Sample-Based Quality Estimation of Query Results in Relational Database Environments

Donald P. Ballou, InduShobha N. Chengalur-Smith, and Richard Y. Wang

Abstract—The quality of data in relational databases is often uncertain, and the relationship between the quality of the underlying base tables and the set of potential query results, a type of information product (IP), that could be produced from them has not been fully investigated. This paper provides a basis for the systematic analysis of the quality of such IPs. This research uses the relational algebra framework to develop estimates for the quality of query results based on the quality estimates of samples taken from the base tables. Our procedure requires an initial sample from the base tables; these samples are then used for all possible information IPs. Each specific query governs the quality assessment of the relevant samples. By using the same sample repeatedly, our approach is relatively cost effective. We introduce the Reference-Table Procedure, which can be used for quality estimation in general. In addition, for each of the basic algebraic operators, we discuss simpler procedures that may be applicable. Special attention is devoted to the Join operation. We examine various, relevant statistical issues, including how to deal with the impact on quality of missing rows in base tables. Finally, we address several implementation issues related to sampling.

Index Terms—Data quality, database sampling, information product, relational algebra, quality control.

1 Introduction

The quality of any information product—the output from an information system that is of value to some user—is dependent upon the quality of data used to generate it. Clearly, decision makers who require a certain quality level for their information products (IPs) would be concerned about the quality of the underlying data. Oftentimes, decision makers may desire to go beyond the preset, standard collection of queries implemented using reporting and data warehousing tools. However, since one cannot predict all the ways decision makers will combine information from base tables, it is not possible a priori to specify data quality requirements for the tables when designing the database.

In this paper, all IPs are the results of queries applied to relational tables, and we use the term *information product* in this sense. In the context of ad hoc IPs generated from multiple base tables, this work provides managers and decision makers with guidelines as to whether the quality of the base tables is sufficient for their needs. The primary contribution of this paper is the application of sampling procedures to the systematic study of the quality of IPs generated from relational databases using combinations of relational algebra operators. The paper yields insights as to how quality estimates for the base tables can be used to provide quality estimates for IPs generated from these base tables. Thus, it is neither necessary nor useful to inspect entire base tables.

- *D.P. Ballou and I.N. Chengalur-Smith are with the Management Science and Information Systems, University at Albany, Albany, NY 12222. E-mail: {d.ballou, shobha}@albany.edu.*
- *R.Y. Wang is with the MIT Information Quality Program MIT E53-320, Sloan School of Management, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142. E-mail: rwang@mit.edu.*

Manuscript received 20 Jan. 2005; revised 1 Aug. 2005; accepted 1 Nov. 2005; published online 17 Mar. 2006.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0028-0105.

Our approach is to take samples from each of the base tables, determine any deficiencies with the data in these samples, and use that information in the context of any given, specific IP, ad hoc or otherwise, to estimate the quality of that IP. Thus, sampling is carried out only once or on some predetermined periodic basis. Since there is an almost unlimited number of potential IPs, a major advantage of our approach is that only the base tables need to be sampled. The relevance of the deficiencies identified in the samples from the base tables is context dependent, i.e., the relevance of a particular deficiency depends upon the IP in question. Thus, the quality measure used for a given base table will vary according to its use.

We examine each of the relational algebra operators as generators of IPs and describe problems that could occur. A general procedure is introduced to overcome these problems, and this and other procedures allow practitioners to estimate the quality of IPs in relational database environments. These procedures have increasing levels of complexity, the choice of which one to use being dependent upon the level of analysis desired.

Section 2 develops the basic framework needed to ascertain the quality of IPs that result from applying the relational algebra operations to base tables. We present and discuss the issues that must be addressed in this context. Section 3 contains material needed to assess the quality implications of applying the relational algebra operators. Section 4 considers the special case of Join Operations. Section 5 presents material regarding sampling that is needed to apply the concepts developed in Sections 2, 3, and 4. The final section contains concluding remarks.

1.1 Related Research

Auditing of computer-based systems is a standard accounting activity with an extensive literature regarding traditional auditing of computer-based systems (e.g., O'Reilly et al. [21], Weber, [30]). But, such work does not tie the

TABLE 1
The Five Orthogonal, Algebraic Operations

Operation	*Purpose*
Restriction (Selection)	Returns a table containing all rows from a specified table that satisfy a specified condition
Projection	Returns a table containing all unique rows retaining only specified columns of the original table
(Cartesian) Product	Returns a table containing all rows formed by every possible combination of rows, one from each of two specified tables
Union	Returns a table containing all possible unique rows that appear in either or both of two specified tables
Difference	Returns a table containing all rows appearing in the first but not the second of two specified tables

quality of the data to the full set of possible IPs that could be generated from that data. There are, however, several works that address problems similar to the ones we consider. A paper that also addresses Join (aggregation) queries using statistical methods is the work by Haas and Hellerstein [12]. Their emphasis is on the performance of query join algorithms in the context of acceptable preciseness of query results. Work by Naumann et al. [20] examines the problem of merging data to form IPs and presents practical metrics to compare sources with respect to completeness. However, their focus is on comparing and contrasting different sources that describe the same or similar entities. Our work, on the other hand, is in the context of multiple tables for disparate entities. Other related works include that of Motro and Rakov [19] and Parssian et al. [23], [24], whose approach tends to be of a more theoretical nature than this work. More specifically, their work is in the context of fixed, known error rates. whereas we address the problem of estimating unknown error rates and dealing with the resulting uncertainty.

Other relevant work includes that of Scannapieco and Batini [28], who examine the issue of completeness in the context of a relational model. In addition, Little and Misra [17] examine various approaches to ensuring the quality of data found in databases. They emphasize the need for effective management controls to prevent the introduction of errors. By contrast, our aim is to estimate the quality of the base tables in the database and then use that information to examine the implications for the quality of any IPs. Acharya et al. [1], [2] use an approach analogous to ours to provide estimates for aggregates (e.g., sums and counts) generated by using samples from base tables rather than the base tables themselves. These papers extend a stream of work aimed at providing estimates for aggregates (e.g., Hellerstein et al. [13]). Their work deals more with sampling issues without addressing the quality of data units in the base tables, as is the focus of this paper. Sampling is also used to support result size estimation in query optimization which involves the use of some of the same statistical methods used here (cf. Mannino et al. [18]). Orr [22] discusses why the quality of data degrades in databases and the difficulty of maintaining quality in databases. In addition, he examines the role of users and system developers in maintaining an adequate level of data quality.

There are various studies documenting error rates for various sets of data. An early paper by Laudon [16] considers data deficiencies in criminal justice data and the implications of the errors. A more recent examination of data quality issues in data warehouse environments is found in Funk et al. [10]. Data quality problems arising from the input process are considerably subtler than simple keying errors and are discussed in Fisher et al. [9]. In addition, Raman et al. [25] describe the endemic problems that retailers have with inaccurate inventory records.

2 Quality of Information Products

We anchor our foundation for estimating the quality of IPs in the relational algebra, which consists of five orthogonal operations: Restriction, Projection, Cartesian product, Union, and Difference (Table 1); other operations (Join, Division, and Intersection) can be defined in terms of these operations (e.g., Klug [15]). For example, Join is defined as a Cartesian product followed by Restriction (e.g., Rob and Coronel [26]).

The focus of this paper is the development of estimates for the quality of IPs generated from relational base tables. As used in this paper, an IP is the output produced by some combination of relational algebraic operations applied to the base tables, which are assumed to be in BCNF. Thus, in this context, an IP is a table. In general, IPs may well involve computations, and the number of ways in which data can be manipulated is almost limitless. Rather than addressing some subset of such activities, we focus on the underlying, fundamental relational database operations, which usually would be implemented through SQL.

This work assumes that the quality of the base tables is not known with precision or certainty, a consequence of the large size and dynamic nature of some of the base tables found in commercial databases. In addition, we posit that the desired quality of the base tables cannot be specified at the design stage due to uncertainty as to how the database will be used, as would be the case, for example, with databases used to generate ad hoc IPs that support decision making.

Our approach to evaluating quality of IPs in a database environment involves taking samples from base tables. The quality of these samples is determined, i.e., all deficiencies (errors) in the sample data are identified. In general, determining all deficiencies is not a trivial task, but it is eased by using relatively small samples, what we call pilot samples below. Information on the quality of the samples is used to estimate the quality of IPs derived from the base tables. This methodology uses one sample from each base table to estimate the quality of any IP. For some IPs, the deficiencies may be material, for others, not. Hence, from

the same sample, the error rate as applied to one IP may differ from that for another IP. Thus, it is important to keep in mind that although only one sample is used, that sample may need to be examined separately in the context of each IP to determine the quality of the data items in the context of that product.

2.1 Basic Framework and Assumptions

This section contains material that forms the basis for our approach and justifies the concepts and approach found in this paper.

Definition. *A* data unit *is the base table level of granularity used in the analysis.*

Thus, for our purpose, a data unit would be either a cell (data element), a collection of cells, or an entire row (record) of a relational table. In terms of granularity, the analyst could operate at any of several levels within the row. Since the relational algebra is in the context of the relational model, which requires unique identifiers for each row of the base tables, any data unit must also possess this property. This implies that to determine quality at the cell level of granularity, the appropriate primary key is conceptually attached to the cell. A consequence is that, in the relational context, the data unit must be a subset of a row or, of course, the entire row. Since all the algebraic operations produce tables, we refer to the primary element of the result of any of the algebraic operations as a row. The determination of the appropriate level of granularity (for the base tables) is context dependent.

For the purposes of this paper, the labels *acceptable* and *unacceptable* are used to capture or represent the quality of the data units.

Definition. *A data unit is deemed to be* acceptable *for a specific IP if it is fit for use in that IP. Otherwise, the data unit is labeled as* unacceptable.

The determination of when a data unit should be labeled as acceptable is context dependent and also depends on the quality dimension of interest. Regarding context dependency, the same data unit may well be acceptable for some IPs but unacceptable for others. For example, the stock price quotes found in today's *Wall Street Journal* are perfectly acceptable for a long-term investor, but are unacceptably out-of-date for a day trader. As discussed below, when the samples taken from base tables are examined, all deficiencies with a particular data unit are recorded. Whether to deem this data unit as acceptable or not depends upon the particular IP in which it will be used.

Regarding quality dimensional dependency, we use *acceptable* in a generic sense to cover each of the relevant data quality dimensions (such as completeness, accuracy, timeliness, and consistency) or some combination of the dimensions. With the extant data, most quality dimensions can be evaluated directly. An exception is completeness and, in Section 5.2, we examine the issue of missing data. (A full examination of the dimensions of data quality can be found in Wang and Strong [29].) In practice, for a given IP, a data unit could be acceptable on some data quality dimensions and unacceptable on others. This leads to the issue of tradeoffs on data quality dimensions (cf., for example, Ballou and Pazer [4]), an issue not explored here due to space limitations.

Definition. *The* measure of the quality of an IP *is the number of acceptable data units found in the IP divided by the total number of data units.*

This measure will always be between 0 and 1. If the IP is empty and there should not be any rows, then the quality measure would be 1; if it should have at least one row, the quality measure would be 0.

Various issues arise with NULL values. If NULL is simply a placeholder for a value that is not applicable, then the NULL value does not adversely affect the acceptability of the data unit. If, on the other hand, NULL signifies a missing value, then it may impact the data unit's acceptability. Context would help determine which is the case, a task that, as indicated, could require some effort. It presumably would be unacceptable for an entire row to be missing, an issue addressed in Sections 5.2 and 5.3.

Inheritance Assumption. *If a data unit is deemed to be unacceptable for a specified IP, any row containing that data unit (for the same IP) would also be unacceptable.*

An implication of removing a deficient data unit is that if it is not required for the IP, then the result could be considered acceptable. Another implication is that when multiple rows are combined, as is the case with the Cartesian product, one unacceptable component causes the entire resulting row to be unacceptable.

Granularity Assumption. *The level of granularity used to evaluate acceptability in the base table needs to be sufficiently fine so that applying any algebraic operation produces data units that can be labeled acceptable or not.*

If, for example, a row level of granularity is used in the base table, then any projection other than the entire row would produce a result whose acceptability or unacceptability could not be determined knowing the values from the base table.

Error Distribution Assumption. *The probability of a data unit being acceptable is constant across the table, and the acceptability of a data unit is independent of the acceptability of any other.*

For relationship tables, the situation may be more complex, and the material found at the end of Section 3.1 would be applicable. This error distribution assumption is definitely not true for columns. For example, some columns may be more prone to missing values than others. When dealing with projection, we limit the evaluation of the deficiencies found in the samples taken from the base tables to the columns of interest. Details are given in Section 5.

Note that we are not concerned with the magnitude of the error, rather, only with whether an error exists that makes the data unit unacceptable for its intended use in a specific IP. Thus, this work makes no normality assumption regarding the errors in the data. We now consider several issues that arise when considering the quality of IPs.

Given an IP (table), the distribution of errors and the level of granularity can impact substantially the measure for the quality of that IP. For instance, if there is at least one unacceptable cell in each row, then the quality measure for the row level of granularity is 0. (By the inheritance assumption, all rows in this case would be of unacceptable quality.) However, if all the unacceptable data units happen to be in the same row, then the quality measure for the row level of granularity would be close to 1 for large tables. It should be noted that the distribution of errors is relevant at the row level but not at the cell level, since, at the cell level, it is the number of unacceptable cells that is the issue, not the rows that they are in.

It may be surprising that base tables of very high quality can yield IPs of very poor quality and vice versa. To see this, suppose that, for a base table of n rows with large n, the level of granularity is row level, and that all rows are acceptable save one. (Hence, the quality of the base table is close to 1.) Suppose that a SELECT retrieves the unacceptable row and no other. The quality of the resulting IP is 0 even though the quality of the base table is arbitrarily close to 1. Similar reasoning justifies the converse. Thus, IPs derived from the same base tables can have widely differing quality measures due to the inherent variability in the quality of the base tables. Hence, knowing the quality of the base tables is not sufficient for knowing the quality of the IPs. This fact motivated our approach for estimating the quality of IPs.

Since knowledge of the quality of an IP is readily obtained, provided we have certain knowledge of the quality of the base tables, it would seem desirable to insist that the quality of the data units in the base tables be determined without uncertainty. However, since many base tables are very large and dynamic, determining with certainty the quality of all the data units in practice is difficult at best. This does **not** imply that we exclude statements regarding the overall quality of a base table. (Statements such as "the data in base table A are 99 percent correct" can be perfectly valid.) Rather, we acknowledge the impossibility of knowing a priori with certainty whether a randomly chosen data unit is acceptable or unacceptable. *This indicates that it is impossible to make definitive statements about the quality of IPs.*

Sample Quality Assumption. *It is possible to determine the quality of each data unit of a sample taken from a base table.*

In practice, it is not always possible to determine the quality of a data unit with certainty. The level of resources the organization commits to data quality assessment essentially determines the effectiveness of the evaluation. We treat the results of that evaluation as correct. If there is a concern that data units may have problems that have not been identified, one can always do sensitivity-type analysis to determine what impact unidentified but suspected deficiencies might have. The resulting information can then be passed on to decision makers.

In theory, the same assumption could apply to entire base tables. Should the base tables be large and dynamic, by the time all the data units had been checked to determine their quality (acceptable or unacceptable), the base table would be different enough so that information regarding quality would be outdated. For stable base tables, this would not be the case. Probably more important, the cost of checking the entire base table could be prohibitive. But, the Sample Quality Assumption implies that these issues do not apply for samples. In Section 5, we discuss issues regarding sampling in a database context, including how large the samples should be.

TABLE 2
Key Notations (in Order of Appearance in the Paper)

T_i	i^{th} base table in the database
N	Number of base tables in the database
I_j	j^{th} IP
M	Number of IPs
$T_{i,j}$	i^{th} base table involved in producing I_j
N(j)	Number of base tables involved in producing I_j
n_i	Size of sample from the i^{th} base table
$\Pi_{i,j}$	True proportion of acceptable data units in $T_{i,j}$
$\Pi(j)$	True proportion of acceptable rows in I_j (created via Cartesian product)
$P_{i,j}$	Estimate for $\Pi_{i,j}$
P(j)	Estimate for $\Pi(j)$
$L_{i,j}$ ($U_{i,j}$)	Lower (upper) limit for $P_{i,j}$
S_i	Sample from i^{th} base table
$S_{i,c}$	Corrected sample from i^{th} base table
A_k	Pre-specified, desired quality level for information product I_k

Under the Sample Quality Assumption, determining a data unit's quality does not result in any classification errors such as an acceptable item being labeled as unacceptable or, conversely, an unacceptable item being labeled as acceptable. This work assumes that, for a given IP the proportion of acceptable data units in the sample is known with certainty, and we then control for variation in the sample. If there are classification errors, the measure of the quality of the sample from the base table is uncertain due to uncertainty in the numerator of the proportion, which would lead to inefficient estimates. Given the small relative magnitude of inspection-induced errors and the fact that this work would be complicated substantially should fallible inspection be incorporated, we limit our work to the case of perfect inspection. Issues related to imperfect inspection can be found in Ballou and Pazer [3] and Klein and Goodhue [14].

As indicated earlier, the same base table can produce IPs of substantially different quality. Since the only quality measure we have is that of the sample, we cannot know with certainty the quality of all IPs. Thus, determining the quality of IPs on the basis of the quality of samples taken from the base tables can be done only in a statistical sense.

3 A Relational Auditing Framework

In this section, we introduce the Reference-Table Procedure, a general approach for estimating the quality of IPs. Next, we examine the five orthogonal operations and, in the following section, we present an in-depth examination of the quality implications of the Join operation.

For this work, let $T_1, T_2, \ldots, T_N$ represent the N base tables in the database. (See Table 2 for notation.) In general, there are multiple IPs, each of which can depend upon multiple base tables. To capture this, let the IPs be designated by $I_1, I_2, \ldots, I_M$. Let $T_{1,j}, T_{2,j}, \ldots, T_{N(j),j}$ represent the base tables that are involved in producing the IPI_j, where $N(j)$ is the number of tables required for I_j. Thus, the first subscript of $T_{i,j}$ identifies a particular table and the

second that that table is one of the tables used to form the jth IP. (Note that $T_{1,j}$ may or may not be the same as T_1.)

Let $\Pi_{i,j}$ represent the true (in general unknown) rate or proportion of acceptable data units in table $T_{i,j}$. The first task is to determine $P_{i,j}$, an estimate for $\Pi_{i,j}$. This is accomplished using a sample of size n_i from the appropriate base table T_i. Recall that this sample is taken independent of any particular information product. As discussed, each member of the sample would have to be examined to determine its quality (acceptability or unacceptability) in the context of information product I_j. Issues related to implementation in general and sampling, including sample size, in particular will be discussed in Section 5.

User requirements guide determination of the appropriate level of granularity and the acceptability of the data units. After completion of the evaluation for acceptability of the members of the sample S_i for information product I_j, the ratio of the number of acceptable items to the size of the sample would be formed, which yields a number $P_{i,j}$ that is used as the estimate for $\Pi_{i,j}$. (Note that $P_{i,j}$ would be the minimum variance unbiased estimator for $\Pi_{i,j}$.) As indicated, it is important to keep in mind that the value for $\Pi_{i,j}$ is a function of the intended IP. For one IP, the value of $\Pi_{i,j}$ might be high, and for another, low. The same sample S_i would be re-evaluated to estimate the $\Pi_{i,j}$ value for each IP.

The standard way of capturing the error in the estimate $P_{i,j}$ is via a $100(1-\alpha)\%$ confidence interval, which, for this context, can be represented by

$$L_{i,j} = P_{i,j} - z_{\alpha/2}s_{i,j} \leq \Pi_{i,j} \leq P_{i,j} + z_{\alpha/2}s_{i,j} = U_{i,j}. \quad (1)$$

Here, $s_{i,j} = (P_{i,j} * (1 - P_{i,j})/n_i)^{1/2}$ represents the standard error of the proportion $P_{i,j}$. This uses the normal approximation to the binomial for sufficiently large samples, an issue discussed in Section 5. For (1) to be valid, the only assumption needed is the Error Distribution Assumption. An excellent source for the statistical background required for our work is Haas and Hellerstein [12, pp. 292-293] and Mannino et al. [18, pp. 200-202].

In reality, $\Pi_{i,j}$ incorporates information not only on acceptable values but also on missing rows. If all data in a given table should be totally acceptable but only 80 percent of the records relevant for the IP in question that should be in the table actually are present, then $\Pi_{i,j} = 0.8$ should hold. Issues such as how to account for missing rows are deferred to Section 5.2.

3.1 Reference-Table Procedure

Before addressing each of the five orthogonal algebraic operations, we begin by introducing the Reference-Table Procedure, which can be applied to all IPs. We then examine each of the orthogonal algebraic operations in turn. For some of these, under certain conditions, simpler approaches are available, although realistically, for any complex IP, the Reference-Table procedure would be used.

The *Reference-Table Procedure* relies upon reference tables to compute quality measures. Essentially, this procedure compares an IP without error to an IP that may well contain unacceptable data and which is used as a surrogate for the "correct" IP. The question arises: How can one measure how well the surrogate IP approximates the "correct" IP? For this, we measure how well the surrogate IP attains the ideal of containing acceptable data units, and only acceptable data units, when viewed from the perspective of the "correct" IP.

To describe the Reference-Table Procedure, let $S_1, S_2, \ldots$ denote the samples from each of the base tables that are involved in creating the IP in question. Let $S_{1C}, S_{2C}, \ldots$ denote the corrected version of $S_1, S_2, \ldots$, respectively, where all deficiencies in the context of the IP have been removed. (Although this may be impossible to do for entire base tables, we believe that this is manageable for samples taken from the base tables.) Apply to $S_1, S_2, \ldots$ the steps required to generate the desired IP. Call the result Table 1. Table 1 is a subset of the IP that would be produced should the steps be applied to the appropriate base tables. Now apply the same steps to $S_{1C}, S_{2C}, \ldots$. Let Table 2 represent the result. Table 2 would be a subset of the desired IP should all deficiencies in the relevant base tables be eliminated. In other words, Table 2 is a sample from the correct (without error) IP. We refer to Table 2 as the reference table. Note that Table 2 could be larger or smaller than Table 1.

Definition. *The appropriate quality measure for Table 1 in the context of Reference Table 2 is*

$$\mathbf{Q(Table\ 1)} = |\mathbf{Table\ 1} \cap \mathbf{Table\ 2}|/\mathbf{max}\{|\mathbf{Table\ 1}|, |\mathbf{Table\ 2}|\}. \quad (2)$$

Here, the vertical lines represent the cardinality (number of data units) of the indicated set. We use Q (Table 1) as the measure of the quality of the IP that has Table 1 as a subset.

The expression for Q(Table 1) satisfies the minimum requirement of being a number between 0 and 1. Also, as required by our definition (see Section 2) for the quality of an IP, the numerator is the number of acceptable data units in Table 1 in the context of the reference table. For the denominator, we now proceed from specific cases to the general expression.

First, suppose that Table 1 $\subset$ Table 2. In this case, each data unit of Table 1 is acceptable, but Table 2, the reference table, contains data units not found in Table 1. Thus, $|\text{Table 1}|/|\text{Table 2}|$ is the appropriate measure of the quality of Table 1, which is (2) for this case. Next, suppose that Table 2 $\subset$ Table 1. Thus, Table 1 contains all the acceptable data units, but other, unacceptable ones as well. For this case, the quality of Table 1 is given by $|\text{Table 2}|/|\text{Table 1}|$, which captures the fact that even though Table 1 contains all the acceptable data units, its quality cannot be 1, as it also contains unacceptable data units. Again, this is (2) for this case. Next, suppose Table 1 = Table 2. Then, the numerator and denominator of (2) are the same, yielding the value 1, as expected. Last, consider the case that Table 1 $\cap$ Table 2 $\neq \emptyset$ and that neither is a proper subset of the other. As mentioned, the numerator of (2) yields the number of acceptable data units in Table 1. The first two cases given above required that the denominator be the larger of $|\text{Table 1}|$ and $|\text{Table 2}|$. Since, for this case, the situation could be arbitrarily close to one or the other of the first two cases, continuity considerations require that the max be used for this situation as well. For example, suppose

that Table 1 would be a subset of Table 2 should exactly one element of Table 1 be corrected. (Without that correction, Table 1 $\cap$ Table 2 $\neq \emptyset$ holds, as the incorrect data unit would not be in Table 2.) As was just discussed, with the unit corrected, expression (2) applies. With the data unit not corrected, one would expect that the value for the quality of Table 1 in reference to Table 2 would be close to that of the corrected case, especially if Table 1 is large. Since, with the data unit corrected, the max function needs to be used, continuity of quality values would require the use of the max function in the case with the data unit not corrected. A mathematical induction-type argument can be applied when there are n data units ($n > 1$) that need to be corrected.

We can build a confidence interval for Q(Table 1) by resampling the samples already drawn from the base tables. Since Q(Table 1) does not have a known distribution, a bootstrapping procedure, which involves sampling with replacement from the original samples, would be appropriate (Efron and Tibshirani, [7]). Q(Table 1) could be calculated for each resample and the set of resulting values could be used to build a confidence interval. By using the initial samples drawn from the base tables as our universe, we can create confidence intervals with little additional effort.

It should be noted that the Reference-Table Procedure is robust and can be applied when various assumptions break down. For example, the Error Distribution Assumption may fail due to lack of independence in the rows of relationship tables and, hence, (1) may not be applicable. However, we can still estimate the quality of the table using the Reference-Table Procedure.

In practice, obtaining a sample is easily implemented, as commercial DBMS packages do this. The chief impediment to applying the Reference-Table Procedure is in determining what precisely the contents of the Reference Table should be, as detecting and correcting errors in data are notoriously difficult. However, this is an issue that has been addressed by the information systems and accounting profession for decades; see Klein and Goodhue [14], Little and Mishra [17], and the references given in them.

3.2 Basic Relational Algebraic Operations

We now consider special cases, some of which provide a more intuitive way to obtain estimates for the quality of the IP by examining, in turn, each of the fundamental algebraic operations applied to base tables. It should be kept in mind that, if multiple algebraic operations are involved in producing an IP, then, in all likelihood, the Reference-Table Procedure would need to be used.

3.2.1 Restriction

Should the Restriction operation be applied to a single table T_k, we use the sample taken from T_k to determine the acceptability (fitness for use) of each data unit of the sample for the IP in question. The fraction of acceptable data units is the estimate for Π_k. A confidence interval for Π_k is $L_k < \Pi_k < U_k$, where L_k and U_k are defined in (1). The more general case where the Restriction operation follows the Join of several tables would require the use of the analogous interval given by (6) in Section 3.2.5.

3.2.2 Projection

Should an IP be formed using the Projection operation, then we would use the appropriate subset of the sample taken from table T_k. Again, each data unit of the sample would be examined for acceptability, and the fraction of acceptable data units would be the estimate for P_k. If there are no duplicates, then this applies. If the projection does not include the primary key, then it is quite likely that duplicates will exist and the Reference-Table Procedure would have to be used. This is needed, as duplicates may be incorrectly retained or incorrectly deleted. Issues involving duplicates are discussed in more detail in the following presentation of the Union operation.

3.2.3 Union

Although the following material is in the context of the Union of two tables, it generalizes in the obvious manner to multiple tables.

No Duplicates Exist. This case is relatively straightforward if there are no duplicate data units. Suppose a table with N data units is combined via Union with a table with M data units. Let P_1 and P_2 represent the estimates, respectively, for the fraction of acceptable data units in the two tables, based on samples of size n_1 and n_2, respectively. An estimate of the fraction of acceptable data units in the union is $P = (n_1 * P_1 + n_2 * P_2)/(n_1 + n_2)$. The standard deviation of the fraction P is given by $s = ((P * (1 - P)/(n_1 + n_2))^{1/2}$. The confidence interval for the true fraction Π of acceptable data units in the union is given by:

$$\mathbf{L} = \mathbf{P} - \mathbf{z}_{\alpha/2}\mathbf{s} \leq \Pi \leq \mathbf{P} + \mathbf{z}_{\alpha/2}\mathbf{s} = \mathbf{U}. \quad (3)$$

Duplicates. Determining the quality of an IP when there are duplicate data units is considerably more difficult. In the union of two tables, duplicates can be incorrectly retained or incorrectly deleted. The Appendix illustrates issues similar to this for the case of Joins. There does not appear to be a simple method for analyzing the quality of IPs in such cases, making it necessary to use the Reference-Table Procedure.

3.2.4 Difference

At first glance, this case appears to be relatively straightforward. Suppose that the IP is formed via $T - S$, where T and S are tables, and let P be the estimate for the proportion of acceptable data units in T. Then, P is also the appropriate estimate for the quality of the IP, assuming that those data units in T that are not in S do not possess a different quality level than do those in both T and S. If one suspects that this assumption is not valid, then it would be necessary to sample from those data units found in T only. However, the situation is subtler as the resulting number of data units in the difference table could be fewer or greater than should be the case. The issues that arise are similar to that of Union with duplicates, and the Reference-Table Procedure can be applied.

3.2.5 Cartesian Product

The case of the Cartesian product is considerably more complex. At this stage, it would be appropriate to examine the product of just two tables. However, when we discuss the Join operation, it is necessary to work with many of

TABLE 3
Summary of Procedures

Algebraic Operation	*Procedure to evaluate quality of IP*
Restriction	Use P_k, the estimate for a sample from base table T_k
Union (if no duplicates exist)	Weighted average estimates of the two or more underlying tables
Union (duplicates deleted)	Reference table procedure
Projection (key included)	Use P_k
Projection (key excluded and duplicates deleted)	Reference table procedure
Difference	Reference table procedure
Cartesian Product	Product of the individual P_k values of the underlying tables

the same concepts in the context of n tables. To facilitate that discussion, we address at this time the more general n-table case.

The product of an *s*-data unit table with a *t*-data unit table is a table with s times t rows. By the Inheritance Assumption, a row of the product table is acceptable if, and only if, each of the component data units is acceptable. This concept generalizes naturally to n tables. Suppose that the information product I_j is formed via a Cartesian product of tables $T_{1,j}, \ldots, T_{N(j),j}$. Then, a row in I_j will be acceptable if, and only if, each of the data units that are concatenated to form the row is acceptable.

The fraction of acceptable values $\Pi(j)$ is given by multiplying together the proportion of acceptable values for the components, i.e.,

$$\Pi(\mathbf{j}) = \Pi_{1,\mathbf{j}} * \Pi_{2,j} * \ldots * \Pi_{\mathbf{N(j)},\mathbf{j}}. \quad (4)$$

It is important to note that $\Pi(j)$ is **not** a population parameter in the statistical sense. Rather, the validity of (4) is a direct consequence of applying the truth values for the logical **and** ($A \wedge B$) with **true** replaced with **acceptable** and **false** with **unacceptable**. Thus, a row in a Cartesian product is acceptable if, and only if, each of the components is acceptable. This is a direct consequence of the Inheritance Assumption. (It is important to note that since $\Pi(j)$ is not a statistic, independence is not relevant to multiplying the components.) Hence, the number of acceptable units in the n-fold Cartesian product is the product of the numerators of the Πs. (This can be established by induction on the number of tables.) The total number of rows of the Cartesian product is simply the product of the denominators of the Πs in (4). Thus, the fraction of acceptable data units is given by (4).

Recall that Π_k represents the true proportion of data units deemed to be acceptable in table T_k. Here, we use the notation $\Pi(j)$ to represent the true proportion of acceptable data units as found in I_j, which, in this context, is formed via a Cartesian product. If the number of terms in the right-hand side of (4) is sizable, then, unless all the $\Pi_{i,j}$ are close to 1, the product will not be large. Thus, it is very difficult for an IP to have a high acceptability value if it is formed via a Cartesian product using many tables.

Definition. *Our estimate for (4) is*

$$\mathbf{P}(\mathbf{j}) = \mathbf{P}_{\mathbf{1},\mathbf{j}} * \mathbf{P}_{\mathbf{2},\mathbf{j}} * \ldots * \mathbf{P}_{\mathbf{N(j)},\mathbf{j}}. \quad (5)$$

As before, $P_{i,j}$ represents the estimate for the true proportion of acceptable data units in $T_{i,j}$.

The validity of (5) depends upon exactly the same chain of reasoning used for (4). Since $P(j)$ is by definition an estimate for $\Pi(j)$, an expression for $\Pi(j)$ analogous to the confidence interval for Π given in (1) is:

$$\mathbf{L}(\mathbf{j}) < \mathbf{\Pi}(\mathbf{j}) < \mathbf{U}(\mathbf{j}), \quad (6)$$

where $L(j)$ is given by

$$\mathbf{L}(\mathbf{j}) = \mathbf{L}_{\mathbf{1},\mathbf{j}} * \mathbf{L}_{\mathbf{2},\mathbf{j}} * \ldots * \mathbf{L}_{\mathbf{N(j)},\mathbf{j}}$$

and $U(j)$ by

$$\mathbf{U}(\mathbf{j}) = \mathbf{U}_{\mathbf{1},\mathbf{j}} * \mathbf{U}_{\mathbf{2},\mathbf{j}} * \ldots * \mathbf{U}_{\mathbf{N(j)},\mathbf{j}}.$$

The probability that $\Pi(j)$ lies outside this interval will be developed below. It should be kept in mind that (6) is **not** a confidence interval for $\Pi(j)$ but rather gives an interval within which, with a certain probability, $\Pi(j)$ lies. We now discuss the computation of the probability that $\Pi(j) \geq U(j)$. The discussion for $\Pi(j) \leq L(j)$ follows the same reasoning.

The probability that $\Pi(j)$ exceeds $U(j)$ is simply the volume bounded by the unit cube and the surface $\Pi_{1,j} * \Pi_{2,j} * \ldots * \Pi_{N(j),j} = U_{1,j} * U_{2,j} * \ldots * U_{N(j),j}$. In general, this would be evaluated using a numerical integration package or via an $N(j)$-fold multiple integral. For the case $n = 2$, the probability (area) that $\Pi_{1,j} * \Pi_{2,j} > U_{1,j} * U_{2,j}$ is given by $1 - U_{1,j}U_{2,j} + U_{1,j}U_{2,j}\ln(U_{1,j}U_{2,j})$. This expression is obtained by evaluating the integral

$$\int_{U_{1,j}*U_{2,j}}^{1} \int_{U_{1,j}*U_{2,j}/\Pi_{1,j}}^{1} 1 \, d\Pi_{2,j} d\Pi_{1,j}. \quad (7)$$

This concludes our discussion of the five orthogonal algebraic operations. The approaches used to estimate the results of these fundamental operations are summarized in Table 3.

4 Join Operation

As most applications of normalized databases involve the use of multiple tables that are joined to produce other tables, the analysis of the quality of IPs involving Joins is perhaps the most critical issue that has not been fully explored. A simple example found in the Appendix illustrates this issue. Because of space limitations, we deal exclusively with inner joins; similar treatment applies to the other types of joins.

The basis for this material was developed for Cartesian products, but the special role of foreign keys complicates

matters substantially. As discussed, a row of an IP would be deficient if the row contains a segment from some table that is not fit for use. In the case of a Join over a nonforeign key field, problems arise whenever rows exist that should not or rows are missing from the IP that should be included. We first address Joins over foreign key fields. We then consider Joins over nonforeign key fields.

4.1 Join over Foreign Key

In this section, we consider the case for which unsuitable records in the IP arise as the result of a Join over a foreign key column. A row in the IP is unacceptable provided at least one of the rows joined to form the row in question is not acceptable. A Join is simply a Cartesian product followed by a Select, each of which we have examined in isolation earlier. The basic idea behind estimating the quality of a Join is as follows: A sample has been taken from each of the two base tables to be joined. As discussed, the quality of each table is estimated in the context of the particular Join. The product of these sample estimates is the estimate for the quality of the Cartesian product. We use that number for the quality of the join that arises when the appropriate Select is applied. Thus, in theory, the quality of a Join over a foreign key is simply the estimate for the quality of the Cartesian product.

The underlying assumption is that the random sampling process averages out atypical behavior. However, this assumption may not be valid in the case of Joins over foreign keys. If there is concern about the quality of the primary and foreign key columns, then it is necessary to resort to the reference table procedure, as is required for Joins over nonforeign keys, details of which follow.

4.2 Join over Nonforeign Key Attributes

When nonforeign key fields are involved in Joins, the situation is considerably less straightforward than the case considered above. To conceptualize the potential difficulties, suppose that Table X, say, with cardinality (number of rows) M is joined to Table Y with cardinality N, yielding Table Z. Then, the cardinality of Table Z can range from 0 to $M * N$. If none of the values in the joining column of Table X match those in the joining column of Table Y, then there would be no rows at all in Table Z. The other extreme arises when the values in each of the joining columns are all the same. Thus, a priori, the size of Table Z can vary considerably. It is possible that high error rates in the two joining columns can have very little or no impact on the error rates of the joined table. To see this, suppose that all values in the joining column in Table X are wrong except for one, and that one is the only value that matches values from the joining column in Table Y. Supposing that the matched rows in Table Y are acceptable, then the joined result, Table Z, will have no errors in spite of the fact that all but one of the values in the joining column of Table X are wrong. The converse situation, namely, that all the values are acceptable save one, and that one is the only matching value, would lead to a result for which all the rows are unacceptable in spite of a high correctness value for the joining columns. This wide variation in possible outcomes requires use of the Reference-Table Procedure.

Fig. 1 contains a summary of the steps of our methodology.

1. Sample each base table
2. Determine the deficiencies in each sample
3. Evaluate all relevant samples in the context of the given specific query
4. Calculate the appropriate L and U values to estimate the quality of the IP
5. Determine whether the quality of the IP meets the pre-specified quality level
6. For subsequent queries re-apply the procedure starting at Step 3.

Fig. 1. Steps of the methodology.

5 Sampling in the Relational Context

Statistical sampling is a well-established field and we draw on some of that work for this paper. Specifically, we make reference to the acceptance sampling procedures used in statistical quality control. Acceptance sampling plans are used to determine whether to accept or reject a lot. In particular, we consider double sampling plans, where the decision to accept or reject the lot is made on the basis of two consecutive samples. The first sample is of fixed size (smaller than the size of a comparable single sampling plan), and if the sample results fall between two predetermined thresholds for acceptance and rejection, then a second sample is taken. The combined sample results are then compared to the rejection threshold. Thus, in addition to being more efficient than single sampling plans, the double sampling plans have "the psychological advantage of giving a lot a second chance" (Duncan [6, p. 185]).

Our approach involves two rounds of sampling from the base tables. In the first round, a random sample is used to obtain an initial, if not especially precise, estimate for the true fraction of acceptable items in each table. The sample must be large enough to satisfy the sample size requirement, as discussed in Section 5.1. Database administrators, users, or other appropriate personnel need to determine what kinds of deficiencies would be sufficient to classify a data unit as unacceptable.

The fraction of acceptable items is used to identify those IPs that meet prespecified (desired) quality levels A_k set by the users of the IPs. If the prespecified quality level A_k is less than the lower limit of intervals such as (1) and (6), then there is strong evidence that whatever the true acceptability rate for the IP is, the true rate is greater than the desired or required quality level. Similarly, for those IPs with the prespecified quality estimate A_k greater than the upper limit of the appropriate interval (e.g., (1) or (6)), we can conclude with a high level of certainty that they do not meet the required quality level. Then, additional sampling is undertaken to determine which of the remaining IPs meet their required quality levels. For this, an approach is used to sample some tables more intensively than others. The goal is to ensure that the enhanced estimates for fraction of acceptable items will contribute most to resolving the remaining ambiguities as to whether or not the specified quality levels are achieved. Issues involved with this second round of sampling are discussed in Section 5.4. It should be kept in mind that, for both rounds of sampling, the sample taken from a particular table is used as part of the evaluation of all IPs that use that table. However, as indicated above, the quality of these samples is dependent upon its use in the IP. (This implies that certain deficiencies that have been identified may not be relevant for certain IPs.)

Although, for some IPs (such as those resulting from restrictions), only a subset of a base table is involved, we do not sample from such subsets, as such a sample would not apply to other IPs. In addition, a subset sample would no longer be random, and our approach relies upon random samples, which are required to avoid potentially serious biases. An exception to this is for the Projection operation, as the assumption of a constant probability of error may not hold across all the columns.

For the Restriction operation, the sample would be taken across the entire table, not just those rows identified by the Restriction condition. Assuming that deficiencies are randomly distributed across the entire table, the same fraction of acceptable items and confidence intervals would result for either case. There are situations, however, when this assumption is not valid. If, for example, the rows in some table are obtained from two different sources and these sources have significantly different error rates, then our Error Distribution Assumption requires that two tables be formed, one for each source. A separate sample would then be taken from each table and separate estimates for the acceptability rates would be generated. At any point, as needed, these two tables could then be combined using the union operation and, as was explained, an error rate for the combined table will be available.

The case for Projection is different. Here, the pilot sample would consist of rows containing only those columns specified by the restricting conditions. The reason is that it is more reasonable to assume homogeneity across rows, which have identical structure, than it is across columns, which inherently tend to have differing error rates.

5.1 Sample Size Issues

How large should the pilot sample be? For this, some rough idea of the underlying (true) error rate is required. This is especially true if the error rate is low, as is likely to be the case. If, for example, the true error rate is 1 percent and a sample of size 10 is taken, then only occasionally will an error show up in the sample. Under this circumstance, a large enough sample needs to be taken so that defective records appear in the sample. In auditing, discovery sampling is used when the population error rate is believed to be very small but critical. Similarly, in statistical quality control, procedures exist for detecting low levels of defects. A standard rule of thumb is that the sample should be large enough so that the expected value of the number of defective items is at least two (Gitlow et al. [11, pp. 229-231]). Since sampling (with replacement) is a binomial process, then n, the size of the sample, must satisfy the inequality $n \geq 2/(1-\Pi)$, where Π represents the true proportion of acceptable data units. Clearly, there needs to be some estimate for the value of Π in order to use this inequality. One way of estimating Π is by taking a preliminary sample before initiating the first round of sampling. If Π is close to 1, then a large sample size would be required. In any case, using just the minimum will prove to be of marginal value, i.e., not yield enough information to allow us to make a decision.

5.2 Missing Rows

Missing or Null values in a particular row could result in that row being labeled as unacceptable, but the assumption to this point has been that if a row should be in the table, then it indeed is. We now consider how to deal with rows that ought to be in a particular table and are not. A table can be such that the data in the existing rows are completely acceptable. Yet, if there are rows that should be in that table but are not present, then a consequence of the missing rows could be a series of deficient IPs.

Two issues need to be addressed. The first is to obtain an estimate for the number of rows that are missing from each table, and the second is to analyze the potential impact of these missing rows on the various IPs. The first issue is handled at this point using an approach employed by statisticians to address similar issues, such as census undercounts. The second is handled via a simulation approach in our discussion of the impact of missing rows on Joins in Section 5.3.

A standard technique used by statisticians to estimate the number of missing objects in a population is capture/recapture sampling. This procedure involves a two-round process. For the first round, a random sample is taken, the captured individuals are tagged, and this tagged sample is then mixed back into the population. At a later point in time, a second sample is taken. The number of tagged individuals in the second sample can be used to estimate the overall population size. If the recapture takes place during a short enough period of time that no additions to or removals from the population have taken place between the samples, then a closed statistical model can be used. The two major assumptions are: 1) a thorough mixing of the sample with the population and 2) the tagging has not affected recapture. This procedure is described in detail in Fienberg and Anderson [8] and the theory in Boswell et al. [5].

In applying these concepts to the determination of the number of missing records in a table, the main obstacle would lie in the capture (first) sample, which would have to be generated in a manner independent from the way the data are obtained for entry into the table in question. The capture sample consists of "tagged" records. The variable n_1 is the size of this independently generated sample. The members of that sample would then be examined to see which ones are also found in the table, which represents the recapture sample. Essentially, in this round, one is counting in the independently generated sample the number of tagged members of the population. The number of records from the sample also found in the stored table whose quality is being determined is the value m_2. If the size of the stored table is n_2, an estimate for the number of missing rows is found from $(n_1 * n_2/m_2) - n_2$.

We illustrate this process through a hypothetical example. Assume that a company database has a stored employee table consisting of 1,000 (n_2) employees. An independent evaluation (perhaps an employee survey) found 100 (n_1) employees. This sample of 100 would be the tagged members of the population that we would try to locate in the database table. If of these 100, 80 (m_2) were in the database, then our estimate for the number of missing employees from the table would be $(100 * 1,000/80) - 1,000 = 250$.

5.3 Impact of Missing Rows on Joins

We continue this discussion by examining the impact of missing rows, an issue for all the algebraic operations but more complicated in the context of Joins. If one wishes to analyze the impact of missing rows on each IP created via Joins, then it is necessary to estimate the number of missing rows for each table using an approach such as the capture/recapture approach discussed in Section 5.2. Should the joining field be a foreign key, then each missing row in the foreign key table would result in exactly one missing row in the IP. (This statement is not exactly correct, as some of the foreign key values may be NULL. For such situations, the same fraction of NULL values found with the extant data should also be used with the missing data, and missing rows with NULL would not be involved in the Join.) Assuming referential integrity, the missing foreign key row would either pair with an existing row or possibly with a missing row. It should be noted that, in either case, there will be one missing row in the resulting table. The item of concern, of course, is the number of missing rows in the IP, each of which clearly would be labeled as unacceptable.

If the Join is over a nonforeign key field, then the situation is more complex. In this case, once the number of missing rows for each of the tables to be joined has been ascertained using a procedure such as the capture/recapture method described below, it would be necessary next to estimate the distribution of values in each of the joining fields. Then, one would employ Monte Carlo-type simulation to populate those fields with values mirroring the original distributions. (See Robert and Casella [27] for an explanation of Monte Carlo simulation.) Once the joining table values have been simulated, one would perform the Join to form the IP, which could contain rows that involve missing rows from the joining tables. The number of additional rows generated in this manner in the IP would then be incorporated into the acceptability measure for the IP in a straightforward manner. For example, if m of N rows are correct prior to the missing row analysis, and M rows are missing, the acceptability measure would be $\mathrm{m}/(\mathrm{N}+\mathrm{M})$.

5.4 Second Round Sampling

For the IPs not eliminated in the first round, an additional sampling must be undertaken to shorten the confidence intervals given in expression (1) so as to evaluate more accurately whether or not they meet the desired quality levels $\mathrm{A_k}$. Note that second-round sampling applies to all base tables needed for the IPs not eliminated from consideration in the first round.

We now discuss issues regarding how to apportion the resources available for the second-round sampling among the tables used to generate the remaining IPs in a way that optimizes usage of these resources. The estimates for Π and the confidence intervals based on them are used to determine if the quality of the IPs is satisfactory or not. Clearly, it is more important to shorten the confidence intervals for some of these estimates as compared to others. There are various reasons why this is so. Some tables may be involved in many IPs, and, accordingly, having a good estimate for their acceptability levels removes more ambiguity. Also, it is probably true that the IPs differ in terms of their importance. If there should be a key IP, it is especially important for the accept/reject decision to have good estimates for the acceptability levels for the tables that are used to form that IP.

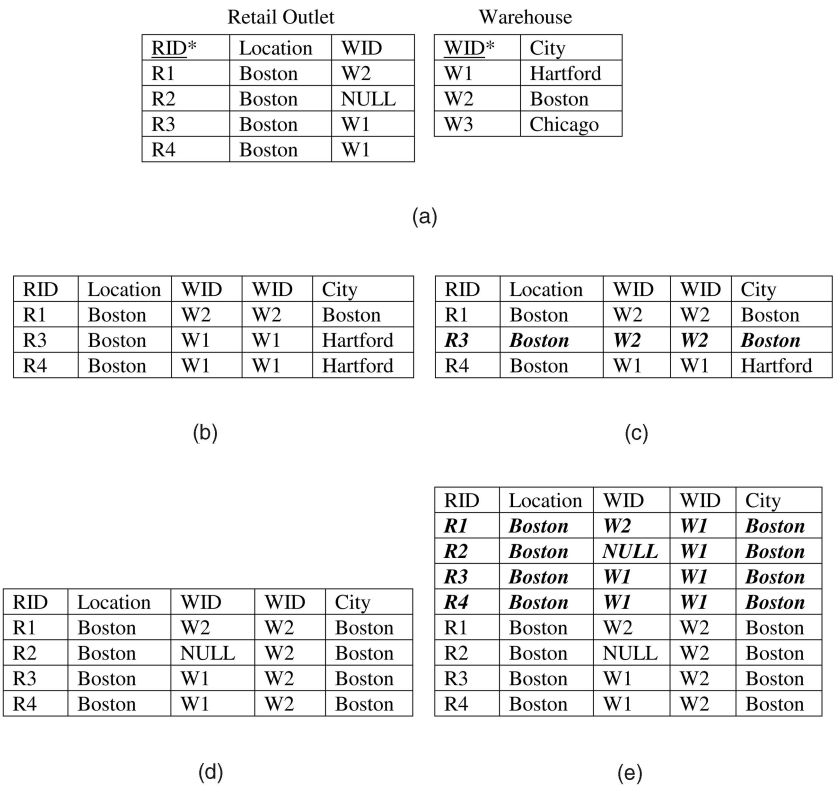
After the additional sampling has been done, the analyst should proceed, as was done with the pilot sample to determine which IPs definitely conform and which definitely do not. For those in the gray area, judgment has to be used. For example, the analyst would consider whether the acceptability level is closer to the product of the L's (accept) or the U's (reject).

6 Concluding Remarks

Managers have always relied on data of less than perfect quality in support of their decision-making activities. Experience and familiarity resulting from use of the data enabled them to develop a feel for its deficiencies and, thus, an ability to make allowances for them. For some time now, computer systems have extracted data from organizational and other databases and manipulated them as appropriate to provide information to managers in support of their activities. As long as the data were extracted from a relatively small number of transaction processing files, it was still possible for management to develop a sense of the quality of the information generated from such sources. However, as the number and diversity of tables available to managers has increased, any hope management might have of intuitively assessing the quality of the information provided to them has pretty much disappeared. The purpose of this paper is to address this need by managers for information regarding the quality of IPs generated for them by computer systems using relational databases.

The problem is exacerbated by the fact that relational tables often contain hundreds of millions of rows in mission critical database applications. Since the diversity of IPs that could be generated from databases is large, to address a manageable subset, we chose to focus on those IPs generated by applying queries formed from the fundamental operations of the relational algebra to relational databases. Since, realistically, it is almost impossible to know the quality of every data unit in a large database with certainty, we use a statistical approach that allows for this. Since it is important to accommodate ad hoc queries, which of course a priori are unknown, one cannot assess data quality in the context of known uses. We address this by taking samples from the base tables independent of any particular use and then identify all possible deficiencies. One limitation of this work is the difficulty in identifying these. Some of these deficiencies will be relevant for certain information products, not for others. Thus, the samples from those base tables involved in producing a certain information product are evaluated for quality in the context of that IP. Statistical procedures, among others, are then used to provide intervals within which, with a known probability, the true but unknown quality measure of the IP would lie. This is the information provided to managers regarding the quality of the IP.

The paper addresses several implementation issues of concern to practitioners. For example, the section on statistical sampling contains a discussion of sample size. Also, there is material as to how to account for the



Retail Outlet

RID*	Location	WID
R1	Boston	W2
R2	Boston	NULL
R3	Boston	W1
R4	Boston	W1

Warehouse

WID*	City
W1	Hartford
W2	Boston
W3	Chicago

(a)

RID	Location	WID	WID	City
R1	Boston	W2	W2	Boston
R3	Boston	W1	W1	Hartford
R4	Boston	W1	W1	Hartford

(b)

RID	Location	WID	WID	City
R1	Boston	W2	W2	Boston
R3	***Boston***	***W2***	***W2***	***Boston***
R4	Boston	W1	W1	Hartford

(c)

RID	Location	WID	WID	City
R1	Boston	W2	W2	Boston
R2	Boston	NULL	W2	Boston
R3	Boston	W1	W2	Boston
R4	Boston	W1	W2	Boston

(d)

RID	Location	WID	WID	City
R1	***Boston***	***W2***	***W1***	***Boston***
R2	***Boston***	***NULL***	***W1***	***Boston***
R3	***Boston***	***W1***	***W1***	***Boston***
R4	***Boston***	***W1***	***W1***	***Boston***
R1	Boston	W2	W2	Boston
R2	Boston	NULL	W2	Boston
R3	Boston	W1	W2	Boston
R4	Boston	W1	W2	Boston

(e)

Fig. 2. Potential data quality problems in the join operation. The "*" indicates that the attribute is the primary key. (a) Two illustrative tables. (b) Correct join. (c) Incorrect row is a placeholder. (d) Correct join. (e) Hartford recorded as Boston.

possibility of missing rows. This paper provides a staged methodology for the estimation of the quality of IPs. For example, suppose that the IP is a result of a union operation. The first stage would be to assess the quality, ignoring the impact of any duplicates that may have been incorrectly retained or deleted. For a more precise estimate, one would proceed to the second stage that would analyze the IP using the Reference-Table Procedure. A still more complete analysis would be in the context of missing rows. Finally, one can employ a second round of sampling.

Note that the Reference-Table Procedure can be avoided only in relatively straightforward cases (see Table 3 for a complete listing). In order to make this process more easily accessible to practitioners, further work is required. For example, the methodology described in this paper could be automated by writing applications using database retrieval languages such as SQL.

APPENDIX

JOIN EXAMPLE

To see the impact of errors when tables are joined together, in Fig. 2a, we have two tables that are linked by a one-to-many relationship through the WID field. Assuming that the data are correct, the Join over the WID attribute would result in a table consisting of three rows, as shown in Fig. 2b. If a value in the joining field is incorrect, then, assuming that referential integrity has been enforced, the Join would result in a row that, while incorrect, at least should be there. To see this, suppose that in the R3 row of Retail Outlet, the correct value W1 is replaced by an incorrect value, say W2. Then, the Join operation would generate an incorrect row, which is a placeholder for a correct one, as shown in Fig. 2c.

If the joining field is not a foreign key, then the resulting table could contain for each incorrect joining value multiple rows that should not exist. For instance, in Fig. 2a, consider joining the two tables over the **Location** field of Retail Outlet and the **City** field of Warehouse. The result would yield four rows, as shown in Fig. 2d. Now, assume that the value Hartford in the W1 row of the Warehouse table is incorrectly replaced by the value Boston. The resulting Join would now have eight rows as shown in Fig. 2e, four of which would be incorrect (in **bold** and *italic*). Another type of error in that field could lead to multiple missing rows. To see this, consider the case where, in the W2 row of the Warehouse table, the value Boston is incorrectly replaced by San Francisco. In this case, the resulting Join would yield an empty table.

The motivating example has used small tables solely for the purpose of illustration. In practice, the IPs would be generated from tables typically containing thousands or even hundreds of millions of rows (Funk et al [10]). Also note that, although the motivating example is in the context of the accuracy dimension, the same methodology applies to any other data quality dimension or combination of dimensions. The rows of the samples must be evaluated as acceptable or unacceptable using the specified dimensions.

REFERENCES

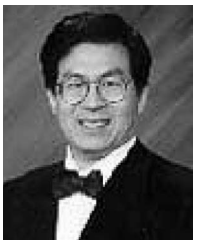
[1] S. Acharya, P.B. Gibbons, V. Poosala, and S. Ramaswamy, "Join Synopses for Approximate Query Answering," *ACM SIGMOD Record, Proc. 1999 ACM SIGMOD Int'l Conf. Management of Data,* vol. 28, no. 2, pp. 275-286, 1999.

[2] S. Acharya, P.B. Gibbons, and V. Poosala, "Congressional Samples for Approximate Answering of Group-by Queries," *ACM SIGMOD Record, Proc. 2000 ACM SIGMOD Int'l Conf. Management of Data,* vol. 29, no. 2, pp. 487-498, 2000.

[3] D.P. Ballou and H.L. Pazer, "Cost/Quality Tradeoffs for Control Procedures in Information Systems," *OMEGA: Int'l J. Management Science,* vol. 15, no. 6, pp. 509-521, 1987.
[4] D.P. Ballou and H.L. Pazer, "Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff," *Information Systems Research,* vol. 6, no. 1, pp. 51-72, 1995.
[5] M.T. Boswell, K.P. Burnham, and G.P. Patil, "Role and Use of Composite Sampling and Capture-Recapture Sampling in Ecological Studies," *Handbook of Statistics,* chapter 19, pp. 469-488, North Holland: Elsevier Science Publishers, 1988.
[6] A.J. Duncan, *Quality Control and Industrial Statistics.* Homewood, Ill.: Irwin, 1986.
[7] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap.* New York: Chapman and Hall, 1993.
[8] S.E. Fienberg and M. Anderson, "An Adjusted Census in 1990: The Supreme Court Decides," *Chance,* vol. 9, no. 3, 1996.
[9] M.L. Fisher, A. Raman, and A.S. McClelland, "Rocket Science Retailing Is Almost Here: Are You Ready?" *Harvard Business Rev.,* pp. 115-124, July-Aug. 2000.
[10] J. Funk, Y. Lee, and R. Wang, "Institutionalizing Information Quality Practice," *Proc. 1998 Conf. Information Quality,* vol. 3, pp. 1-17, 1998.
[11] H. Gitlow, S. Gitlow, A. Oppenheim, and R. Oppenheim, *Tools and Methods for the Improvement of Quality.* Boston: Irwin, 1989.
[12] P.J. Haas and J.M. Hellerstein, "Ripple Joins for Online Aggregation," *Proc. ACM-SIGMOD Int'l Conf. Management of Data,* pp. 287-298, 1999.
[13] J.M. Hellerstein, P.J. Haas, and H.J. Wang, "Online Aggregation," *SIGMOD Record (ACM Special Interest Group on Management of Data),* vol. 26, no. 2, pp. 171-182, 1997.
[14] B.D. Klein and D.L. Goodhue, "Can Humans Detect Errors in Data? Impact of Base Rates, Incentives, and Goals," *MIS Quarterly,* vol. 21, no. 2, pp. 169-195, 1997.
[15] A. Klug, "Equivalence of Relational Algebra and Relational Calculus Query Languages Having Aggregate Functions," *J. ACM,* vol. 29, pp. 699-717, 1982.
[16] K.C. Laudon, "Data Quality and Due Process in Large Interorganizational Record Systems," *Comm. ACM,* vol. 29, no. 1, pp. 4-11, 1986.
[17] D. Little and S. Misra, "Auditing for Database Integrity (IS Management)," *J. Systems Management,* vol. 45, no. 8, pp. 6-11, 1994.
[18] M.V. Mannino, P. Chu, and T. Sager, "Statistical Profile Estimation in Database Systems," *ACM Computing Surveys,* vol. 20, no. 3, pp. 191-221, 1988.
[19] A. Motro and I. Rakov, "Estimating the Quality of Databases," *Flexible Query Answering Systems,* pp. 298-307, Berlin: Springer Verlag, 1988.
[20] F. Naumann, J.C. Freytag, and U. Leser, "Completeness of Integrated Information Sources," *Information Systems,* vol. 29, no. 7, pp. 583-615, 2004.
[21] V.M. O'Reilly, P.J. McDonnell, B.N. Winograd, J.S. Gerson, and H.R. Jaenicke, *Montgomery's Auditing.* New York: Wiley, 1998.
[22] K. Orr, "Data Quality and Systems Theory," *Comm. ACM,* vol. 41, no. 2, pp. 66-71, 1998.
[23] A. Parssian, S. Sarkar, and V.S. Jacob, "Assessing Data Quality for Information Products," *Proc. 20th Int'l Conf. Information Systems,* pp. 428-433, 1999.
[24] A. Parssian, S. Sarkar, and V.S. Jacob, "Assessing Information Quality for the Composite Relational Operation Join," *Proc. Seventh Int'l Conf. Information Quality,* pp. 225-237, 2002.
[25] A. Raman, N. DeHoratius, and Z. Ton, "Execution: The Missing Link in Retail Operations," *California Management Rev.,* vol. 43, no. 3, pp. 136-152, 2001.
[26] P. Rob and C. Coronel, *Database Systems: Design, Implementation, and Management,* fourth ed. Cambridge, Mass.: Course Technology, 2000.
[27] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods.* Springer Verlag, 2004.
[28] M. Scannapieco and C. Batini, "Completeness in the Relational Model: A Comprehensive Framework," *Proc. Int'l Conf. Information Quality,* pp. 333-345, 2004.
[29] R.Y. Wang and D.M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *J. Management Information Systems (JMIS),* vol. 12, no. 4, pp. 5-34, 1996.
[30] R. Weber, *Information Systems Control and Audit.* Upper Saddle River, N.J.: Prentice Hall, 1999.

Donald P. Ballou received the PhD degree in applied mathematics from the University of Michigan. He is now a Professor Emeritus in the Information Technology Management Department in the School of Business at the University at Albany, State University of New York. His research focuses on information quality with special emphasis on its impact on decision making and on ensuring the quality of data. He has published in various journals including *Management Science*, *Information Systems Research*, *MIS Quarterly*, and *Communications of the ACM*.

InduShobha N. Chengalur-Smith received the PhD degree from Virginia Tech in 1989. She is an associate professor in the Information Technology Management Department in the School of Business at the University at Albany, State University of New York. Her research interests are in the areas of information quality, decision making, and technology implementation. She has worked on industry-sponsored projects that ranged from best practices in technology implementation to modeling the costs of bridge rehabilitation. She has published in various journals including *Information Systems Research*, *Communications of the ACM*, and several IEEE transactions.



Richard Y. Wang received the PhD degree in information technology from the Massachusetts Institute of Technology (MIT). He is director of the MIT Information Quality (MITIQ) Program and codirector of the Total Data Quality Management Program at MIT. He also holds an appointment as a university professor of information quality, University of Arkansas at Little Rock. Before heading the MITIQ program, Dr. Wang served as a professor at MIT for a decade. He was also on the faculty of the University of Arizona and Boston University. Dr. Wang has put the term information quality on the intellectual map with a myriad of publications. In 1996, Professor Wang organized the premier International Conference on Information Quality, for which he has served as the general conference chair and currently serves as chairman of the board. Dr. Wang's books on information quality include *Quality Information and Knowledge* (Prentice Hall, 1999), *Data Quality* (Kluwer Academic, 2001), and *Journey to Data Quality* (MIT Press, forthcoming).