (1.9)*, respectively. Then*

$$E\left[f(\tilde{Y}'_{\mathbf{n}})\right] \leq E\left[f(\tilde{Y}_{\mathbf{n}})\right] \tag{1.10}$$

for $\mathbf{n} \in \mathbf{N}$ and any convex function f. In particular, $\mathrm{Var}[\tilde{Y}'_{\mathbf{n}}] \leq \mathrm{Var}[\tilde{Y}_{\mathbf{n}}]$.

Since $\tilde{Y}_{\mathbf{n}}$ and $\tilde{Y}'_{\mathbf{n}}$ are each unbiased, it follows from Theorem 1 that $\tilde{Y}'_{\mathbf{n}}$ has a lower mean squared error than $\tilde{Y}_{\mathbf{n}}$.

Our next result, Theorem 2 below, generalizes the inequalities in (1.5) and (1.6) and bounds the probability that $\tilde{Y}_{\mathbf{n}}$ (resp., $\tilde{Y}'_{\mathbf{n}}$) deviates from μ by more than a specified amount. The inequalities require *a priori* knowledge only of lower and upper bounds a and b, respectively, on the function v:

$$a \leq v(l_1, l_2, \ldots, l_K) \leq b \tag{1.11}$$

for $(l_1, l_2, \ldots, l_K) \in \Lambda$. For $\mathbf{n} \in \mathbf{N}$, set

$$m(\mathbf{n}) = \min_{1 \leq k \leq K} n_k$$

and

$$m'(\mathbf{n}) = \min_{1 \leq k \leq K} n'_k,$$

where

$$n'_k = \begin{cases} n_k & \text{if } n_k < m_k; \\ +\infty & \text{if } n_k \geq m_k \end{cases}$$

for $1 \leq k \leq K$.

Theorem 2. *Let $\tilde{Y}_{\mathbf{n}}$ and $\tilde{Y}'_{\mathbf{n}}$ be defined as in* (1.8) *and* (1.9)*, respectively, and let a and b satisfy* (1.11)*. Then*

$$P\left\{\, |\tilde{Y}_{\mathbf{n}} - \mu| \geq t \,\right\} \leq 2e^{-2m(\mathbf{n})t^2/(b-a)^2} \tag{1.12}$$

and

$$P\left\{\, |\tilde{Y}'_{\mathbf{n}} - \mu| \geq t \,\right\} \leq 2e^{-2m'(\mathbf{n})t^2/(b-a)^2} \tag{1.13}$$

for $t > 0$ and $\mathbf{n} \in \mathbf{N}$.

Clearly, the tighter the bounds a and b on the function v, the tighter the above inequalities.

Suppose as a worst-case scenario that the function v depends only upon the first of its K arguments and that $n_1 \le n_k$ for $2 \le k \le K$. Then the cross-product averages defined in (1.8) and (1.9) reduce to ordinary averages as in (1.1) and (1.2), and the inequalities (1.12) and (1.13) reduce to (1.5) and (1.6). These latter inequalities represent the best available Hoeffding bounds for this situation. In this sense, the inequalities in (1.12) and (1.13) can be viewed as tight worst-case Hoeffding bounds.

The bound in (1.13) sometimes can be tightened as follows. Fix $\mathbf{n} \in \mathbf{N}$ and suppose that for some positive integer $r = r(\mathbf{n}) < K$ we have $n_k < m_k$ for $1 \le k \le r$ and $n_k \ge m_k$ for $r < k \le K$. Set

$$w_{\mathbf{n}}(l_1, l_2, \ldots, l_r) = \frac{1}{m_{r+1} m_{r+2} \cdots m_K} \sum_{l_{r+1}=1}^{m_{r+1}} \sum_{l_{r+2}=1}^{m_{r+2}} \cdots \sum_{l_K=1}^{m_K} v(l_1, l_2, \ldots, l_r, l_{r+1}, l_{r+2}, \ldots, l_K) \tag{1.14}$$

for $(l_1, l_2, \ldots, l_r) \in \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_r$. Let $a(\mathbf{n})$ and $b(\mathbf{n})$ be lower and upper bounds, respectively, on the function $w_{\mathbf{n}}$. Applying Theorem 2 to the r-dimensional cross-product average of the function $w_{\mathbf{n}}$, we find that

$$P\left\{ |\tilde{Y}'_{\mathbf{n}} - \mu| \ge t \right\} \le 2e^{-2m'(\mathbf{n})t^2/(b(\mathbf{n})-a(\mathbf{n}))^2} \tag{1.15}$$

for $t > 0$ and $\mathbf{n} \in \mathbf{N}$. The key point is that it is sometimes possible to choose $a(\mathbf{n})$ and $b(\mathbf{n})$ such that $a(\mathbf{n}) > a$ and/or $b(\mathbf{n}) < b$, so that the bound in (1.15) is tighter than the bound in (1.13); see Section 3.1 below for an example. Of course, this approach to tightening the bound in (1.13) can be applied with obvious modifications when $\{ k\colon\ n_k < m_k \}$ is an arbitrary strict subset of $\{ 1, 2, \ldots, K \}$.

When $\mathbf{n} = (n, n, \ldots, n)$ for some $n \ge 1$, we write $\tilde{Y}_n$ and $\tilde{Y}'_n$ instead of $\tilde{Y}_{\mathbf{n}}$ and $\tilde{Y}'_{\mathbf{n}}$, respectively, so that

$$\tilde{Y}_n = \frac{1}{n^K} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_K=1}^{n} v(L_{1,i_1}, L_{2,i_2}, \ldots, L_{K,i_K}) \tag{1.16}$$

and

$$\tilde{Y}'_n = \left(\prod_{k=1}^{K} (n \wedge m_k) \right)^{-1} \sum_{i_1=1}^{n \wedge m_1} \sum_{i_2=1}^{n \wedge m_2} \cdots \sum_{i_K=1}^{n \wedge m_K} v(L'_{1,i_1}, L'_{2,i_2}, \ldots, L'_{K,i_K}). \tag{1.17}$$

In this case, the inequalities (1.12) and (1.13) take the form

$$P\left\{ |\tilde{Y}_n - \mu| \ge t \right\} \le 2e^{-2nt^2/(b-a)^2} \tag{1.18}$$

and

$$P\left\{ |\tilde{Y}'_n - \mu| \ge t \right\} \le \begin{cases} 2e^{-2nt^2/(b-a)^2} & \text{if } n < \max(m_1, m_2, \ldots, m_K); \\ 0 & \text{if } n \ge \max(m_1, m_2, \ldots, m_K) \end{cases}$$

for $t > 0$ and $n \geq 1$.

Hoeffding actually establishes his inequalities (1.3), (1.5), and (1.6) for an arbitrary collection of mutually independent, real-valued random variables (not necessarily discrete). The inequalities in (1.10), (1.12), and (1.13) also can be shown to hold at this level of generality. Hoeffding also establishes one-sided bounds, such as the following one-sided analogues of (1.5):

$$P\left\{ \overline{Y}_n - \mu \geq t \right\} \leq e^{-2nt^2/(b-a)^2}$$

and

$$P\left\{ \mu - \overline{Y}_n \geq t \right\} \leq e^{-2nt^2/(b-a)^2}.$$

For each two-sided bound presented in this paper, there exists a pair of one-sided analogues as in the above example.

In applications, the foregoing bounds are often "inverted" for purposes of obtaining confidence intervals. For example, it follows from (1.13) that for fixed $p \in (0,1)$ and $\mathbf{n} \in \mathbf{N}$

$$P\left\{ |\tilde{Y}'_{\mathbf{n}} - \mu| \leq \epsilon \right\} \geq p, \tag{1.19}$$

where

$$\epsilon = (b-a)\left(\frac{1}{2m'(\mathbf{n})}\ln\Big(\frac{2}{1-p}\Big)\right)^{1/2}, \tag{1.20}$$

Our final result gives an analogue of (1.19) for the ratio of two cross-product averages. Consider two real-valued functions f and g, each defined on Λ, along with finite constants a_f, b_f, a_g (≥ 0), and b_g such that

$$a_f \leq f(l_1, \ldots, l_K) \leq b_f \qquad \text{and} \qquad a_g \leq g(l_1, \ldots, l_K) \leq b_g \tag{1.21}$$

for $(l_1, \ldots, l_K) \in \Lambda$. For $p \in (0,1)$ and $\mathbf{n} \in \mathbf{N}$, set

$$\gamma_{\mathbf{n},p} = \left(\frac{1}{2m'(\mathbf{n})}\ln\Big(\frac{4}{1-p}\Big)\right)^{1/2}.$$

Also set

$$\epsilon_{\mathbf{n},p} = \gamma_{\mathbf{n},p}\left(\frac{\tilde{Y}'_{\mathbf{n}}(g)(b_f - a_f) + |\tilde{Y}'_{\mathbf{n}}(f)|(b_g - a_g)}{\tilde{Y}'_{\mathbf{n}}(g)\big(\tilde{Y}'_{\mathbf{n}}(g) - (b_g - a_g)\gamma_{\mathbf{n},p}\big)}\right) \tag{1.22}$$

if $\tilde{Y}'_{\mathbf{n}}(g) > (b_g - a_g)\gamma_{\mathbf{n},p}$; otherwise, set $\epsilon_{\mathbf{n},p} = \infty$. In (1.22), $\tilde{Y}'_{\mathbf{n}}(f)$ is defined as in (1.9), but with v replaced by f, and similarly for $\tilde{Y}'_{\mathbf{n}}(g)$. Define corresponding population averages $\mu(f)$ and $\mu(g)$ analogously. Take $0/0 = 0$.

Theorem 3. *Suppose that* (1.21) *holds and* $\mu(g) > 0$. *Then*

$$P\left\{\left|\frac{\tilde{Y}'_{\mathbf{n}}(f)}{\tilde{Y}'_{\mathbf{n}}(g)} - \frac{\mu(f)}{\mu(g)}\right| \le \epsilon_{\mathbf{n},p}\right\} \ge p$$

for $p \in (0,1)$ *and* $\mathbf{n} \in \mathbf{N}$, *where* $\epsilon_{\mathbf{n},p}$ *is defined by* (1.22).

Theorem 3 applies to the case of sampling without replacement, but analogous results hold for sampling with replacement. The previously-mentioned techniques for tightening bounds also can be applied in the current setting. Confidence intervals for other aggregates such as VARIANCE also can be obtained by adapting the techniques used to establish Theorem 3.

1.2. Subpopulations

We now consider the problem of estimating the mean, variance, and standard deviation of the numbers $\{\, v(i):\ i \in S \,\}$, where v is a real-valued function defined on $\{\, 1, 2, \ldots, m \,\}$ (with $m > 1$) and S is a nonempty subset of $\{\, 1, 2, \ldots, m \,\}$. We assume that the set S is specified by means of an indicator function u:

$$u(i) = \begin{cases} 1 & \text{if } i \in S; \\ 0 & \text{if } i \in \{\, 1, 2, \ldots, m \,\} - S \end{cases} \tag{1.23}$$

for $1 \le i \le m$. We also assume that it is not possible to sample directly from S; as in previous sections, each element in the sample is selected randomly and uniformly from the set $\{\, 1, 2, \ldots, m \,\}$. The function u is then applied to determine whether the sampled element is a member of S.

1.2.1. Mean

Denote by $\mu(S)$ the average of the function v over the set S:

$$\mu(S) = \frac{1}{|S|}\sum_{i \in S} v(i) = \frac{\sum_{i=1}^{m} u(i)v(i)}{\sum_{i=1}^{m} u(i)}, \tag{1.24}$$

where $|S|$ denotes the number of elements in S. We refer to $\mu(S)$ as the *subpopulation mean* over S.

Define random indexes $L_1, L_2, \ldots, L_n$ and $L'_1, L'_2, \ldots, L'_n$ as at the beginning of Section 1. For $n \ge 1$, two possible estimators of $\mu(S)$ are

$$\overline{Y}_n(S) = \frac{1}{I_n}\sum_{i=1}^{n} u(L_i)v(L_i) \tag{1.25}$$

and

$$\overline{Y}'_n(S) = \frac{1}{I'_n} \sum_{i=1}^{n \wedge m} u(L'_i) v(L'_i), \tag{1.26}$$

where

$$I_n = \sum_{i=1}^{n} u(L_i) \tag{1.27}$$

and

$$I'_n = \sum_{i=1}^{n \wedge m} u(L'_i). \tag{1.28}$$

We take $\overline{Y}_n(S) = 0$ when $I_n = 0$ and similarly for $\overline{Y}'_n(S)$. Neither $\overline{Y}_n(S)$ nor $\overline{Y}'_n(S)$ is unbiased for μ. As n becomes large, however, the bias of each estimator approaches 0 and each estimator converges to μ with probability 1.

Because $\overline{Y}_n(S)$ and $\overline{Y}'_n(S)$ are each biased estimators, it appears quite difficult to develop Hoeffding inequalities analogous to (1.12) and (1.13). In practice, however, it suffices to bound the *conditional* probability that $\overline{Y}_n(S)$ (resp., $\overline{Y}'_n(S)$) deviates from μ by more than a specified amount, given the observed value of I_n (resp., I'_n). The key observation (cf Section 2.12 in Cochran [2]) is as follows: given that $I_n = k$ (where $k > 0$), the estimator $\overline{Y}_n(S)$ is distributed as $(1/k) \sum_{i=1}^{k} v(L^*_i)$, where $\{ L^*_1, L^*_2, \ldots, L^*_k \}$ is a random sample of size k drawn from the set S uniformly with replacement. An analogous statement holds for the conditional distribution of $\overline{Y}'_n(S)$ given $I'_n = k$. Thus, $\overline{Y}_n(S)$ and $\overline{Y}'_n(S)$ are conditionally unbiased for $\mu(S)$ and, using the inequalities in (1.5) and (1.6), we obtain the bounds

$$P \left\{ |\overline{Y}_n(S) - \mu(S)| \geq t \;\middle|\; I_n = k \right\} \leq 2e^{-2kt^2/(b-a)^2}$$

for $t > 0$, $n \geq 1$, and $1 \leq k \leq n$, and

$$P \left\{ |\overline{Y}'_n(S) - \mu(S)| \geq t \;\middle|\; I'_n = k \right\} \leq 2e^{-2kt^2/(b-a)^2} \tag{1.29}$$

for $t > 0$, $n \geq 1$, and $1 \leq k \leq |S| \wedge n$.

1.2.2. Variance and Standard Deviation: Sampling with Replacement

Denote by $\sigma^2(S)$ the variance of the function v over the set S:

$$\sigma^2(S) = \frac{1}{|S|} \sum_{i \in S} \big(v(i) - \mu(S)\big)^2 = \frac{\sum_{i=1}^{m} u(i) \big(v(i) - \mu(S)\big)^2}{\sum_{i=1}^{m} u(i)}. \tag{1.30}$$

We refer to $\sigma^2(S)$ and $\sigma(S)$ as the *subpopulation variance* over S and *subpopulation standard deviation* over S, respectively.

First suppose that $S = \{ 1, 2, \ldots, m \}$ and that this fact is known *a priori*. (The value of m need not be known.) In this case we write μ for $\mu(S)$ and σ^2 for $\sigma^2(S)$. Fix $n \geq 2$ and define the sample average $\overline{Y}_n$ as in (1.1). It is well-known that the estimator

$$Z_n = \frac{1}{n-1} \sum_{i=1}^{n} (v(L_i) - \overline{Y}_n)^2 \tag{1.31}$$

is unbiased for σ^2; see, for example, Cramér [3, p. 347]. As pointed out by Hoeffding [10], Z_n can be written in the form

$$Z_n = \frac{1}{n(n-1)} \sum_{i \neq j} g\big(v(L_i), v(L_j)\big),$$

where $g(x, y) = (x - y)^2/2$; that is, Z_n is a "one-sample U-statistic" as defined in [10, 11]. Hoeffding's inequalities for U statistics [11, Section 5a] therefore can be applied in a straightforward manner to yield inequalities for Z_n. These computations, which have been carried out by Krafft and Schmitz [13], yield the inequalities

$$P \{ Z_n - \sigma^2 \geq t \} \leq e^{-8 \lfloor n/2 \rfloor t^2/(b-a)^4}, \tag{1.32}$$

$$P \{ \sigma^2 - Z_n \geq t \} \leq e^{-8 \lfloor n/2 \rfloor t^2/(b-a)^4}, \tag{1.33}$$

and

$$P \{ |Z_n - \sigma^2| \geq t \} \leq 2e^{-8 \lfloor n/2 \rfloor t^2/(b-a)^4} \tag{1.34}$$

for $t > 0$ and $n \geq 2$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x.

The random variable $\sqrt{Z_n}$ often serves as an estimator of the standard deviation σ. To obtain a bound analogous to that in (1.34), observe that

$$\sqrt{x} - \sqrt{y} = (x - 2\sqrt{xy} + y)^{1/2} \leq \sqrt{x - y}$$

for $0 \leq y \leq x$. Using (1.32), we find that

$$P \left\{ \sqrt{Z_n} - \sigma \geq t \right\} \leq P \left\{ \sqrt{Z_n - \sigma^2} \geq t \right\} = P \{ Z_n - \sigma^2 \geq t^2 \} \leq e^{-8 \lfloor n/2 \rfloor t^4/(b-a)^4}. \tag{1.35}$$

A similar argument using (1.33) yields the same bound for $P \{ \sigma - \sqrt{Z_n} \geq t \}$, and combination of the two bounds yields the inequality

$$P \left\{ \left| \sqrt{Z_n} - \sigma \right| \geq t \right\} \leq 2e^{-8 \lfloor n/2 \rfloor t^4/(b-a)^4}. \tag{1.36}$$

Now suppose that S is an arbitrary nonempty subset of $\{1, 2, \ldots, m\}$. The quantity $\sigma^2(S)$ can be estimated by

$$Z_n(S) = \frac{1}{I_n - 1} \sum_{i=1}^{n} u(L_i)\big(v(L_i) - \overline{Y}_n(S)\big)^2,$$

where u, $\overline{Y}_n(S)$, and I_n and are defined as in (1.23), (1.25), and (1.27), respectively. Using the above results and arguing as in the case of the subpopulation mean, we obtain the conditional inequalities

$$P\left\{ |Z_n(S) - \sigma^2(S)| \geq t \;\middle|\; I_n = k \right\} \leq 2e^{-8\lfloor k/2 \rfloor t^2/(b-a)^4}$$

and

$$P\left\{ \left| \sqrt{Z_n(S)} - \sigma(S) \right| \geq t \;\middle|\; I_n = k \right\} \leq 2e^{-8\lfloor k/2 \rfloor t^4/(b-a)^4}$$

for $t > 0$, $n \geq 2$, and $2 \leq k \leq n$.

1.2.3. Variance and Standard Deviation: Sampling without Replacement

Suppose at first that $S = \{1, 2, \ldots, m\}$, that this fact is known *a priori*, and that m is known. Fix $n \geq 2$ and define the sample average $\overline{Y}'_n$ as in (1.2). It is well-known that the estimator

$$Z'_n = \left(\frac{m-1}{m}\right) \frac{1}{(n \wedge m) - 1} \sum_{i=1}^{n \wedge m} (v(L'_i) - \overline{Y}'_n)^2 \tag{1.37}$$

is unbiased for σ^2; see, for example, [2, Theorem 2.4]. In analogy with (1.3), we have the following result.

Proposition 1. *Let Z_n and Z'_n be defined as in* (1.31) *and* (1.37)*, respectively. Then*

$$E\left[f(Z'_n)\right] \leq E\left[f(Z_n)\right]$$

for $n \geq 2$ and any convex function f. In particular, $\mathrm{Var}\left[Z'_n\right] \leq \mathrm{Var}\left[Z_n\right]$.

Pathak [18] gives a proof of this result based on the Rao-Blackwell inequality. In Section 4 we give a "first principles" proof along the lines of the original argument used by Hoeffding to establish (1.3).

As in Hoeffding's proof of (1.6), the inequality in Proposition 1 can be combined with the results in Section 1.2.2 to establish inequalities for Z'_n and $\sqrt{Z'_n}$.

Theorem 4. *Let Z'_n be defined as in* (1.37)*, and let a and b satisfy* (1.4)*. Then*

$$P\left\{\,|Z'_n-\sigma^2|\ge t\,\right\}\le 2e^{-8\lfloor n/2\rfloor t^2/(b-a)^4} \tag{1.38}$$

and

$$P\left\{\,\left|\sqrt{Z'_n}-\sigma\right|\ge t\,\right\}\le 2e^{-8\lfloor n/2\rfloor t^4/(b-a)^4} \tag{1.39}$$

for $t>0$ and $2\le n\le m$.

The inequality (1.38) is stated in [13], but the supporting proof is incomplete.

Now suppose that S is an arbitrary nonempty subset of $\{\,1,2,\ldots,m\,\}$. If $|S|$ is known, then for $n\ge 2$ the estimator

$$Z'_n(S)=\left(\frac{|S|-1}{|S|}\right)\frac{1}{I'_n-1}\sum_{i=1}^{n\wedge m}u(L'_i)\bigl(v(L'_i)-\overline{Y}'_n\bigr)^2.$$

is conditionally unbiased for $\sigma^2(S)$, given the value of I'_n. Arguing as in previous sections, we obtain the conditional inequalities

$$P\left\{\,|Z'_n(S)-\sigma^2|\ge t\;\bigm|\;I'_n=k\,\right\}\le 2e^{-8\lfloor k/2\rfloor t^2/(b-a)^4}$$

and

$$P\left\{\,\left|\sqrt{Z'_n(S)}-\sigma\right|\ge t\;\bigm|\;I'_n=k\,\right\}\le 2e^{-8\lfloor k/2\rfloor t^4/(b-a)^4}$$

for $t>0$, $n\ge 2$, and $2\le k\le|S|\wedge n$. If $|S|$ is unknown, we cannot use the estimator $Z'_n(S)$. If S is known to contain a sizable number of elements, however, a reasonable estimator can be obtained by ignoring the troublesome bias-correction term $(|S|-1)/|S|$ and simply computing the sample variance over all elements of S observed so far. This approach yields a biased estimator of $\sigma^2(S)$ given by

$$Z''_n(S)=\frac{1}{I'_n-1}\sum_{i=1}^{n\wedge m}u(L'_i)\bigl(v(L'_i)-\overline{Y}'_n(S)\bigr)^2, \tag{1.40}$$

where $\overline{Y}'_n(S)$ and I'_n are defined as in (1.26) and (1.28), respectively. Because Z''_n also is biased when given the value of I'_n, it is difficult to obtain even conditional Hoeffding inequalities without some extra information. When $|S|$ is large relative to $(b-a)^2$ and a good lower bound $d\le|S|$ is available, the following inequality can be useful.

Theorem 5. *Let $Z''_n(S)$ be defined as in* (1.40) *and let $d\le|S|$. Then*

$$P\left\{\,|Z''_n(S)-\sigma^2(S)|\ge t\;\bigm|\;I'_n=k\,\right\}\le 2e^{-8\lfloor k/2\rfloor\tau^2/(b-a)^4} \tag{1.41}$$

and

$$P\left\{\,|\sqrt{Z_n''(S)}-\sigma(S)|\ge t\;\Big|\;I_n'=k\,\right\}\le 2e^{-8\lfloor k/2\rfloor\nu^2/(b-a)^4} \tag{1.42}$$

for $t>0$, $n\ge 2$, *and* $2\le k\le |S|\wedge n$, *where*

$$\tau=\tau(t,d,a,b)=\frac{d-1}{d}t-\frac{(b-a)^2}{4(d-1)}$$

and

$$\nu=\nu(t,d,a,b)=\frac{d-1}{d}t^2-\frac{(b-a)^2}{4(d-1)}.$$

Note that, given $I_n'=k$, we can set $d=k$ if k is sufficiently large.

1.3. Some Further Refinements

Krafft and Schmitz [13] provide several techniques that can be used to tighten the various bounds given in this section. The first (trivial) observation is that if $a\le v\le b$, then, for example, $P\{\,|\tilde{Y}_{\mathbf{n}}-\mu|\ge t\,\}=0$ for $t>b-a$, where $\tilde{Y}_n$ and μ are as in Theorem 2. Similarly, it follows from the inequality in (4.5) below that $|Z_n-\sigma^2|\le t_n^*$ for $n\ge 0$, where Z_n and σ^2 are as in (1.34) and

$$t_n^*=\frac{n(b-a)^2}{4(n-1)}.$$

Thus, $P\{\,|Z_n-\sigma^2|\ge t\,\}=0$ for $t>t_n^*$ and $P\{\,|\sqrt{Z_n}-\sigma|\ge t\,\}=0$ for $t>\sqrt{t_n^*}$. All of the other inequalities in this section can be tightened in a similar manner.

A less trivial result is obtained after rewriting (1.12), (1.34), and (1.36) in the form

$$P\{\,|\tilde{Y}_{\mathbf{n}}-\mu|\ge t\,\}\le 2e^{-m(\mathbf{n})\theta(t/(b-a))},$$

$$P\{\,|Z_n-\sigma^2|\ge t\,\}\le 2e^{-\lfloor n/2\rfloor\theta(2t/(b-a)^2)},$$

and

$$P\left\{\,\left|\sqrt{Z_n}-\sigma\right|\ge t\,\right\}\le 2e^{-\lfloor n/2\rfloor\theta(2t^2/(b-a)^2)},$$

respectively, where $\theta(x)=2x^2$. It follows from results in [13] that the above inequalities hold with θ replaced by θ_0, where

$$\theta_0(x)=2x^2+\frac{4}{9}x^4+\frac{2}{9}x^6.$$

All of the other inequalities in this section can be rewritten and tightened in a similar manner.

2. Application to Join-Selectivity Estimation

A (select-)join query with K (≥ 2) input relations $R_1, R_2, \ldots, R_K$ specifies a subset of the cross product $R_1 \times R_2 \times \cdots \times R_K$. For each element $(j_1, j_2, \ldots, j_K)$ in this subset, tuples $j_1, j_2, \ldots, j_K$ are concatenated to form a tuple of the output relation $R_{1,2,\ldots,K}$. The "selectivity of the join" is the number of tuples in $R_{1,2,\ldots,K}$ divided by the number of elements in the cross product of the input relations. Selectivity estimates play a key role in query optimization for ORDBMS's as well as for capacity planning, cost estimation for online queries, system access control, load balancing, and statistical studies.

Our formulation of the selectivity-estimation problem follows [7]. For $j_1 \in R_1, j_2 \in R_2, \ldots, j_K \in R_K$, set $v_0(j_1, j_2, \ldots, j_K) = 1$ if tuples $j_1, j_2, \ldots, j_K$ join (that is, if these tuples are concatenated to form a tuple of $R_{1,2,\ldots,K}$); otherwise, set $v_0(j_1, j_2, \ldots, j_K) = 0$. The function v_0 is determined by the join and selection predicates that make up the query. We wish to estimate the *selectivity* μ, defined as

$$\mu = \frac{|R_{1,2,\ldots,K}|}{|R_1 \times \cdots \times R_K|} = \frac{1}{|R_1 \times \cdots \times R_K|} \sum_{j_1 \in R_1} \sum_{j_2 \in R_2} \cdots \sum_{j_K \in R_K} v_0(j_1, j_2, \ldots, j_K).$$

A naïve method for estimating the selectivity μ is to use a "tuple level, independent" sampling scheme, denoted by *t_indep*. At the nth sampling step of *t_indep*, a tuple is selected randomly and uniformly from each input relation. The observation X_n is then computed from the K selected tuples, where $X_n = 1$ if the tuples join, and $X_n = 0$ otherwise. In other words, X_n is the selectivity of the K-way join of the randomly selected tuples. The tuples are then discarded prior to the next sampling step. For each $n \geq 1$, the observations $X_1, X_2, \ldots, X_n$ are identically distributed, and the average of these observations is an unbiased estimator of μ; if samples are drawn with replacement, then these observations are also mutually independent.

An alternative approach is to estimate μ using the "page-level, cross-product" sampling scheme proposed in Hou, Ozsoyoglu, and Taneja [12] and analyzed in [7, 19]. We denote this scheme by *p_cross*. At each sampling step of *p_cross* and for each of the K input relations, a page of tuples is selected randomly and uniformly from among the pages that make up the relation; this randomly selected page is stored in main memory. At the nth sampling step, $n^K - (n-1)^K$ observations are generated by computing

(i) the selectivity of the K-way join of the pages selected at the current sampling step; and

(ii) the selectivities of all possible K-way joins among pages selected at the current sampling step and pages selected at previous sampling steps.

Although *p_cross* examines many more tuples per sampling step than *t_indep*, the resulting observations are not mutually independent, even when samples are drawn with replacement.

Haas et al. [7] compare *p_cross* and *t_indep* when samples are drawn with replacement. They show that, for any fixed number of sampling steps, selectivity estimators based on

the *p_cross* scheme always have variance less than or equal to that of estimators based on the *t_indep* scheme. In practice, the variance of the selectivity estimator can be smaller by orders of magnitude when *p_cross* is used. We therefore focus throughout on the *p_cross* sampling scheme and its variants.

The sampling cost of *p_cross* sometimes can be reduced by using "index-assisted" sampling as proposed by Lipton, Naughton, and Schneider [14, 15] and extended in [7]. Suppose, for example, that we wish to estimate the selectivity of an equijoin of relations R_1 and R_2 using the *t_indep* scheme and that R_2 has an index on its join attribute; that is, the join predicate is "$R_1.a = R_2.a$" and R_2 has an index on attribute a. At each sampling step, one tuple is selected randomly and uniformly from relation R_1 and the total number of tuples from R_2 that join with this random tuple is computed using the index; no sampling from R_2 is required. Thus, we obtain $|R_2|$ observations at each sampling step. This idea extends in a straightforward manner to the *p_cross* scheme and to general K-way joins when one or more of the input relations has a combined index on (the concatenation of) all relevant join and selection attributes.

In the following, we assume that samples are drawn with replacement. We also assume that tuples are stored and brought into main memory in pages, where each page contains N (≥ 1) tuples. To discuss page-level and index-assisted sampling schemes in a unified way, we consider a generalized scheme in which, for $1 \leq k \leq K$, the tuples in relation R_k are partitioned into m_k *blocks* $B(k,1), B(k,2), \ldots, B(k,m_k)$ with t_k ($= |R_k|/m_k$) tuples per block. For each relation R_k, a block of tuples is selected at each sampling step, and all of the tuples in the block are brought into main memory. When $t_1 = \cdots = t_K = N$ we have pure page-level sampling. When $t_j = |R_j|$ for one or more values of j we have index-assisted sampling. (As indicated above, we don't actually bring all of the tuples in an indexed relation into memory, we just perform an index lookup. Such a lookup, however, is equivalent to examining all of the tuples in the indexed relation, as far as their join and selection attributes are concerned. In general, one or more I/O's may be required to perform the lookup; see [7] for a detailed discussion of sampling costs.) We assume that at most $K-1$ indexes are available; Ganguly, Gibbons, Matias, and Silberschatz [4] provide an estimation procedure when $K = 2$ and there is an index on each input relation.

For $(l_1, l_2, \ldots, l_K) \in \Lambda$, denote by $v(l_1, l_2, \ldots, l_K)$ the selectivity of the join of blocks $B(1,l_1), B(2,l_2), \ldots, B(K,l_K)$:

$$v(l_1, l_2, \ldots, l_K) = \frac{1}{t_1 t_2 \cdots t_K} \sum_{j_1 \in B(1,l_1)} \sum_{j_2 \in B(2,l_2)} \cdots \sum_{j_K \in B(K,l_K)} v_0(j_1, j_2, \ldots, j_K).$$

The function v is called the *selectivity function* for the join. Observe that the selectivity μ can be represented as a cross-product average of the form (1.7). Denote by $L_{k,i}$ (resp., $L'_{k,i}$) the random index of the block of tuples selected from relation R_k at the ith sampling step when samples are drawn with (resp., without) replacement. Then for $n \geq 1$ the estimators $\tilde{Y}_n$ and $\tilde{Y}'_n$ defined by (1.16) and (1.17), respectively, are each unbiased for μ.

Suppose that we wish to estimate the selectivity μ to within $\pm\epsilon$ with probability p, where $\epsilon > 0$ and $p \in (0, 1)$. In general, it is impossible to satisfy the precision criterion with a probability exactly equal to p. To address this problem, Haas et al. [7] introduced a fixed-precision estimation procedure called *f_p_cross*. In this procedure, the basic *p_cross* sampling procedure is executed for a random number of sampling steps. After each sampling step of *p_cross*, a stopping rule is used to determine whether to continue sampling or to stop sampling and return the cross-product average based on all of the blocks sampled so far. The final estimate $\tilde{Y}(\epsilon)$ is given by $\tilde{Y}(\epsilon) = \tilde{Y}_{N(\epsilon)}$, where $\tilde{Y}_n$ is defined by (1.16) for $n \geq 1$ and $N(\epsilon)$ is the random number of sampling steps executed by the procedure.

The stopping rule for *f_p_cross* is obtained by approximating the distribution of $\tilde{Y}(\epsilon)$ by a normal distribution. Use of this approximation leads to an estimate of the probability that the precision criterion is satisfied. If the estimated probability is greater than or equal to the prespecified probability, then no further samples are taken; see [7] for details. It is shown in [7] that the *f_p_cross* procedure is "asymptotically consistent" in the sense that the probability of satisfying the precision criterion converges to the prespecified value as the precision criterion becomes increasingly stringent: $\lim_{\epsilon \to 0} P\{ |\tilde{Y}(\epsilon) - \mu| \leq \epsilon \} = p$. Moreover, *f_p_cross* is "asymptotically efficient" in the sense that the total number of sampling steps converges to the theoretical minimum number of required steps as ϵ becomes small. The number of sampling steps required by *f_p_cross* is independent of the size of the input relations. Thus, for a fixed precision criterion, the cost of sampling relative to the cost of computing μ exactly decreases as the size of input relations increases; see Haas and Naughton [8] for further discussion.

One shortcoming of *f_p_cross* is that there is no guaranteed upper bound on the number of sampling steps. This can be an issue because the normal approximation that underlies the stopping rule is exact only in the limit as $\epsilon \to 0$. For fixed positive ϵ, the probability that the precision criterion is satisfied can be underestimated, resulting in too many sampling steps.

We can use the results in Section 1.1 to alleviate this problem and provide a guaranteed upper bound on the number of sampling steps. (See [15] for a related discussion of "sanity bounds.") Set

$$n_0 = n_0(\epsilon, p) = \left\lceil \frac{1}{2\epsilon^2} \ln\left(\frac{2}{1-p}\right) \right\rceil,$$

where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Then the idea is to stop sampling in *f_p_cross* after $\min\big(n_0, N(\epsilon)\big)$ steps, where $N(\epsilon)$ is the number of steps executed in the original *f_p_cross* procedure. It follows from (1.12) that no more than n_0 sampling steps are ever needed to satisfy the precision criterion.

The bound n_0 can be improved upon when there is additional *a priori* information on the selectivity of the join. For example, consider a join query consisting of the single join predicate

$$R_1.a_1 = R_2.a_1 \textbf{ and } R_2.a_2 = R_3.a_2 \textbf{ and } \cdots \textbf{ and } R_{K-1}.a_{K-1} = R_K.a_{K-1}.$$

Suppose it is known *a priori* that each tuple in relation R_k $(1 \leq k \leq K-1)$ joins with at most γ_{k+1} tuples in relation R_{k+1}, where $\gamma_{k+1} < t_{k+1}$. Then $0 \leq v(l_1, l_2, \ldots, l_K) \leq u$ for all $l_1, l_2, \ldots, l_K$, where $u = \prod_{k=2}^{K}(\gamma_k/t_k) < 1$. It follows from (1.18)that

$$P\left\{ |\tilde{Y}_{n_0(u)} - \mu| \leq \epsilon \right\} \geq p,$$

where

$$n_0(u) = n_0(u; \epsilon, p) = \left\lceil \frac{u^2}{2\epsilon^2} \ln\left(\frac{2}{1-p}\right) \right\rceil \leq n_0.$$

3. Application to Online Aggregation

Users of an ORDBMS often execute "aggregation queries" in order to obtain statistical summaries of large, complex data sets. Aggregation queries are processed by first executing a (typically complex) sequence of joins and selections on the base relations to create an output relation; each tuple of the output relation is then mapped to a real number and aggregate quantities such as the sum, mean, or variance are computed from the numbers. The rows of the output table sometimes are divided into groups based on values of the data attributes and aggregates are computed separately for each group.

We focus on scientific and decision-support applications in which the user explores a data set by executing a sequence of aggregation queries in an interactive manner. The formulation of each successive query is influenced by the results of previous queries. In this setting it is crucial that the processing time for each query be as short as possible. Moreover, since the typical goal is simply to get a rough feel for the data, approximate results often suffice. Unfortunately, current database systems do not adequately support such interactive exploration: the result of an aggregation query is not returned to the user until the query has run to completion and the exact answer has been computed. Such an exact computation, which can involve the processing of many terabytes of data, can take an extremely long time. Moreover, there is no feedback during processing; a user can lose much valuable time before discovering that a particular query is misguided or uninformative.

In an effort to address these problems, Hellerstein, Haas, and Wang [9] have proposed an *online aggregation* interface to an ORDBMS. This interface lets users both observe the progress of their aggregation queries and control the execution of these queries on the fly. The idea is to retrieve the pages of each base table in random order, so that the rows retrieved so far can be viewed as a random sample. At each time point, the system displays a running estimate of the final value of the aggregate based on all of the pages retrieved so far. The system also indicates the estimated proximity of each running estimate to the corresponding final result by means of a running confidence interval. The user can abort processing of the query as soon as the intervals become sufficiently short. When aggregates are being computed for multiple groups, some online-aggregation prototype systems maintain a running confidence interval for each group and permit the user to increase/decrease

the relative processing speed for an individual group on the fly or abort processing for an individual group.

Classical confidence-interval formulas based on results for i.i.d. observations cannot be applied in the setting of online aggregation because of the complicated correlation structure induced by the joins and selections executed prior to the final aggregation step. Large-sample confidence intervals suitable for online aggregation are given in [5, 6]; such intervals are based on central limit theorems. Although the large-sample intervals are useful in the earlier stages of query processing, the set of auxiliary statistics needed to compute the intervals can become very large as the processing proceeds. Also, the intervals rest on the approximating assumption that samples are obtained with replacement, and this assumption becomes untenable as the sample size (number of records scanned) becomes large. Finally, the actual coverage probability for the intervals can be less than the nominal value.

The results in Section 1 can be used to obtain *conservative* running confidence intervals for a variety of online aggregation queries. That is, for a prespecified parameter $p \in (0, 1)$, a number ϵ is displayed such that the current value of the running estimate is within $\pm\epsilon$ of the final answer μ with probability $\geq p$. Although conservative confidence intervals are wider in general than large-sample intervals, the computations for the conservative intervals require minimal memory and CPU time and avoid the undercoverage problem alluded to above.

In the following subsections, we illustrate the application of our results to online aggregation processing by means of an extended example. Consider an online aggregation interface to a relational database system containing the following three relation schemes:

$$\begin{aligned} \textit{Supplier-scheme} &= (\textit{part-num, supplier, price}) \\ \textit{Inventory-scheme} &= (\textit{part-num, location, quantity}) \\ \textit{Sales-scheme} &= (\textit{item, month, day, location, number-sold}) \end{aligned}$$

Assume that there are m_1 pages of tuples (with t_1 tuples per page) in the *Supplier* relation, m_2 pages of tuples (with t_2 tuples per page) in the *Inventory* relation, and m_3 pages of tuples (with t_3 tuples per page) in the *Sales* relation. Tuples are retrieved from each input relation a page at a time. Denote by *Supplier*(i) the ith tuple in the *Supplier* relation, and similarly for the *Inventory* and *Sales* relations.

3.1. Complex SUM, COUNT, and AVERAGE Queries

As a first example, consider the query that returns the total value μ of inventory stored at the San Jose warehouse:

```
SELECT SUM(Supplier.price * Inventory.quantity)
FROM Supplier, Inventory
WHERE Supplier.part-num = Inventory.part-num
AND Inventory.location = ´San Jose´;
```

For $1 \le i \le m_1 t_1$ and $1 \le j \le m_2 t_2$, set

$$v_0(i,j) = \big(\mathit{Supplier}(i).\mathit{price}\big) \times \big(\mathit{Inventory}(j).\mathit{quantity}\big)$$

if $\mathit{Supplier}(i).\mathit{part\text{-}num} = \mathit{Inventory}(j).\mathit{part\text{-}num}$ and $\mathit{Inventory}(j).\mathit{location} =$ "San Jose"; otherwise, set $v_0(i,j) = 0$. Also set

$$v(l_1,l_2) = m_1 m_2 \sum_{i \in B(1,l_1)} \sum_{j \in B(2,l_2)} v_0(i,j), \tag{3.1}$$

where $B(1,k)$ and $B(2,k)$ denote the kth page of tuples from the *Supplier* relation and *Inventory* relation, respectively. Observe that μ is of the form (1.7) with v defined as above. Suppose it is known that

(i) no part has a price greater than p; and

(ii) no more than q units of a given part type are stored at the San Jose warehouse simultaneously.

Such information often can be deduced from statistics that are maintained in the database system catalog for use by the query optimizer. It follows that $\max_{i,j} v_0(i,j) \le pq$ and the function v satisfies the bounds $a \le v \le b$ with $a = 0$ and $b = m_1 m_2 t_1 t_2 pq$. Tighter bounds on v can be obtained by observing that the *part-num* attribute is in fact a key for the *Supplier* table; that is, there are no duplicate *part-num* values in *Supplier*. (Otherwise, the query is not well posed.) We can take $b = m_1 m_2 t_2 pq$, since each row in the *Inventory* table joins with at most one row in the *Supplier* table.

Suppose that at a given time point n_1 pages have been retrieved from the *Supplier* relation and n_2 pages have been retrieved from the *Inventory* relation. Then the estimator $\tilde{Y}'_{\mathbf{n}}$ defined by (1.9) is unbiased for μ, where $\mathbf{n} = (n_1, n_2)$. It follows from (1.13) that a conservative $100p\%$ confidence interval for μ is given by $[\tilde{Y}'_{\mathbf{n}} - \epsilon, \tilde{Y}'_{\mathbf{n}} + \epsilon]$, where ϵ is defined by (1.20) with a and b as above. If the assumptions in (i) and (ii) hold and

(iii) all of the pages in the *Supplier* relation and a portion of the pages in the *Inventory* relation have been retrieved, so that $n_1 = m_1$ and $n_2 < m_2$,

then we can obtain a tighter interval than the one given above. In analogy to (1.14), set

$$w_{\mathbf{n}}(l_2) = \frac{1}{m_1} \sum_{l_1=1}^{m_1} v(l_1,l_2) = m_2 \sum_{l_1=1}^{m_1} \sum_{i \in B(1,l_1)} \sum_{j \in B(2,l_2)} v_0(i,j)$$

for $1 \le j \le m_2$. Since at most t_2 terms in the above sum are positive (with the value of a positive term equal to at most pq), we can obtain a tighter interval $[\tilde{Y}'_{\mathbf{n}} - \epsilon(\mathbf{n}), \tilde{Y}'_{\mathbf{n}} + \epsilon(\mathbf{n})]$ by setting

$$\epsilon = \big(b(\mathbf{n}) - a(\mathbf{n})\big) \left(\frac{1}{2m'(\mathbf{n})} \ln\Big(\frac{2}{1-p}\Big) \right)^{1/2}, \tag{3.2}$$

where $a(\mathbf{n}) = 0$ and $b(\mathbf{n}) = m_2 t_2 pq$.

The formula (1.20) for the precision parameter ϵ assumes that pages, but not necessarily tuples, are retrieved in random order. That is, the attribute values of a tuple may depend on the page on which the tuple resides. When such dependence is present, the tuples are said to be *clustered* on the pages; otherwise, they are said to be *unclustered*. If it is known that rows are unclustered, the lengths of the conservative confidence intervals can be considerably reduced. To see this, suppose that the assumptions in (i) and (ii) hold. Also suppose that n_1 ($< m_1$) pages have been retrieved from the *Supplier* table and n_2 ($< m_2$) pages have been retrieved from the *Inventory* table. If the rows are clustered on the pages then ϵ is computed from (1.20) with $\mathbf{n} = (n_1, n_2)$ as discussed above. If the rows are unclustered, however, then rows (and not just pages) are retrieved in random order, and ϵ can be computed using the same formula as in the clustered case except that $\mathbf{n} = (t_1 n_1, t_2 n_2)$. Similarly, if the assumptions in (i)–(iii) hold, then ϵ is computed from (3.2) with $\mathbf{n} = (n_1, n_2)$ in the clustered case and $\mathbf{n} = (t_1 n_1, t_2 n_2)$ in the unclustered case.

As a second example of an aggregation query, consider the query that returns the total number μ of part types at the San Jose warehouse that are supplied by the Acme company:

```
SELECT COUNT(*)
FROM Supplier, Inventory
WHERE Supplier.part-num = Inventory.part-num
AND Supplier.supplier =´Acme´ AND Inventory.location = ´San Jose´;
```

We assume here that the combined attribute *(part-num, location)* is a key for the *Inventory* relation. Set

$$v_0(i, j) = 1$$

if *Supplier*(i).*part-num* = *Inventory*(j).*part-num*, *Supplier*(i).*supplier* = "Acme", and *Inventory*(j).*location* = "San Jose"; otherwise, set $v_0(i, j) = 0$. Defining $v(i, j)$ as in (3.1), we see that μ is of the form (1.7). With no additional assumptions about the data, the function v satisfies the bounds $a \le v \le b$ with $a = 0$ and $b = m_1 m_2 t_1 t_2$. As with the SUM query discussed above, tighter bounds on v may be obtainable. Since μ has the same mathematical form as the result of the SUM query discussed above, the previous discussion for the SUM query carries over to the current setting almost unchanged; only the specific form of the function v and the value of b are different. In particular, the estimator $\tilde{Y}'_{\mathbf{n}}$ defined by (1.9) is unbiased for μ and conservative choices for the confidence interval half-width ϵ are of the form (1.20). Moreover, the result in (1.15) can be applied as described for SUM queries.

Finally, consider the query that returns the average price μ of the part types supplied by the Acme company and stored in San Jose:

```
SELECT AVERAGE(Supplier.price)
FROM Supplier, Inventory
WHERE Supplier.part-num = Inventory.part-num
AND Supplier.supplier =´Acme´ AND Inventory.location = ´San Jose´;
```

For $1 \leq i \leq m_1 t_1$ and $1 \leq j \leq m_2 t_2$, set

$$f_0(i,j) = \textit{Supplier}(i).\textit{price}$$

if $\textit{Supplier}(i).\textit{part-num} = \textit{Inventory}(j).\textit{part-num}$, $\textit{Supplier}(i).\textit{supplier} =$ "Acme", and $\textit{Inventory}(j).\textit{location} =$ "San Jose"; otherwise, set $f_0(i,j) = 0$. Also set

$$f(l_1, l_2) = \sum_{i \in B(1,l_1)} \sum_{j \in B(2,l_2)} f_0(i,j).$$

Similarly, set

$$g_0(i,j) = 1$$

if $\textit{Supplier}(i).\textit{part-num} = \textit{Inventory}(j).\textit{part-num}$, $\textit{Supplier}(i).\textit{supplier} =$ "Acme", and $\textit{Inventory}(j).\textit{location} =$ "San Jose"; otherwise, set $g_0(i,j) = 0$. Also set

$$g(l_1, l_2) = \sum_{i \in B(1,l_1)} \sum_{j \in B(2,l_2)} g_0(i,j).$$

Then μ is of the form $\mu(f)/\mu(g)$ as in Theorem 3. Under assumption (i) above, a conservative $100p\%$ confidence interval for μ is given by $[\tilde{Y}'_{\mathbf{n}} - \epsilon_{\mathbf{n},p}, \tilde{Y}'_{\mathbf{n}} + \epsilon_{\mathbf{n},p}]$, where $\epsilon_{\mathbf{n},p}$ is given by (1.22) with $a_f = 0$, $b_f = t_2 p$, $a_g = 0$, and $b_g = t_2$. (As above, we have used the fact that the *part-num* attribute is a key for the *Supplier* table.)

3.2. Summary Statistics Defined on Selections

We assume throughout that the online aggregation system retrieves tuples from each relation in random order. As discussed above, this assumption is stronger than the assumption that the system retrieves pages in random order and implies that tuples are unclustered.

Consider the query that returns the average daily number $\mu(S)$ of widgets sold in December:

```
SELECT AVG(Sales.number-sold)
FROM Sales
WHERE Sales.month=´December´
AND Sales.item=´widget´;
```

For $1 \leq i \leq m_3 t_3$, set $v(i) = \textit{Sales}(i).\textit{number-sold}$. Also set $u(i) = 1$ if $\textit{Sales}(i).\textit{item} =$ "widget" and $\textit{Sales}(i).\textit{month} =$ "December"; otherwise, set $u(i) = 0$. Then $\mu(S)$ is of the form (1.24) with u and v defined as above. Suppose that no location stocks more than w widgets on any day. Then the function v satisfies the bounds $a \leq v \leq b$ with $a = 0$ and $b = w$.

Suppose that at a given time point n tuples (equivalently, n/t_3 pages) have been retrieved from the *Sales* relation and k (> 0) of these tuples correspond to widgets sold in December.

The estimator $\overline{Y}'_n(S)$ defined by (1.26) (with $I'_n = k$) is conditionally unbiased for $\mu(S)$. By (1.29), a conservative $100p\%$ confidence interval is given by $[\overline{Y}'_n(S) - \epsilon, \overline{Y}'_n(S) + \epsilon]$, where

$$\epsilon = (b-a)\left(\frac{1}{2k}\ln\Big(\frac{2}{1-p}\Big)\right)^{1/2}.$$

Note that an alternative confidence interval is given by Theorem 3. Neither interval dominates the other in all cases, and the shorter of the two intervals should be used.

Now consider the query that returns the variance $\sigma^2(S)$ of the daily number of widgets sold in December:

```
SELECT VARIANCE(Sales.number-sold)
FROM Sales
WHERE Sales.month='December'
AND Sales.item='widget';
```

Defining u and v as above, we see that $\sigma^2(S)$ is of the form (1.30). As above, suppose that n tuples have been retrieved from the *Sales* relation and k of these tuples correspond to widgets sold in December. We can estimate $\sigma^2(S)$ by $Z''_n(S)$, where this estimator is defined by (1.40) (with $I'_n = k$). It follows from (1.41) that a conservative $100p\%$ confidence interval is given by $[Z''_n(S) - \epsilon_0, Z''_n(S) + \epsilon_0]$, where

$$\epsilon_0 = \frac{d(b-a)^2}{4(d-1)^2} + \frac{d(b-a)^2}{d-1}\left(\frac{1}{8\lfloor k/2\rfloor}\ln\Big(\frac{2}{1-p}\Big)\right)^{1/2}.$$

In the above expression, d is a lower bound on the total number $|S|$ of tuples that correspond to widgets sold in December. The parameter d can be taken equal to k; if a larger lower bound on $|S|$ is available, then d should be taken equal to this lower bound. If the quantity of interest is the standard deviation rather than the variance, then $\sqrt{Z''_n(S)}$ can be used to estimate $\sigma(S)$. By (1.42), a conservative $100p\%$ confidence interval is given by $[Z''_n(S) - \epsilon, Z''_n(S) + \epsilon]$, where $\epsilon = \sqrt{\epsilon_0}$.

4. Proofs

Proof of Theorem 1. Fix a vector $\mathbf{n} \in \mathbf{N}$. For $1 \le k \le K$ let N_k $(\ge n_k)$ be the unique random integer such that $S_k = \{ L_{k,i} : 1 \le i \le N_k \}$ contains exactly n_k distinct values; denote the set of these distinct values by $T_k = \{ L'_{k,i} : \ 1 \le i \le n_k \}$. Observe that T_k coincides with a random sample of size n_k drawn uniformly without replacement from Λ_k. Set $S = (S_1, S_2, \ldots, S_K)$ and $T = (T_1, T_2, \ldots, T_K)$. We can view $\tilde{Y}_{\mathbf{n}}$, $\tilde{Y}'_{\mathbf{n}}$, and T as functions of S. Write

$$E\left[\tilde{Y}_{\mathbf{n}} \mid T\right] = \frac{1}{n_1 \cdots n_K}\sum_{i_1=1}^{n_1}\cdots\sum_{i_K=1}^{n_K} E\left[v(L_{1,i_1}, \ldots, L_{K,i_K}) \mid T\right]. \tag{4.1}$$

It follows from symmetry considerations that

$$P\{ L_{k,i} = l \mid l \in T_k \} = \frac{1}{n_k},$$

for $1 \le l \le m_k$, $1 \le i \le n_k$, and $1 \le k \le K$, so that

$$E[v(L_{1,i_1}, \ldots, L_{K,i_K}) \mid T] = \frac{1}{n_1 \cdots n_K} \sum_{i_1=1}^{n_1} \cdots \sum_{i_K=1}^{n_K} v(L'_{1,i_1}, \ldots, L'_{K,i_K}). \tag{4.2}$$

Substituting (4.2) into (4.1), we find that $\tilde{Y}'_{\mathbf{n}} = E[\tilde{Y}_{\mathbf{n}} \mid T]$. The desired result now follows from the Rao-Blackwell Theorem, because T is a sufficient statistic for the sampling scheme that generates S; see Pathak [17] for definitions and details. □

We now prove Theorem 2 using an approach developed by Hoeffding in the setting of U-statistics and averages of m-dependent random variables; see [11, Section 5].

We first recall the following elementary but useful inequality, attributed by Hoeffding [11] to S. N. Bernstein:

$$P\{ X \ge 0 \} \le \inf_{h \ge 0} E\left[e^{hX}\right] \tag{4.3}$$

for any random variable X. The inequality in (4.3) holds since $P\{ X \ge 0 \} = E[g(X)]$, where

$$g(x) = \begin{cases} 1 & \text{if } x \ge 0; \\ 0 & \text{if } x < 0, \end{cases}$$

and $g(x) \le e^{hx}$ for all real x and $h \ge 0$. We also need the following proposition.

Proposition 2. *Let $Y = p_1 U_1 + p_2 U_2 + \cdots + p_m U_m$, where each of $U_1, U_2, \ldots, U_m$ is the average of $n \ge 1$ independent random variables taking values in $[0, 1]$ and $p_1, p_2, \ldots, p_m$ are nonnegative numbers such that $p_1 + p_2 + \cdots + p_m = 1$. The random variables $U_1, U_2, \ldots, U_m$ need not be mutually independent, but are assumed to have common mean μ. Then*

$$\inf_{h \ge 0} E\left[e^{h(Y - \mu - t)}\right] \le e^{-2nt^2}$$

for $t > 0$.

This proposition follows almost immediately from results in Sections 1, 4 and 5 in [11]. Indeed,

$$E\left[e^{h(nY - n\mu - nt)}\right] \le \sum_{i=1}^{m} p_i E\left[e^{h(nU_i - n\mu - nt)}\right].$$

for $h \geq 0$ by the inequality that is stated just prior to (5.2) in [11]. On the other hand, it follows from (1.8) and (4.16) in [11] that

$$E\left[e^{h(nU_i - n\mu - nt)}\right] \leq e^{-hnt + h^2 n/8}$$

for $1 \leq i \leq m$ and $h \geq 0$. The right side of the above inequality is minimized when $h = 4t$, so that

$$\inf_{h \geq 0} E\left[e^{h(nU_i - n\mu - nt)}\right] \leq e^{-2nt^2}$$

for $1 \leq i \leq m$, and the asserted inequality follows directly.

Proof of Theorem 2. Fix $t > 0$ and $\mathbf{n} \in \mathbf{N}$, and set

$$m = m(\mathbf{n}) = \min(n_1, n_2, \ldots, n_K).$$

Assume without loss of generality that $n_1 = m$. Also assume without loss of generality that $0 \leq v \leq 1$; the general result follows by considering the function $v'(\cdot) = \big(v(\cdot) - a\big)/(b - a)$. Set

$$U(i_2, i_3, \ldots, i_K) = \frac{1}{n_1} \sum_{j=1}^{n_1} v(L_{1,j}, L_{2,i_2+j}, \ldots, L_{K,i_K+j})$$

for $(i_2, i_3, \ldots, i_K) \in \{1, 2, \ldots, n_2\} \times \{1, 2, \ldots, n_3\} \times \cdots \times \{1, 2, \ldots, n_K\}$, where an index of the form "$i_k + j$" is interpreted as $\big((i_k + j - 1) \bmod n_k\big) + 1$ (that is, the indices "wrap around"). For example, when $K = 3$, $n_1 = 3$, $n_2 = 4$, and $n_3 = 3$, we have

$$U(3, 1) = \frac{v(L_{1,1}, L_{2,4}, L_{3,2}) + v(L_{1,2}, L_{2,1}, L_{3,3}) + v(L_{1,3}, L_{2,2}, L_{3,1})}{3}.$$

Observe that

$$\tilde{Y}_{\mathbf{n}} = \frac{1}{n_2 n_3 \cdots n_K} \sum_{i_2=1}^{n_2} \sum_{i_3=1}^{n_3} \cdots \sum_{i_K=1}^{n_K} U(i_2, i_3, \ldots, i_K).$$

Indeed, for specified values of $i_1, i_2, \ldots, i_K$ the quantity $v(L_{1,i_1}, L_{2,i_2}, \ldots, L_{K,i_K})$ appears exactly once as the i_1-st term in the representation of $U(i_2 - i_1, i_3 - i_1, \ldots, i_K - i_1)$, where indices wrap around as described above. By construction, each $U(i_2, i_3, \ldots, i_K)$ is the average of n_1 i.i.d. random variables. It follows from (4.3) and Proposition 2 that

$$P\{\tilde{Y}_{\mathbf{n}} - \mu \geq t\} \leq \inf_{h \geq 0} E\left[e^{h(\tilde{Y}_{\mathbf{n}} - \mu - t)}\right] \leq e^{-2nt^2}.$$

A symmetric argument establishes the same bound for $P\{\mu - \tilde{Y}_{\mathbf{n}} \geq t\}$, and (1.12) follows.

We can now use (1.12) to establish (1.13) in the same way that Hoeffding uses (1.5) to establish (1.6). First suppose that $n_k < m_k$ for $1 \le k \le K$, and define $m = m(\mathbf{n})$ as above. Observe that m is also equal to $m'(\mathbf{n})$ in this case. Since the function $g(x) = e^{hx}$ is convex for $h \ge 0$, it follows from (4.3) and Theorem 1 that

$$P\left\{ \tilde{Y}' - \mu \ge t \right\} \le E\left[e^{h(\tilde{Y}'_{\mathbf{n}} - \mu - t)}\right] = e^{-h(\mu+t)} E\left[e^{h\tilde{Y}'_{\mathbf{n}}}\right] \le e^{-h(\mu+t)} E\left[e^{h\tilde{Y}_{\mathbf{n}}}\right]$$

for $h \ge 0$, so that, by Proposition 2,

$$P\left\{ \tilde{Y}' - \mu \ge t \right\} \le \inf_{h \ge 0} E\left[e^{h(\tilde{Y}_{\mathbf{n}} - \mu - t)}\right] \le e^{-2nt^2}.$$

A symmetric argument establishes the same bound for $P\left\{ \mu - \tilde{Y}'_{\mathbf{n}} \ge t \right\}$, and (1.13) follows. Now suppose without loss of generality that for some integer $r = r(\mathbf{n})$ with $1 \le r < K$ we have $n_k < m_k$ for $1 \le k \le r$ and $n_k \ge m_k$ for $r+1 \le k \le K$. Observe that we can write

$$\tilde{Y}'_{\mathbf{n}} = \frac{1}{n_1 n_2 \cdots n_r} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_r=1}^{n_r} w_{\mathbf{n}}(L'_{1,i_1}, L'_{2,i_2}, \ldots, L'_{r,i_r}),$$

where $w_{\mathbf{n}}$ is defined as in (1.14). The desired result follows by applying the previous argument to the r-dimensional cross-product average of the function $w_{\mathbf{n}}$. The only case not considered so far is when $n_k \ge m_k$ for $1 \le k \le K$, but for this case the desired result follows trivially. □

To prove Theorem 3, we need the following lemma.

Lemma 1. *Suppose that there exist real numbers Y_i, μ_i, ϵ_i $(i = 1, 2)$ such that $\epsilon_1, \epsilon_2 > 0$, $Y_2 > \epsilon_2$ and $Y_i - \epsilon_i \le \mu_i \le Y_i + \epsilon_i$ for $i = 1, 2$. Then*

$$\left| \frac{\mu_1}{\mu_2} - \frac{Y_1}{Y_2} \right| \le \frac{Y_2\epsilon_1 + |Y_1|\epsilon_2}{Y_2(Y_2 - \epsilon_2)}.$$

Proof. We have

$$\left| \frac{\mu_1}{\mu_2} - \frac{Y_1}{Y_2} \right| = \left| \frac{Y_2\mu_1 - Y_1\mu_2}{Y_2\mu_2} \right| \le \frac{|Y_2\mu_1 - Y_1\mu_2|}{Y_2(Y_2 - \epsilon_2)}.$$

The set $A = \{ Y_1 - \epsilon_1, Y_1 + \epsilon_1 \} \times \{ Y_2 - \epsilon_2, Y_2 + \epsilon_2 \}$ contains the value of (μ_1, μ_2) that maximizes the numerator of the rightmost term. Observe that $|Y_2\mu_1 - Y_1\mu_2| \le Y_2\epsilon_1 + |Y_1|\epsilon_2$ for any $(\mu_1, \mu_2) \in A$. □

Proof of Theorem 3. By (1.19), we have

$$P\left\{ \left| \tilde{Y}'_{\mathbf{n}}(h) - \mu(h) \right| \le (b_h - a_h)\gamma_{\mathbf{n},p} \right\} \ge (1+p)/2$$

for $h = f, g$. It then follows from Bonferroni's inequality (see Miller [16, p. 8]) that

$$P\left\{ \left| \tilde{Y}'_{\mathbf{n}}(h) - \mu(h) \right| \le (b_h - a_h)\gamma_{\mathbf{n},p} \text{ for } h = f, g \right\} \ge p.$$

The desired result now follows from Lemma 1. □

Proof of Proposition 1. Set $Y_i = v(L_i)$ and $Y_i' = v(L_i')$ for $1 \le i \le n$. Throughout, we use the algebraic identities

$$Z_n = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} Y_i Y_j$$

and

$$Z_n' = \frac{m-1}{m}\left(\frac{1}{n'}\sum_{i=1}^{n'}(Y_i')^2 - \frac{2}{n'(n'-1)}\sum_{i=1}^{n'}\sum_{j=i+1}^{n'} Y_i' Y_j'\right),$$

where $n' = n \wedge m$.

Fix a convex function f and assume first that $n \ge m$. Using Jensen's inequality (see [1, p. 283]) and the fact that $Z_n' \equiv \sigma^2$ when $n \ge m$, we have

$$E\left[f(Z_n')\right] = f(\sigma^2) = f\left(E\left[Z_n\right]\right) \le E\left[f(Z_n)\right].$$

Now assume that $n < m$. We establish the desired inequality by mimicking the argument in Section 6 of [11]. For $y = (y_1, y_2, \ldots, y_n) \in \Re^n$, set

$$z_n(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} y_i y_j$$

and

$$z_n'(y) = \frac{m-1}{m} z_n(y).$$

Thus, $Z_n = z_n(Y)$ and $Z_n' = z_n'(Y')$, where $Y = (Y_1, Y_2, \ldots, Y_n)$ and $Y' = (Y_1', Y_2', \ldots, Y_n')$. Also set

$$Q_n = \left\{\, q = (q_1, q_2, \ldots, q_n) \in \{0, 1, \ldots, n\}^n \colon\ q_1 + q_2 + \cdots + q_n = n \,\right\}$$

and denote by Π_n the set of permutations of $\{\, 1, 2, \ldots, n \,\}$. For $\pi \in \Pi_n$ and $y = (y_1, y_2, \ldots, y_n) \in \Re^n$, we abuse notation slightly and denote the vector $(y_{\pi(1)}, y_{\pi(2)}, \ldots, y_{\pi(n)})$ by $\pi(y)$. Finally, for $y = (y_1, y_2, \ldots, y_n) \in \Re^n$ and $q = (q_1, q_2, \ldots, q_n) \in Q_n$, set $\widehat{z}_n(q, y) = z_n(\widehat{y})$, where $\widehat{y}_i = y_1$ for $1 \le i \le q_1$, $\widehat{y}_i = y_2$ for $q_1 < i \le q_1 + q_2$, and so forth. For example, when $n = 5$ and $q = (2, 0, 2, 1, 0)$, we have $\widehat{z}_n(q, y) = z_n(y_1, y_1, y_3, y_3, y_4)$. Similarly to Equation (6.6) in [11], we can write

$$E\left[f(Z_n)\right] = E\left[f\left(z_n(Y)\right)\right] = E\left[\overline{g}_n(Y'; f)\right], \tag{4.4}$$

where $\overline{g}_n$ is a function of the form

$$\overline{g}_n(y; f) = \sum_{q \in Q_n}\sum_{\pi \in \Pi_n} p(q, \pi) f\left(\widehat{z}_n\left(q, \pi(y)\right)\right)$$

with each coefficient $p(q,\pi)$ independent of f and

$$\sum_{q\in Q_n}\sum_{\pi\in\Pi_n} p(q,\pi) = 1.$$

The representation in (4.4) can be interpreted as follows: a realization y of the random vector Y can be obtained by generating a realization y' of the random vector Y', permuting the components of y' according to a randomly selected permutation $\pi \in \Pi_n$, and then forming a new vector of length n by replacing each component of $\pi(y')$ by 0 or more copies of the component in accordance with a randomly selected vector $q \in Q_n$. The quantity $p(q,\pi)$ is the probability that the components of y' are permuted according to π and then the components of $\pi(y')$ are duplicated in accordance with q.

Symmetry considerations imply that there exist numbers $\{\, p(q)\colon\ q \in Q_n \,\}$ such that $p(q,\pi) = p(q)$ for $\pi \in \Pi_n$ and $q \in Q_n$, so that

$$\overline{g}_n(y;f) = \sum_{q\in Q_n} p(q)\left(\sum_{\pi\in\Pi_n} f\left(\widehat{z}_n\big(q,\pi(y)\big)\right)\right).$$

For fixed $q = (q_1, q_2, \ldots, q_n) \in Q_n$ and $y = (y_1, y_2, \ldots, y_n) \in \Re^n$, observe that

$$\begin{aligned}
&\sum_{\pi\in\Pi_n} \widehat{z}_n\big(q,\pi(y)\big) \\
&\quad= \sum_{\pi\in\Pi_n}\left(\frac{1}{n}\sum_{i=1}^n q_i y_{\pi(i)}^2 - \frac{2}{n(n-1)}\sum_{i=1}^n\sum_{j=i+1}^n q_i q_j y_{\pi(i)} y_{\pi(j)} - \frac{2}{n(n-1)}\sum_{i=1}^n \frac{q_i(q_i-1)}{2} y_{\pi(i)}^2\right) \\
&\quad= (n-1)!\sum_{i=1}^n y_i^2 - \left(\frac{4(n-2)!}{n(n-1)}\sum_{i=1}^n\sum_{j=i+1}^n q_i q_j\right)\sum_{i=1}^n\sum_{j=i+1}^n y_i y_j \\
&\qquad\qquad - \left(\frac{(n-1)!}{n(n-1)}\sum_{i=1}^n q_i(q_i-1)\right)\sum_{i=1}^n y_i^2 \\
&\quad= \frac{n!}{n-1}\left(n - \frac{1}{n}\sum_{i=1}^n q_i^2\right) z_n(y).
\end{aligned}$$

Denoting the identity function by h, we see that $\overline{g}_n(y;h) = c z_n(y)$ for some constant c. It follows from (4.4) and the unbiasedness of both Z_n and Z_n' for σ^2 that

$$E\left[\overline{g}_n(Y';h)\right] = E\left[z_n(Y)\right] = \frac{m-1}{m} E\left[z_n(Y')\right],$$

so that $c = (m-1)/m$ and thus $\overline{g}_n(y;h) = z_n'(y)$. Using Jensen's inequality, we have

$$\begin{aligned}
\overline{g}_n(y;f) &= \sum_{q\in Q_n}\sum_{\pi\in\Pi_n} p(q,\pi) f\Big(\widehat{z}_n\big(q,\pi(y)\big)\Big) \\
&\geq f\left(\sum_{q\in Q_n}\sum_{\pi\in\Pi_n} p(q,\pi)\widehat{z}_n\big(q,\pi(y)\big)\right) \\
&= f\big(\overline{g}_n(y;h)\big) \\
&= f\big(z_n'(y)\big).
\end{aligned}$$

Since y is arbitrary,

$$E\left[f(Z_n)\right] = E\left[\overline{g}_n(Y';f)\right] \geq E\left[f\big(z_n'(Y')\big)\right] = E\left[f(Z_n')\right],$$

and the desired result follows. □

Proof of Theorem 4. Fix $t > 0$ and $n \in \{2,3,\ldots,m\}$. We have

$$P\left\{ Z_n' - \sigma^2 \geq t \right\} \leq e^{-h(\sigma^2+t)} E\left[e^{hZ_n'}\right] \leq e^{-h(\sigma^2+t)} E\left[e^{hZ_n}\right]$$

for $h > 0$, where the first inequality follows from (4.3) and the second inequality follows from Proposition 1. Krafft and Schmitz [13] show that

$$\inf_{h\geq 0} e^{-h(\sigma^2+t)} E\left[e^{hZ_n}\right] \leq e^{-8\lfloor n/2\rfloor t^2/(b-a)^4},$$

so that

$$P\left\{ Z_n' - \sigma^2 \geq t \right\} \leq e^{-8\lfloor n/2\rfloor t^2/(b-a)^4}.$$

A symmetric argument establishes the same bound for $P\left\{ \sigma^2 - Z_n' \geq t \right\}$, and the inequality in (1.38) follows. An argument as in (1.35) then establishes (1.39). □

Proof of Theorem 5. Fix $\tau > 0$, $n \in \{2,3,\ldots,m\}$, and $k \in \{2,3,\ldots,|S| \wedge n\}$. Since for any random variable $0 \leq X \leq 1$ we have

$$\mathrm{Var}\left[X\right] = E\left[X^2\right] - E^2\left[X\right] \leq E\left[X\right] - E^2\left[X\right] \leq \max_{0\leq u\leq 1}(u-u^2) = \frac{1}{4}, \tag{4.5}$$

it follows that

$$\sigma^2(S) \leq \frac{(b-a)^2}{4} \leq \frac{|S|(b-a)^2}{4(|S|-1)} \leq \frac{d(b-a)^2}{4(d-1)}.$$

Using this result, Theorem 4, and an argument as in the derivation of (1.29), we find that

$$\begin{aligned}
&P\left\{ Z_n''(S) - \sigma^2(S) \geq \frac{d}{d-1}\tau + \frac{d(b-a)^2}{4(d-1)^2} \;\middle|\; I_n' = k \right\} \\
&\quad \leq P\left\{ Z_n''(S) - \sigma^2(S) \geq \frac{d}{d-1}\tau + \frac{\sigma^2(S)}{(d-1)} \;\middle|\; I_n' = k \right\} \\
&\quad = P\left\{ \frac{d-1}{d} Z_n''(S) - \sigma^2(S) \geq \tau \;\middle|\; I_n' = k \right\} \\
&\quad \leq P\left\{ \frac{|S|-1}{|S|} Z_n''(S) - \sigma^2(S) \geq \tau \;\middle|\; I_n' = k \right\} \\
&\quad \leq e^{-8\lfloor k/2 \rfloor \tau^2/(b-a)^4}
\end{aligned}$$

Similarly,

$$\begin{aligned}
&P\left\{ Z_n''(S) - \sigma^2(S) \leq -\frac{d}{d-1}\tau - \frac{d(b-a)^2}{4(d-1)^2} \;\middle|\; I_n' = k \right\} \\
&\quad \leq P\left\{ Z_n''(S) - \sigma^2(S) \leq -\tau \;\middle|\; I_n' = k \right\} \\
&\quad \leq P\left\{ \frac{|S|-1}{|S|} Z_n''(S) - \sigma^2(S) \leq -\tau \;\middle|\; I_n' = k \right\} \\
&\quad \leq e^{-8\lfloor k/2 \rfloor \tau^2/(b-a)^4}
\end{aligned}$$

Thus,

$$P\left\{ |Z_n''(S) - \sigma^2(S)| \geq \frac{d}{d-1}\tau + \frac{d(b-a)^2}{4(d-1)^2} \;\middle|\; I_n' = k \right\} \leq 2e^{-8\lfloor k/2 \rfloor \tau^2/(b-a)^4}$$

and (1.41) follows directly. The inequality in (1.42) now follows by an argument as in (1.35). □

Acknowledgements

The author wishes to thank Joe Hellerstein, whose work helped motivate this paper. The author also benefited from comments by Hadas Shachnai and an anonymous referee of an earlier version of this paper.

References

[1] P. Billingsley, "Probability and Measure," second ed., Wiley, New York, 1986.

[2] W. G. Cochran, "Sampling Techniques," third ed., Wiley, New York, 1977.

[3] H. Cramér, "Mathematical Methods of Statistics," Princeton University Press, Princeton, New Jersey, 1946.

[4] S. Ganguly, P. B. Gibbons, Y. Matias, and A. Silberschatz, Bifocal sampling for skew-resistant join size estimation, *in* "Proc. 1996 ACM SIGMOD Intl. Conf. Managment of Data," pp. 271–281, ACM Press, 1996.

[5] P. J. Haas, Large-sample and deterministic confidence intervals for online aggregation, *in* "Proc. Ninth Intl. Conf. Scientific and Statist. Database Management," pp. 51–63, IEEE Computer Society Press, 1997.

[6] P. J. Haas and J. M. Hellerstein, Ripple joins for online aggregation, *in* "Proc. 1999 ACM SIGMOD Intl. Conf. Managment of Data," pp. 287–298, ACM Press, 1999.

[7] P. J. Haas, J. F. Naughton, S. Seshadri, and A. N. Swami, Selectivity and cost estimation for joins based on random sampling, *J. Comput. System Sci.* **52** (1996), 550–569.

[8] P. J. Haas, J. F. Naughton, and A. N. Swami, On the relative cost of sampling for join selectivity estimation, *in* "Proc. Thirteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems," pp. 14–24, ACM Press, 1994.

[9] J. M. Hellerstein, P. J. Haas, and H. J. Wang, Online aggregation, *in* "Proc. 1997 ACM SIGMOD Intl. Conf. Managment of Data," pp. 171–182, ACM Press, 1997.

[10] W. Hoeffding, A class of statistics with asymptotically normal distribution, *Ann. Math. Statist.* **19** (1948), 293–325.

[11] W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.* **58** (1963), 13–30.

[12] W. Hou, G. Ozsoyoglu, and B. Taneja, Processing aggregate relational queries with hard time constraints, *in* "Proc. 1989 ACM SIGMOD Intl. Conf. Managment of Data," pp. 68–77, ACM Press, 1989.

[13] O. Krafft and N. Schmitz, A note on Hoeffding's inequality, *J. Amer. Statist. Assoc.* **64** (1969), 907–912.

[14] R. J. Lipton and J. F. Naughton, Query size estimation by adaptive sampling, *in* "Proc. Ninth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems," pp. 40–46, ACM Press, 1990.

[15] R. J. Lipton, J. F. Naughton, and D. A. Schneider, Practical selectivity estimation through adaptive sampling, *in* "Proc. 1990 ACM SIGMOD Intl. Conf. Managment of Data," pp. 1–11, ACM Press, 1990.

[16] R. G. Miller, Jr., "Simultaneous Statistical Inference," second ed., Springer-Verlag, New York, 1981.

[17] P. K. Pathak, Sufficiency in sampling theory, *Ann. Math. Statist.* **35** (1964), 795–808.

[18] P. K. Pathak, An extension of an inequality of Hoeffding, *Studia Sci. Math. Hungar.* **10** (1975), 73–74.

[19] S. Seshadri, "Probabilistic Methods in Query Processing," Ph.D. Dissertation, Department of Computer Sciences, University of Wisconsin, Madison, Wisconsin, 1992.