Statistical Inference of Unknown Attribute Values In Databases

Wen-Chi Hou[1], Zhongyang Zhang[1], and Nong Zhou[2]

Abstract

In this paper, we propose to use statistical methods to estimate unknown attribute values in databases, as compared to assigning possible values at users' discretion in common practice. Regression models and classification analysis are introduced for estimating continuous and categorical unknown attribute values, respectively. Procedures for selecting relevant attributes in a relation and preliminary experimental results of the proposed models on a real life database are also presented.

Keywords : Relational Database, Unknown Attribute Value, Statistical Inference, Estimation.

1. Introduction

It is generally assumed that a database has complete information about the world it models. However, in reality, there may sometimes be information missing in the database. Various types of missing data, such as "not existing", "not applicable", and "existing but unknown", are often seen in the database. In this paper, we are concerned only with estimation of "existing but unknown" attribute values in a relation, which is hereafter referred to as the unknown (attribute) values.

Various attempts have been made to represent and manipulate unknown values so that databases can capture more meanings. An earlier approach [Codd 79, Bisk 83, etc.] used a special null symbol to represent an unknown attribute value in a tuple. Other representations such as distinguished nulls [Maie 83] and disjunctive values [Reit 86, ImVa89, etc.] have also been devised. Another approach illustrated by [BuPe 84, Zema 85, RaMa 88 etc.] applied fuzzy set concept to describe the vagueness of unknown attribute values.

Recently, it has been suggested that, due to the atomic data type of the attribute, unknown values be represented as mutually exclusive disjunctive values [BGMP 90, Ola 92, Lee 92, etc.], perhaps with some probabilistic measures on the possible values. For example, an employee's age may be described as [30, 0.4] and [31, 0.6], which indicate that the probability of the age being 30 is 0.4, and being 31 is 0.6. Relational algebra operations have been extended to manipulate such data. Although this approach allows quantitative description of the possible values, a fundamental problem, namely how to assign possible attribute values with/without probabilities (e.g., [30, 0.4] and [31, 0.6]), remains unsolved. To the best of our knowledge, there has been virtually no work done in estimating unknown attribute values of base tuples in the database community. The possible values and their associated probabilities are currently assigned at users' discretion.

A relation scheme in the relational database is a collection of related attributes, as suggested by the word "relation". Considering a relation as a statistical population, then it would seen quite natural to use statistical methods to explore relationships among attributes. In this paper, we propose a framework for exploring statistical relationships among attributes and accordingly estimating the unknown attribute values. We believe that the proposed approach not only allows databases to capture more meanings, but also, perhaps more importantly, to capture more reliable information. In order not to misuse statistical inference, however, we should also point out that statistical inference is useful only when statistical relationships do exist among attributes, at least approximately. Therefore, when attributes are not statistically related or data has been encrypted, the proposed approach may not offer much help.

1 Department of Computer Science, Southern Illinois University at Carbondale, IL 62901.

2 Department of Computer Science, New Mexico Tech, Socorro, NM 87801.

CIKM '93 - 11/93/D.C., USA

© 1993 ACM 0-89791-626-3/93/0011 ....$1.50

In this paper, we shall mainly take up two issues in the statistical inference of unknown attribute values. That is, to identify attributes that are relevant to the unknown attribute values (in Section 4), and to approximate existing relationships among relevant attributes using statistical models (in Section 3). We have selected several most commonly used statistical methods that are suitable for our applications to develop this methodology. Preliminary experimental results on a real life database are also presented. The contribution of this paper includes a formal application of statistical methods to the estimation of unknown attribute values and the incorporation of such methodology into a DBMS for automatic inference. Our approach can also be applied to areas such as knowledge mining, uncertainty reasoning, etc.

The rest of the paper is organized as follows. In Section 2, we briefly describes some statistical terminology and notations. Section 3 introduces several statistical models for estimating unknown attribute values as they pertain to different occasions. Section 4 discusses procedures for identifying relevant attributes. Section 5 presents preliminary results of the proposed estimation models on a real life database. Section 6 is the conclusion.

2. Terminology and Notations

First, we briefly describe some statistical terminology. Let ψ be a parameter of interest, such as a population mean or a population total. A *point estimator* of ψ, denoted by $\hat{\psi}$, is a function that returns a value serving as a guess for the value of ψ based on the observation on the sample data. An estimator $\hat{\psi}$ is said to be *unbiased* if the expected value of $\hat{\psi}$ equals the true value of ψ for all possible values of ψ. However, in terms of inference, an estimated interval, called *confidence interval*, together with a statistical *confidence level* shall disclose more information. Confidence interval is an interval of plausible values for the parameter being estimated. Confidence level is the degree of plausibility of such an interval.

Let $\mathbf{X}$ be a matrix. Then, $\mathbf{X}^T$ and $\mathbf{X}^{-1}$ denote the transpose and inverse of the matrix $\mathbf{X}$, respectively. $|\mathbf{X}|$ is the determinant of $\mathbf{X}$.

3. Statistical Models for Estimating Unknown Attribute Values

Statistical estimation methods (e.g., regression estimation, categorical analysis) have been used widely and successfully in various areas such as biology, education, business, agriculture, ..., etc., as they can often produce good approximations to existent relationships or reveal influence of attributes on the phenomena concerned. To establish a statistical estimation model, the first step is usually to identify the set of factors (i.e., a set of relevant attributes) that may contribute to the effect (i.e., the attribute with a unknown value in our context). Normally, identifying relevant attributes is not a tough job, since users may usually have good knowledge of the database. When the knowledge is lacking, standard statistical procedures for identifying relevant attributes are also available. We will take up this issue in more detailed in Section 4 after some statistical notions are described in this section. Here, we shall focus on how to derive the relationships and estimate the unknown values accordingly under the assumption that statistical relationships do exist, at least approximately.

In relational databases, a relation can be viewed as a table with rows (called tuples) and columns (called attributes). Each column describes a certain property of the tuples, and has its own domain of values. A row in the table represents a relationship among a set of attribute values. Consequently, an attribute can be treated as a variable, a tuple as an observation, and a relation as a statistical population. With the analogy, one can readily apply statistical methods to explore the relationships among attributes (if exist), and draw statistical inference accordingly on the unknown attribute values. In this paper, we use tuple and observation, attribute and variable, population and relation interchangeably.

In the following, we shall exemplify the use of various estimation methods using the relation scheme EMP(Name, Salary, Dept, Exp(erience), Rank, Major). An attribute of a relation can be either *continuous* or *categorical*. A continuous attribute is one on which subjects differ in amount or degree, e.g., the experience or salary of an employee. A categorical attribute is one on which the subjects differ in type or kind, e.g., the rank, department or major of an employee. In this section, we shall discuss several statistical models as they per-

tain to estimation of different types of unknown attribute values. For simplicity, hereafter, *an unknown attribute* refers to an attribute whose value in the tuple concerned is unknown. Also, for illustration of various methods, different assumptions on the underlying relationship among attributes may be made.

3.1 Regression Models For Continuous Unknown Attributes

First, we consider the situations where the unknown attribute values are continuous, e.g., estimation of the salary of an employee. The most commonly used method to reveal statistical relationships among continuous variables is the *Regression Estimation* [Devo 84, Jowi 92]. In a regression model, variables are usually classified into two classes, i.e., the *explanatory and dependent* variables. Roughly speaking, a dependent variable corresponds to an unknown attribute in our context, whereas explanatory variables correspond to those attributes which are relevant to the dependent variable and thus whose values can be used to "explain" (or estimate) the value of the dependent variable.

To apply regression estimation, the first step is to select an appropriate model to approximate the existent relationship, as it is crucial to draw correct inference. In general, a regression model can be specified either as *a linear regression model* or *a nonlinear regression model*. In the following, we discuss how to apply linear and nonlinear techniques to the estimation of unknown attribute values.

3.1.1 Linear Regression Model with Continuous Explanatory Attributes

Consider the relation scheme EMP mentioned earlier. Assume that the Salary of an employee is related solely to the Experience, and yet an approximately linear relationship exists between them. If one is to estimate the salary of an employee, then using a linear regression model with the Salary and Exp as the dependent and explanatory variables, respectively, may closely approximate such a relationship. In the following, we formally describe the derivation of relationship among attributes from (a sample of) a relation, and discuss its properties on estimation.

Let Y be an unknown attribute (i.e., a dependent variable), and X be a set of attributes $X_1, X_2, \ldots, X_m$, identified to be relevant to Y (i.e., explanatory variables). Both X and Y are assumed to be continuous. Let's further assume that an approximately "linear" relationship is believed to exists between X and Y. Let $(x_{1i}, x_{2i}, \ldots, x_{mi}, y_i)$ be the value of a tuple t_i projected onto the set of attributes X and Y. Considering $(x_{1i}, x_{2i}, \ldots, x_{mi}, y_i)$ as an observation, then the relation between X and Y in the linear regression model is formulated as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} + e_i \qquad i = 1, 2, \ldots, n \tag{1}$$

where β_0 is the intercept; β_j, $j = 1, 2, \ldots, m$, are the regression coefficients of Y on X_j; e_i is the error term designed to account for all sources of variability of Y, other than X; and n is the number of observations (i.e., the size of a sample). It is user's preference to express the intercept β_0 either explicitly (as shown in equation (1)) or implicitly (i.e., without β_0) in the regression model. However, whatever one chooses should not affect the expressive power of a regression model and the following discussion.

One can rewrite equation (1) in matrix form as

$$\mathbf{Y} = \mathbf{X}\,\beta + \mathbf{e} \tag{2}$$

where $\mathbf{Y}$ is the $n \times 1$ column vector $(y_1, y_2, \ldots, y_n)^T$; $\mathbf{X}$ is an $n \times (m+1)$ matrix with the i^{th} row being the vector $(1, x_{1i}, x_{2i}, \ldots, x_{mi})$, $i = 1, 2, \ldots, n$; β is the column vector $(\beta_0, \beta_1, \beta_2, \ldots, \beta_m)^T$ representing the relationship to be estimated; and $\mathbf{e}$ is the error term vector $(e_1, e_2, \ldots, e_n)^T$. In the following, we describe how to derive an estimator for β that minimizes the sum of the square errors (i.e., $\sum_i e_i^2$) and then draw an estimation of the unknown attribute value accordingly.

3.1.1.1 Estimating Linear Relation and the Unknown Value

From equation (2), it can be easily shown that $(\mathbf{Y} - \mathbf{X}\,\beta)^T(\mathbf{Y} - \mathbf{X}\,\beta)$ is the *Sum of Square Errors* (i.e., $\sum_i e_i^2$). Then, by minimizing the sum of square errors, an estimator, denoted by $\hat{\beta}$, is obtained as

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\mathbf{Y} \tag{3}$$

The estimator $\hat{\beta}$ so obtained, called *the best linear*

unbiased estimator, has minimum variance and is unbiased. Given a tuple with an unknown value y for attribute Y, then, the best linear unbiased point estimator of y, denoted by $\hat{y}$, is

$$\hat{y} = \mathbf{x}\,\hat{\beta} \tag{4}$$

where $\mathbf{x}$ is the explanatory attribute values of the tuple, represented as a row vector. Under the assumptions that the explanatory variables are normally distributed, it follows that $\hat{\beta}$ has a normal distribution [Chou 75, Devo 84]. Let α represent the desired confidence level, then the 100 (1- α) percent confidence interval for the estimator $\hat{y}$ is

$$\hat{y} \pm t_{\alpha/2,n-m}\hat{\sigma}_{\hat{y}} \tag{5}$$

where $\hat{\sigma}_{\hat{y}}$ is an estimate of the standard deviation of the estimator $\hat{y}$, which can be obtained from the sample; $t_{\alpha/2,n-m}$ is the t distribution value at $\alpha/2$ with an $n-m$ degree of freedom. When the normality assumption on the explanatory variables does not hold, one can apply transformation procedures [Judge 85, John 89] to normalize them, and then draw inference as above.

3.1.2 Linear Regression Model with Categorical Explanatory Attributes

In the previous discussion, we have assumed that all the explanatory attributes were continuous. Nevertheless, in many database applications, (some of) the explanatory attributes may be categorical. Therefore, the regression model discussed in the previous section has to be extended to accommodate categorical explanatory variables.

A regression model with categorical explanatory variables can best be understooded by an example. Consider the relation scheme EMP discussed earlier. Assume that salary is not only related to the experience, but also related to the department one serves. For simplicity, we assume that Dept has a domain {CS, EE, ME}. A common way to accommodate a categorical variable in a regression model is to represent each distinct categorical value by a dummy variable. The method is thus called the *dummy coding method*. For instance, dummy variables D_1, D_2, and D_3 are introduced for CS, EE, and ME, respectively. D_i has the value 1 if the tuple of concern has the corresponding domain value as its Dept value; otherwise D_i has a value 0. For example, if the employee works in CS department, then $D_1 = 1, D_2 = D_3 = 0$. With dummy variables, the regression model can be expressed as

$$y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 Exp_i + e_i \qquad i = 1, 2, \cdots, n \tag{6}$$

Again, one can rewrite the above equation in matrix form as

$$\mathbf{Y} = \mathbf{X}\,\beta + \mathbf{e} \tag{7}$$

where $\mathbf{Y}$, $\mathbf{X}$ and β are defined as in the last section. Now, we can apply the same estimation technique discussed in the last section to pursue the analyses. In general, we need k dummy variables to substitute for an attribute with a domain of k values. The dummy coding method can be readily generalized to accommodate multiple categorical variables. Let $|A_i|$ be the number of distinct values in the domain of an explanatory categorical attribute A_i. Then, the number of dummy variables required to encode all of the explanatory categorical variables is $\sum_i |A_i|$.

3.1.3 Nonlinear Regression Model With Continuous Explanatory Attributes

Nonlinearities may appear in various forms in the relations. Some of the nonlinear relationships can well be transformed into linear relations and fit into linear models, while others have to be dealt with using different techniques. In the following, we show some examples of the transformations from nonlinearity to linearity.

Example. Consider the following "nonlinear" models

$$Y = \beta_0 + \beta_1 \frac{1}{X} \tag{8}$$

$$Y^2 = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2^3 \tag{9}$$

$$Y = \beta_0 X^{\beta_1} \tag{10}$$

If we substitute X' for $1/X$ in equation (8), a linear model $Y = \beta_0 + \beta_1 X'$ is obtained. By the same token, in equation (9) we can substitute X_1' for X_1^2, X_2' for X_2^3, Y' for Y^2, and obtain a linear model $Y' = \beta_0 + \beta_1 X_1' + \beta_2 X_2'$. As for equation (10), if we take a logarithm on both sides of the equation, it should not be difficult to see that a linear model $Y' = \beta_0' + \beta_1 X'$ is derived, where $Y' = \log Y$, $X' = \log X$, and $\beta_0' = \log \beta_0$. Of course, the linear

regression techniques discussed in the previous sections can be directly applied to resulting models. Hereafter, nonlinear models refer to those which can not be converted into linear ones.

3.1.4 Estimating Nonlinear Regression Model

A nonlinear model is generally specified as

$$\mathbf{Y} = f(\mathbf{X}, \beta) + \mathbf{e} \quad (11)$$

where **Y**, **X**, β, and **e** are defined as before, and their respective statistical assumptions remain valid. As in the cases of linear models, an estimate of the β value can be obtained by minimizing the sum of square errors.

The nonlinear least square estimate for parameter vector β is obtained by minimizing the sum of square errors $\sum e_i^2$, i.e., $(\mathbf{Y}-f(\mathbf{X}, \beta))^T (\mathbf{Y}- f(\mathbf{X}, \beta))$. Unfortunately, the solution to β can not be expressed in analytical form as in the linear case, and has to resort to a numerical method such as Gauss-Newton (or Newton-Raphson) algorithm to obtain the estimates.

In Gauss-Newton method, the nonlinear function $f(\mathbf{X}, \beta)$ is approximated by the first-order Taylor series expansion around an initial point β^0, that is

$$f(\mathbf{X}, \beta) = f(\mathbf{X}, \beta^0) + D(\beta^0)\,(\beta - \beta^0) + \mathbf{e} \quad (12)$$

where $D(\beta^0)$ is $\dfrac{\partial f(\mathbf{X}, \beta^0)}{\partial \beta^T}$, the matrix of the first-order derivative around β^0. The numerical solution to β at the $n+1^{th}$ iteration, denoted by β^{n+1}, is given by

$$\beta^{n+1} = \beta^n + [D(\beta^n)]^{-1} D(\beta^n)^T [\mathbf{Y} - f(\mathbf{X}, \beta^n)] \quad (13)$$

The iterative process stops and converges to a local minimum when the difference between β^{n+1} and β^n is less than a predefined threshold. However this solution does not guarantee a global minimum in terms of sum of square errors. In general, one should try a number of different initial values β^0 and then select the one with the least sum of square errors [Judg 88]. Once we obtain the estimated β vector, we can then obtain a point or interval estimate of the corresponding unknown Y value and draw statistical inference as we did in the previous sections.

3.2 Classification Analysis for Categorical Unknown Attribute

In the linear and nonlinear regression models, it is assumed that the dependent variable is continuous. However, there are many cases in the databases where the unknown attributes (e.g., the Rank, Dept or Major of an employee) are categorical. There have been various methods proposed in the literature to deal with estimation of categorical values. For example, one can first transform a categorical attribute into a normally distributed continuous attribute, and then apply the regression analysis discussed in the last section to obtain an estimation [John 89]. However, we think it is more natural in general to consider the estimation of categorical values as a classification problem. Therefore, in this section we propose the most commonly used classification and resemblance techniques to estimate the unknown categorical attribute value.

Classification is a grouping method based on the measure of resemblance, which is defined as a measure of profile similarity or dissimilarity [Tats 88] by comparing the characteristics of objects. The entire classification analysis can be carried out in two steps. First, all of the complete tuples are classified into distinct groups according to the values of the dependent attribute. For example, assume that we are to estimate the Dept value of a tuple in the EMP relation. Then, the tuples would naturally fall into three categories, i.e., CS, EE, and ME, according to their Dept values. In the second step, we compute the dissimilarity coefficients of the tuple with respect to each individual group. The dissimilarity coefficient indicates how far away a tuple deviates from a group, and thus can be used as a criterion for the classification.

3.2.1 Classification with Continuous Explanatory Attribute

First, we discuss the situations where the explanatory attributes are continuous. For example, if one wish to estimate the Rank of an employee, and it has been known that Rank is closely related to one's Experience. Then we can use Exp as explanatory variable to estimate the Rank value by the following classification technique.

We preassume that the dependent attribute Y has a domain of K distinct values. Thus, the remaining tasks of the classification process are to measure the dissimilarity of a tuple relative to each individual group,

and then to predict the membership of the tuple. Let $X_1, X_2, \ldots, X_m$ be the set of continuous explanatory variables, and $(x_{1i}, x_{2i}, \ldots, x_{mi})$ (x_i for short) be the $X_1, X_2, \ldots, X_m$ values of a given tuple t_i. Let $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)$ be the *centroid* of the k^{th} group (for simplicity, without specifying the k), where $\bar{x}_j$, $j = 1, 2, \ldots, m$, are the means of the X_j values of the group members. The centroid vector indicates the average (or standard) characteristics of the concerned group. The dissimilarity of a tuple t_i with respect to the k^{th} group is usually measured by the "squared distance" defined as

$$D_{ik}^2 = (x_{1i} - \bar{x}_1, \cdots, x_{mi} - \bar{x}_m) \times C_k^{-1} \times (x_{1i} - \bar{x}_1, \ldots, x_{mi} - \bar{x}_m)^T \quad (14)$$

where C_k^{-1} is the inverse of the covariance matrix of the k^{th} group. Under the normality assumption for the variables, D_{ik}^2 has a Chi-square distribution with $m - 1$ degree of freedom, and is thus referred to as a Chi-square (χ^2)-statistic. Both the centroid $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)$ and the covariance matrix C_k of a group can be obtained from (a sample of) the relation. It should be noted that the larger the χ^2 statistic, the farther away (in the generalized distance sense) is the point $(x_{1i}, x_{2i}, \ldots, x_{mi})$ from the centroid $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)$ of the reference group. Thus, the tuple may be said to be more deviant from the "average member" of the group. Conversely, a small χ^2 statistic indicates that the tuple resembles that group closely.

Although χ^2 statistic is a good indication of how deviant a tuple is from a given group, it is not possible in general to calculate the probability of a tuple belonging to a group, unless some assumption on the distribution of the variables is made. Let G_k represent the k^{th} group of tuples. Then $p(t_i|G_k)$ is the conditional probability that a tuple of the k^{th} group is t_i (i.e., having x_i as its value). Assume that variables in each group can be characterized by a multivariate normal density function, then,

$$p(t_i|G_k) = (2\pi)^{-m/2} |C_k|^{-1/2} e^{-(\chi_{ik}^2)/2} \quad (15)$$

where $|C_k|$ denotes the determinant of C_k [Jame 85, Tats 88]. The right hand side of equation (15) denotes the multivariate normal density function for a member of the k^{th} group having the value $(x_{1i}, x_{2i}, \ldots, x_{mi})$ (i.e., x_i). Then, by Bayes' theorem on posterior probability [Tats 88, JoWi 92], the probability of t_i (having value x_i) being a member of the group k, denoted $p(G_k|t_i)$, is computed as

$$p(G_k|t_i) = \frac{p(G_k)\ p(t_i|G_k)}{\sum_{j=1}^{k} p(G_j)\ p(t_i|G_j)} \quad (16)$$

where $p(G_k)$ is the prior probability of a tuple being a member of the k^{th} group, which can be obtained from (a sample of) the relation.

"Assign the object to the group with the highest conditional probability" [Jame 85, Tats 88] has been known as Bayes' rule. This rule minimizes the total error of classification. Thus, one should assign the tuple to the k^{th} group which has the largest $p(G_k|t_i)$ value.

3.2.2 Classification with Categorical Explanatory Attributes

In the above discussion, we have assumed that all of the explanatory attributes are continuous. When the assumption does not hold, a different metric for the resemblance coefficient may be necessary. For example, if one is to estimate the department an employee works in, and it has been found that "Major" is strongly related to one's "Dept". In this situation, the explanatory variable is categorical, and thus a different similarity metric on categorical values is needed. Here, we introduce *Jaccard's Similarity Coefficient* [Rome 90] to measure the resemblance on the categorical explanatory variables. Readers may choose to use other measures, such as Yule's, Hamann's coefficients [Rome 90], wherever appropriate.

The Jaccard's resemblance (or similarity) coefficient between two tuples t_u and t_v, denoted by J_{uv}, is defined to be the frequency of identical explanatory attribute values between the two tuples. For example, assume that Dept and Rank are the explanatory attributes, then two tuples projected onto the attributes Dept and Rank (i.e., (d_1, r_1) and (d_2, r_1)) yield $J_{uv} = 0.5$, as they have identical values on one of the explanatory attributes Rank (i.e., one out of two explanatory attributes).

Let $X_1, X_2, \cdots, X_m$ be a set of categorical explanatory attributes. As in Section 3.2.1, we assume that the entire population is divided into K groups according to its values of the dependent variable. Let t_u

be the tuple of concern (i.e., to be classified), and t_v be any tuple in the k^{th} group. Then, the mean of the similarity coefficient between t_u and the k^{th} group, is computed as

$$J_k = \frac{\sum_v J_{uv}}{N_k} \tag{17}$$

where N_k be the number of tuples in the k^{th} group. It should be clear that the higher the value of J_k is, the closer the tuple t_u resembles the k^{th} group, since they share more identical attribute values. Consequently, it would be quite natural to draw an inference that tuple t_u is more likely to be a member of the group with which it shares a higher similarity coefficient value. However, the probability that such an inference statement being correct is determined not only by the similarity coefficient computed with respect to each individual group, but also by the dissimilarity of the K groups. The more dissimilar the groups, the higher is the probability of having a right classification on tuple t_u. In this regard, we adopt the χ^2 test, as in Section 3.2.1, as the significance test for the classification.

The expected similarity coefficient for the entire population can be calculated as

$$J = \frac{\sum_k J_k}{K} \qquad k = 1, 2, \ldots, K \tag{18}$$

Thus, the Chi-square statistic is constructed as

$$\chi^2 = \sum_k \frac{(J_k - J)^2}{J} \qquad k = 1, 2, \ldots, K \tag{19}$$

with (K-1) degree of freedom.

4. Selection of Relevant Attributes

Given a dependent variable, an estimation model can be established by first identifying the set of explanatory variables, which should include all the relevant attributes, yet exclude any irrelevant attributes for correctness and efficiency of the estimation model. The explanatory attributes should well explain the behavior of the dependent variable. Including insignificant explanatory variables may degrade the efficiency of the model, while excluding significant ones may cause distortion of the inference. Once the explanatory attributes are identified, one can proceed to estimate by choosing an appropriate model as discussed in Section 3.

The selection of explanatory variables can be based on one's understanding of the data. For example, if one is to estimate an employee's salary in relation EMP, then it would seen quite natural, from our experience, to assume that the salary is related to one's Rank, Exp, Dept, and perhaps some other attributes. In case that knowledge is not available, some standard statistical procedures can be applied. In the following, we discuss two selection procedures as they pertain to different types of explanatory variables.

4.1 Selection of Continuous Explanatory Variables

Stepwise Forward Selection, is one of the most commonly used methods in selecting continuous explanatory variables. This method is based on the measure of R^2, which is defined as

$$R^2 = 1 - \frac{SSE\ /\ (n-m)}{TSS\ /\ (n-1)} \tag{20}$$

where SSE is *Sum of Square Errors* (i.e.,$\sum_i \hat{e}_i^2$), representing the portion of variation in dependent variable *not explained* by the explanatory variables; TSS is the *Total Sum of Square* ($\sum_i (y_i - \bar{y})^2$), representing the total variation of the dependent variable; n is the number of sample tuples, and m is the number of explanatory attributes in the model. Note that the higher the R^2 value is, the better the model fits a linear relation. Therefore, R^2 is a good indicator of whether an attribute is relevant or not.

Stepwise forward selection starts with an empty explanatory variable set. By pairing a candidate explanatory variable with the dependent variable, an R^2 is computed. The method selects the attribute which yields the highest R^2 value when paired with the dependent variable into the explanatory variable set. In the second step, each remaining candidate variable is again paired with the dependent and the already selected explanatory variable(s), and the respective R^2 value is calculated. As before, the one adds the largest value to R^2 is selected into the explanatory variable set. This process continues until no more explanatory variable can yield a significant change to R^2.

4.2 Selection of Categorical Explanatory Variables

When the dependent variable is categorical, *Chi-square Statistic Analysis* [Free 87, Suit 85] is among the most commonly used methods for selecting both continuous and categorical variables. When an explanatory variable is continuous, its χ^2 value can be directly computed by Equation (14). When an explanatory variables is categorical, the Chi-square value, however, has to be obtained through other measures. *Two-Way Frequency (Contingency) Table* can be used for this purpose (i.e., to derive the Chi-square statistic) [Devo 84, HoWo 73]. Values of the dependent variable form the rows of the table, and values of the explanatory variable form the columns. Let n_{ij} be the value of a cell in the i^{th} row and j^{th} column, then n_{ij} signifies the frequency that the i^{th} dependent variable value and j^{th} explanatory variable value appear together in the observations. Let $n_j = \sum_i n_{ij}$ (i.e., the sum of the j^{th} column), $n_i = \sum_j n_{ij}$ (i.e., the sum of the i^{th} row), and $m_{ij} = n_i\, n_j/n$. Then the Chi-square statistic is computed as

$$\chi^2 = \frac{\sum_i \sum_j (n_{ij} - m_{ij})^2}{m_{ij}} \qquad (21)$$

with $(K_1\text{-}1)(K_2\text{-}1)$ degree of freedom, where K_1 and K_2 are numbers of distinct values in the domains of dependent and explanatory variables, respectively. A high χ^2 value usually indicates a close correlation. Thus, variables with low χ^2 values are considered to be less correlated with the dependent variable.

4.3 Selection of Appropriate Estimation Models

Once the explanatory variables are determined, one can proceed to select statistical models for approximating the intended relationship. For example, suppose that one is to estimate a faculty member's Salary, and Rank, Exp, and Dept have been identified to be relevant to Salary. Then, a possible statistical model may be constructed as

$$Salary_i = \beta_1\, Rank_i + \beta_2\, Exp_i + \beta_3\, Dept_i + e_i \qquad i = 1, \ldots, n \qquad (22)$$

where β_j, $1 \le j \le 3$, are the coefficients for explanatory variables. For simplicity, in the above model, we did not replace the categorical attributes (i.e., Rank and Dept) by sets of dummy variables. Usually, users can try several potential models (e.g., linear and nonlinear ones) and pick the one that best approximate the existent relationship.

5. Preliminary Experimental Results

We have incorporated the statistical methods discussed in Section 3 and 4 into a DBMS, called CASE-DB [HoOz 93], by taking advantages of a commonly used statistical software called SAS [SAS 91]. Given a dependent variable, relevant attributes (if exist) are first identified, and then a statistical relationship is derived and stored in a system catalog called STAT. When an unknown attribute value is encountered, the system catalog is consulted automatically to see if there exists a relationship that can be used to estimate the unknown value.

The experiments are performed by first marking out the value of a selected attribute (i.e., as the unknown attribute) of a tuple, and then, using the statistical relationship stored in STAT, an estimate of the unknown attribute value is obtained and compared with the real value. The relation in experiments has the scheme Emp (SS#, Name, Address, Phone#, Gender, Age, Salary, Degree, Dept, College, Exp, Rank). Data is collected on the faculty members of a university in the United State. Due to space limitation, we present only two type of experiments, one with continuous unknown attribute, and the other with categorical unknown attribute. The explanatory attributes consist of both continuous and categorical attributes.

Each experiment runs two hundred trials. Two indices are set up to evaluate the performance. One, called the prediction biasness (*pb*), is designed for the continuous dependent attribute, and is defined as

$$pb = \left| \frac{y_0 - \hat{y}}{y_0} \right| \qquad (23)$$

where y_0 is the true value, and $\hat{y}$ is the estimate. This index gives us percent deviation of the estimator from the true value. Another index, called the ratio of success (rs), is designed for the cases with a categorical dependent attribute, and is defined as

$$rs = \frac{number\ of\ correct\ predictions}{number\ of\ predictions} \qquad (24)$$

5.1 An Experiment On A Continuous Unknown Attribute

Given the relation scheme Emp (SS#, Name, Address, Phone, Gender, Age, Salary, Degree, Dept, College, Exp, Rank), assume that one is to estimate the salary of an employee. There are totally 15 departments and 4 different ranks (i.e., lecturer (Lect), assistant professor (AstP), associate professor (AsoP) and professor (Prof)). First, we need to identify the set of attributes relevant to Salary. According to some preliminary tests, as described in Section 4, it has been shown that Dept, Rank, and Exp have significant effect on Salary. Therefore, the following model is established.

$$Salary_i = \beta_1 Exp_i + \beta_2 Rank_i + \beta_3 Dept_i \quad i = 1, \ldots, n \qquad (25)$$

Note that Exp is a continuous attribute, while Rank and Dept are categorical attributes. It has been calculated based on the data set that β_1 is 0.16; β_2 is 20.0, 35.6, 39.9, and 48.2 for Lect, AstP, AsoP, and Prof, respectively; and β_3 is 6.26, 6.49, ..., for CS, EE, ... departments, respectively. Thus, the salary of an employee is estimated by substituting one's Exp, Rank, Dept information and their respective coefficients into the above equation. For instance, given a confidence level, say 95%, John, who is an associate professor in CS with 10 years' experience, is estimated to have a salary of 47.8±6.4K. The mean prediction bias (pb) obtained from 200 experiments is 8.7%. For 92% of our 200 experiments, the actual salaries of the employees do fall in the estimated 95% confidence intervals. Overall, the model illustrated by Equation (25) has effectively described the relationship between dependent and explanatory attributes, though it may not be perfect (if ever exists one).

5.2 An Experiment On A Categorical Unknown Attribute

Assume that one is to estimate the rank of an employee. The attributes identified to be relevant to the Rank are Salary, Exp, and Dept, where Dept is a categorical attribute, whereas Salary and Exp are continuous attributes.

The Rank has a domain of 4 distinct values; in other words, the entire relation is subdivided into 4 groups. Here the value of the Dept is used to refine each group. For instance, knowing that the tuple in question works in the CS department, each group should therefore be refined and contain only those people who work in the CS department. Note that, initially each group may contain employees of the same rank from various Depts.

The experimental results have shown that 83% of the time the correct classifications are made (i.e., rs=83%). For another 15% of the experiments, we predict the right ranks as the second choice (i.e., the ones with the second highest probability). In other words, 98% of the time the right ranks appear as the top two choices in the estimation. However, we would like to also point out that the performance of an estimation model is largely dependent on how closely the dependent variables are related to the independent variables. If explanatory attributes contain decisive factors, higher success rates are expected; otherwise, a poor performance is also possible.

6. Conclusion and Future Work

In this paper, we have introduced several statistical estimation methods that have been used widely and successfully in various areas, such as biology, education, business, etc.. Good performance has also been demonstrated in our preliminary experimental results on the estimation of unknown attribute values. The proposed approach is systematic, as compared to assigning possible values at users' discretion in current practice, and is thus free of personal bias. We believe that statistical methods are useful tools for exploring relationships among attributes, and can facilitate the database to capture more reliable meanings. Right now, we store statistical relationships obtained in a system catalog so that whenever an unknown value is to be estimated, the relationships can be consulted automatically. The contribution of this paper includes a formal application of statistical methods to the estimation of unknown attribute values and the incorporation of such methodology into a DBMS for automatic inference.

A potential application of the method developed here is in the probabilistic reasoning. In order to make probabilistic reasoning in belief networks [Pear 88], conditional probability distributions among related variables must be provided, which are usually expressed in tables with discrete values. This is often cumbersome even for variables with categorical domains, and is not feasible for variables with continuous domains. The

developed method can be used to estimate the relationships among variables and express it as a function, instead of discrete data values. This will make the probability calculation easier. Also, we are exploring potential applications of our approach in areas such as uncertainty reasoning and knowledge mining.

Acknowledgement

We would like to thank Yong Qiang Gao who first introduced the idea of using statistical methods to estimate unknown attribute values.

References

[Bisk 83] Biskup, J., "A Foundation of Codd's Relational Maybe-Operation", ACM TODS, Vol. 8, No. 4, Dec 1983, pp. 608-636.

[BGMP 90] Barbara, D., Garcia-Molina, H., Porter, D., "The Management of Probabilistic Data", Proc. EDBT, Venice 1990, pp. 60-74.

[BuPe 84] Buckles, B. and Petry, F., "Extending the Fuzzy Database with Fuzzy Numbers", Info. Sci. Vol. 34, No. 2, 1984, pp. 145-155.

[Chou 75] Chou, Y-I. "Statistical Analysis", Holt, Rinehart and Winston, 1975.

[Codd 79] Codd, E., "Extending the Database Relational Model to Capture More Meaning", ACM Trans. on Database Systems, Vol. 4, No. 4, Dec 1979, pp. 397-434.

[Devo 84] Devore, J., "Probability & Statistics for Engineering and the Sciences", Brooks/Cole Publishing, 1984.

[Free 87] Freeman, D., "Applied Categorical Data Analysis", Dekker, 1987.

[HoWo 73] Hollander, M., and Wolfe, D., "Nonparametric Statistical Methods", John Wiley, 1973.

[HoOz 93] Hou, W-C., Ozsoyoglu, G., *"Processing Real-Time Aggregate Relational Queries in CASE-DB"*, ACM Transactions on Database Systems Vol. 18, No. 2, June, 1993.

[ImVa 89] Imielinski, T., Vadaparty, K., "Complexity of Query Processing in Databases with OR-Objects", PODS, 1989, pp. 51-65.

[Jame 85] James, M., "Classification Algorithems", John Wiley & Sons, New York, 1985.

[John 89] Johnston, J., "Econometric Methods", McGraw Hill, 1989.

[JoWi 92] Johnson, R. and Wichern, D., "Applied Multivariate Statistical Analysis", 3rd ed. Prentice-Hall Inc., Englewood Cliffs, 1992.

[Judg 85] Judge, G., Griffiths, W.E., etc., "The Theory and Practice of Econometrics", Wiley, 1985.

[Lee 92] Lee, S., "Imprecise and Uncertain Information in Databases: An Evidential Approach", Proc. of Data Engineering, 1992, pp. 614-621.

[Maie 83] Maier, D., "The Theory of Relational Databases", Computer Science Press, 1983.

[Ola 92] Ola, A., "Relational Databases with Exclusive Disjunctions", Proc., 8th Intl. Conf. on Data Engineering, Feb 1992, Tempe, AZ.

[Pear 88] Pearl, J., "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference" Morgan Kaufmann publisher, 1988.

[RaMa 88] Raju, K., Majumdar A., "Fuzzy Functional Dependencies and Lossless Join Decomposition of Fuzzy Relational Database Systems", ACM TODS, Vol. 13, No. 2, June 1988, pp 129-166.

[Reit 86] Reiter, R., "A Sound and Sometimes Complete Query Evaluation Algorithm for Relational Databases with Null Values", JACM Vol. 33, No. 2, April 1986, pp. 349-370.

[Rome 90] Romesburg, H., "Cluster Analysis for Researchers", Rober E. Krieger Publishing, 1990.

[SAS 91] "SAS/STAT User's Guide", Release 6.03 Ed., SAS Institute Inc., North Carolina.

[Suit 85] Suits, D. "Statistics : An Introduction to Quantitative Economic Research", Halyburton Press, 1985.

[Tats 88] Tatsuoka, M., "Multivariate Analysis", Macmillan Publishing, 1988.

[Zema 85] Zemankova, M., "Implementing Imprecision in Information Systems", Info. Sci. Vol. 37, 1985, pp. 107-141.