A Statistical Theory for Quantitative Association Rules

Yonatan Aumann*
Bar-Ilan University
aumann@cs.biu.ac.il

Yehuda Lindell†
The Weizmann Institute of Science
lindell@wisdom.weizmann.ac.il

Abstract

Association rules are a key data-mining tool and as such have been well researched. So far, this research has focused predominantly on databases containing categorical data only. However, many real-world databases contain quantitative attributes and current solutions for this case are so far inadequate. We introduce a new definition of quantitative association rules based on statistical inference theory. Our definition reflects the intuition that the goal of association rules is to find extraordinary and therefore interesting phenomena in databases. We present rigorous experimental evaluation on real-world datasets, demonstrating the usefulness and characteristics of rules mined according to our definition.

1 Introduction

Association Rules. The goal of data mining is to extract higher level information from an abundance of raw data. Association rules are a key tool used for this purpose. An association rule is a rule of the form $X \Rightarrow Y$, where X and Y are events, which states that when X occurs in the database so does Y, with a certain probability (coined the *confidence* of the rule). A well-known application of association rules is in market basket data analysis.

The problem of mining association rules was first introduced by Agrawal et. al. in [1], and later broadened by Agrawal et. al. in [2], for the case of databases consisting of categorical attributes alone. Categorical association rules are rules where the events X and Y, on both sides of the rule, are the appearance of given categorical items. In this case, we wish to find all rules with confidence and support above user-defined thresholds (*minconf* and *minsup*). Several efficient algorithms for mining categorical association rules have been published (see [2], [6], [8] for just a few examples). A variation of categorical association rules was recently introduced by Brin et. al. in [3]. Their new definition is based on relating to associations as statistically interesting correlations. In all, the problem of mining categorical association rules is well understood and extensively researched, on both the algorithmic and conceptual levels.

Quantitative Association Rules. In practice the information in many, if not most, databases is not limited to categorical attributes, but also contains much quantitative data. Unfortunately, the definition of categorical association rules does not translate directly to the case of quantitative attributes. It is therefore necessary to provide a definition of association rules for the case of a database containing quantitative attributes. Srikant et. al. [7] extended the categorical definition to include quantitative data. The basis for their definition is to build categorical events from the quantitative data by considering *intervals* of the numeric values. Thus, each basic event is either a categorical item or a range of numerical values. An example of a rule according to this definition would be:

$$\text{sex} = \text{female and age} \in [20,30] \Rightarrow$$
$$\text{wage} \in [\$5, \$10] \text{ (conf. 85\%)}$$

Given this definition, [7] provides an algorithm which approximately finds all rules by employing a discretization technique. In addition, [7] provides an *interest* filter, aimed at reducing the problem of many similar rules.

While the [7] definition and algorithm for quantitative association rules provides a strong tool for mining quantitative data, there are also several drawbacks to their approach. First and foremost, the use of intervals as the basic means for describing the distribution of quantitative values can be limited and, at times, mis-

*Address: Bar-Ilan University, Department of Computer Science, Ramat Gan, Israel 52900.

†Work carried out while at Bar-Ilan University. Address: The Weizmann Institute of Science, Faculty of Mathematics and Computer Science, Rehovot 76100, Israel.

KDD-99 San Diego CA USA

leading. For example, the rule "height $\in$ [100cm,150cm] $\Rightarrow$ age $\in$ [0,14] (70%)" may be true even though few, if any, children under the age of one are 100cm tall (see Section 4.3 for real-world examples). In addition, the [7] definition often results in an exponential blowup of the number of rules, as the left-hand side of any given rule can always be enlarged. Hence, they place an a priori restriction on the maximum support of a rule (*maxsup*), partially solving this problem. Finally, the discretization employed in the mining algorithm results in loss of information. In particular, the algorithm can only approximate the best rules (see [7] for details).

Other work on this problem includes Zhang et. al. in [10] who use clustering methods to improve the partitioning of the quantitative attributes in the algorithm. Fukuda et. al. in [4] and Yoda et. al. in [9] also worked on the quantitative associations problem. However, their work is related to a different version of the problem and is focused more on prediction, rather than association rules.

A New Definition. In this paper we introduce a new definition of quantitative association rules, based on the *distribution* of values of the quantitative attributes. The new definition is a natural generalization of the categorical definition, when interpreted in the proper statistical terms. An example of a rule according to our new definition would be:

$$\text{sex} = \text{female} \Rightarrow \text{Wage: mean} = \$7.90 \text{ p/hr}$$
$$(\text{overall mean wage} = \$9.02)$$

saying that the average wage for females is \$7.90 dollars per hour. This rule is *interesting* as it reveals a group of people earning a significantly lower than average wage (\$9.02 p/hr). Our definition captures the notion of finding "interesting behavior", generating rules revealing extraordinary phenomena. We use accepted statistical tests to confirm the validity of the discovered rules. We present algorithms that do not use discretization, but rather view the quantitative attributes as continuous. Finally, we validate our definition through an in-depth evaluation of results.

Before giving our new definition, it would be helpful to backtrack a little and discuss the goal and structure of association rules in general. Association rules are designed to help us discover "interesting" phenomena or behavior in databases. This is accomplished by locating sets of transactions containing unexpected behavior. Each rule is comprised of a *left-hand side* and a *right-hand side*:

- The left-hand side of the rule is a description of a subset of the population.
- The right-hand side of the rule is a description of interesting behavior particular to the population described on the left-hand side.

Thus, the general structure of an association rule is:

$$\text{population-subset} \Rightarrow \text{interesting-behavior}$$

It is necessary to define what denotes "interesting behavior". In the categorical case, interesting behavior is a higher than usual incidence of certain attributes. Thus, for categorical attributes, *behavior* is naturally described by a list of items and the probability of their appearance. Statistically, this description is the probability distribution of the set of items, for the given population. So too, we argue, for a set of quantitative values the best description of its behavior is its distribution. For numerical values, mean and variance are comprehensive measures for describing a distribution. We therefore choose to describe the behavior of a set of quantitative values by calculating their *mean* and *variance*.

In order to ensure that we obtain rules that truly inform us of remarkable phenomena, we consider the behavior of a subset to be *interesting* if its distribution stands out from the rest of the population. We therefore say that a subset of the population displaying a distribution significantly different from that of its complement, either in terms of the mean or the variance, is recognized as interesting and noteworthy. For example, a possible association rule under the new definition would be:

$$\text{non-smoker and wine-drinker} \Rightarrow$$
$$\text{life expectancy} = 85 \text{ (overall} = 80)$$

Here, the interesting behavior is expressed in a dramatic increase in the mean. We use standard statistical methods to measure the significance of disparity between the distributions.

In summary, an association rule under the new definition is a rule of the form:

$$\textbf{population-subset} \Rightarrow$$
$$\textbf{mean or variance} \text{ values for the subset}$$

A rigorous definition is provided in Section 2. In Section 3 we show that this definition is also computationally workable. We need not use discretization and we have no exponential blowup in the number of rules.

Other Statistical Measures. We note that a similar definition can be established using any other measure of the statistical distribution (e.g. median). Thus, our definition actually provides a framework for an entire family of association rules. Athough mean-based rules are the most natural, other measures provide important information. The variance of a subset, for example, points to the homogeneity of those included in the subset. We choose to focus on the mean and variance measures, as they are the most commonly used measures, and tend to provide most of the interesting information regarding the distribution.

Sample Results

In the coming sections we will present exact definitions, algorithms and evaluation for our new concepts. However, before delving into these details, we first present some sample results, obtained from an actual database.

We applied our algorithm to a database called *Determinants of Wages from the 1985 Current Population Survey in the United States* (the database may be found at http://lib.stat.cmu.edu/datasets). The database contains 534 transactions and 11 attributes (7 categorical and 4 quantitative). Here are some of the rules discovered with 95% statistical confidence. Note that the mean wage overall is \$9.02 p/hr.

Sex = female ⇒
Wage: mean = \$7.90 p/hr
Sex = female and South = Yes ⇒
Wage: mean = \$6.30 p/hr

The second rule is a sub-rule (defined in section 2.2) of the first and shows that although females overall are paid lower wages, in the South of the USA the situation is much worse. We hope that things have improved somewhat since 1985. The other side of the coin of these two rules is the next rule also found:

Sex = male and Race = White ⇒
Wage: mean = \$10.33 p/hr

Other rules linked Education (as in years of formal education) to Wage and justify the argument that on average, education improves earning power:

Education ∈ [2, 13] years ⇒
Wage: mean = \$7.52 p/hr

Education ∈ [14, 18] years ⇒
Wage: mean = \$11.64 p/hr

In a different and somewhat unexpected direction, a rule connecting Education to Age was found. The mean age of the population was 37 years.

Education ∈ [3, 10] years ⇒
Age: mean = 46 years old

This rule shows us that those with very little schooling are on average far older, a sign of positive progress in society. Usually, most research on this database would be limited to factors affecting a person's wage. Through our data-mining technique we exposed interesting information which we would not initially have thought to look for. The *Wages* database provided us with interesting results. Clearly, this gives only a flavor of the rules. A rigorous evaluation of the quality of the rules discovered is provided in Section 4.

Outline. In the next section we develop a formal framework for our definition. In section 3 we present efficient algorithms for some cases of quantitative rules and demonstrate the computational viability of mining for these rules. In section 4 we present experimental evaluation and analysis of results obtained, including comparisons to [7]. We use real-world databases and present evaluations by domain experts who provided us with the databases and used the results for their research. Finally, we present open questions and discussion for further work.

2 Definitions

An association rule contains a left-hand side and a right-hand side. In the most general form, the left-hand side of the rule is a description of a subset of the database, while the right-hand side provides a description of outstanding behavior of this subset. This general structure gives rise to many different concrete rule types, depending on the type of subset used in the left-hand side, and the description used for the right-hand side. In this paper we focus on two specific types, which we found to be most useful in practice, and algorithmically manageable. We also provide a general definition of sub-rules, which, as we shall see, are essential for providing both comprehensive and exact information.

Notations. Let $E = e_1, ..., e_m$ be the set of attributes (or fields) for a database D of transaction. Let $E_Q \subseteq E$ be the set of quantitative attributes, $E_C \subseteq E$ the set of categorical attributes, and C the set of all possible categorical values. Each transaction in D is an m-tuple $t = (<e_1,v_1>,...,<e_m,v_m>)$, of attributes and corresponding values (i.e. for each i, if e_i is categorical then $v_i \in C$, and if e_i is quantitative then $v_i \in \mathcal{R}$.

2.1 Basic Rule Types

2.1.1 Categorical ⇒ Quantitative Rules

The first type of rule we consider are rules where the left-hand side is a set of categorical attributes, and the right-hand side is the mean value for some quantitative attributes.

The Left-Hand Side. The left-hand side of the rule is a set $X \subseteq E_C \times C$ of categorical attributes and matching categorical values. The set X, which is the *profile*, defines a subset of the database. For a transaction $t = (<e_1,v_1>,...,<e_m,v_m>)$, we say that t *has profile* X if $X \subseteq t$, i.e. t coincides with X whenever X is defined. We denote the set of transactions with profile X by T_X.

The Right-Hand Side. The right-hand side of a rule provides the mean values of part or all of the quantitative attributes, with the mean taken over the transactions which match the profile of the left-hand side. Formally, for a list of quantitative attributes J, and a set of transactions T, we denote by $Mean_J(T)$ the vector of mean values of the attributes in J for the set T. The right-hand side of the rule is $Mean_J(T_X)$ for some $J \subseteq E_Q$.

Significance. A rule is only interesting if the mean for the subset T_X is significantly different from the rest and is therefore unexpected. We therefore compare the mean in T_X to the mean of the complement, i.e. $D - T_X$. Note, however, that although the two means may be numerically different in the database, we may not have any statistical evidence to infer a difference in the real populations. Thus, we use statistical tests to establish the significance level of the difference. In the case of mean-values we use the standard Z-test, to establish significance of the inequality of the means. We test the hypothesis that the mean of the two subsets are not equal (the null hypothesis) against the hypothesis claiming a difference of means. A rule is then considered *significant* if the null hypothesis is rejected with confidence above a set threshold (usually 95%). Formally, we say that $Mean_J(T_X)$ is *significantly different from* $Mean_J(T_Y)$, denoted $Mean_J(T_X) \not\approx Mean_J(T_Y)$, if for every $e \in J$ the means of attribute e in T_X and T_Y are different, based on the appropriate statistical test.

We are now ready to define mean-based quantitative association rules.

Definition 1 *A (mean-based)* categorical to quantitative association rule *is of the form* $X \Rightarrow Mean_J(T_X)$, *where* X *is a profile of categorical attributes* ($X \subseteq E_C \times C$), J *is a set of quantitative attributes* ($J \subseteq E_Q$), *and* $Mean_J(T_X) \not\approx Mean_J(D - T_X)$.

Minimum Difference. Sometimes, finding populations for which the means are merely different does not lead to interesting information. If we were to discover, for example, a group of people with life expectancy three days more than the overall population, it may not be of interest to us even if it passes a statistical test. We therefore allow a user-defined minimum difference parameter, denoted *mindif*, and we say that $Mean_i(T_X) \not\approx Mean_i(T_Y)$ if there is statistical support for inferring that $|Mean_i(T_X) - Mean_i(T_Y)| > mindif$.

Categorical $\Rightarrow$ Quantitative Rules Based on other Distribution Measures. The rules defined here provide a tool to discover interesting behavior of the distribution with regards to its mean value. An analogous definition can be provided given any other measure of the distribution, e.g. variance, median. For a given measure M (e.g. M = Variance), an *M-based association rule* is of the form $X \Rightarrow M_i(T_X)$. The rest of the definitions carry over directly from the basic mean-based rules by changing Mean to M throughout. For a given measure M, significance of inequality, denoted $M_J(T_X) \not\approx M_J(T_Y)$, is defined with the proper test for this measure, e.g. the F-test for variance. The algorithm outlined in section 3.2 is correct for any measure and has been implemented for variance-based rules as well.

2.1.2 Quantitative $\Rightarrow$ Quantitative Rules

Next, we consider rules for which both the left-hand side and the right-hand side are comprised of quantitative attributes. We provide a definition for the case where both sides are comprised of a *single* quantitative attribute.

The **left-hand side** of the rule is a triplet $X = (e, r_1, r_2)$, where e is a quantitative attribute, and r_1, r_2 are real values, $r_1 \leq r_2$. We call X the *profile*. We say that a transaction t has profile X if the value of t for the attribute e is within the interval $[r_1, r_2]$. We denote by T_X the transactions with profile X. The **right-hand side** of the rule is a quantitative attribute j ($j \neq e$) and its mean value $Mean_j(T_X)$. As before, **significance** is ensured by demanding that $Mean_j(T_X) \not\approx Mean_j(D - T_X)$.

At this stage we would like to simply define a rule as one of the form $X \Rightarrow Mean_j(T_X)$ where $Mean_j(T_X) \not\approx Mean_j(D - T_X)$. However, in this case, not all rules of this type are desirable. We thus introduce additional concepts.

Consider the following (fictitious) database. Assume that the average weight of the entire population is 80 kilograms. The rules below may be deduced from the database.

Age	Weight
...	...
50	80
60	90
70	90
80	90

age $\in$ [60,80] $\Rightarrow$ average = 90.0
age $\in$ [70,80] $\Rightarrow$ average = 90.0
age $\in$ [50,80] $\Rightarrow$ average = 87.5

It is clear that the first rule is the only one we wish to obtain. The second rule is contained in the first rule adding no new information and is therefore superfluous. We point out that if the mean of two adjacent ranges is above average, then the union is definitely above average. We therefore wish to obtain the widest possible rule (if a part of the range contains a different average, then this will be revealed as a sub-rule).

Notice that even though the third rule is even wider than the first one, it is nevertheless undesirable. This is because it is generated by appending a non-interesting region to another rule. This clearly detracts from the accuracy and leads to misleading rules. This motivates the following formal definitions.

Irreducible and Maximal Rules. Consider a rule $X \Rightarrow Mean_j(T_X)$ with $X = (e, a, b)$, and suppose that $Mean_j(T_X)$ is above average. We now define the notions *irreducibility* and *maximality*. Intuitively, the rule is *irreducible* if we cannot cut the interval $[a, b]$ in

the middle and obtain an interval which is not above average. This property ensures that we do not have non-interesting regions appended to the edge of the rule (the third rule in the above example is reducible). A rule is *maximal* if we cannot enlarge the interval $[a, b]$ either to the right or the left and still remain with an irreducible rule with above average distribution. Maximal rules are therefore the largest "good" rules and provide the most concise presentation (the second rule in the above example is not maximal).

Formally, we say that the rule is *irreducible* if for any $a < c < b$, setting $Y = (e, a, c)$, and $Z = (e, c, b)$, then both $Mean_j(T_Y)$ and $Mean_j(T_Z)$ are above average. We say that the rule is *maximal* if for any $c < a$ $(c > b)$ setting $Y = (e, c, b)$ $(Y = (e, a, c))$ then $Mean_j(T_Y)$ is either not above average or reducible. The definition for below average rules is analogous. A rule of this type must be both maximal and irreducible.

We now have all the necessary concepts to define rules of this type:

Definition 2 *A (mean-based)* quantitative to quantitative association rule *is a maximal and irreducible rule of the form* $X \Rightarrow Mean_j(T_X)$ *where* X *is a profile for a single quantitative attribute* $(X \in E_Q \times \mathcal{R} \times \mathcal{R})$ *,* j *is a quantitative attribute* $(j \in E_Q)$*, and* $Mean_j(T_X) \not\approx Mean_j(D - T_X)$.

Remark. The profile of a rule of this type is in the form of a range. The profile classifies the transactions for which the rule refers to and a range provides a clear partition into those belonging and those not. For example, by saying that a phenomenon occurs for people between the ages of 10 and 20, it is clear to whom it applies. However, if we were to say that a phenomenon occurs for people of average age 15, then it would not be clear exactly what population this rule is based on and who it describes.

2.2 Sub-Rules

We have so far provided a framework for defining rules, and definitions for two important categories of them. However, not all rules are desired. We are interested in finding the key factors of extraordinary behavior of a population. Consider the following set of rules, where the overall life expectancy is 70 years:

smoker $\Rightarrow$	life expectancy = 60
male and smoker $\Rightarrow$	life expectancy = 60
smoker and wine-drinker $\Rightarrow$	life expectancy = 70

Both the second and third rules are more specific than the first rule and are therefore *contained* in it. Intuitively, it is clear to us that the second rule is undesirable as it introduces a factor that does not contribute to the interesting phenomenon, while the third one is desirable as it adds new information. The third rule qualifies as a *sub-rule* of the original rule as it is statistically different from the original rule. We now formalize these ideas.

The concept of *rule containment* is defined as follows. Let X and Y be profiles containing categorical attributes only (as in 2.1.1). Then we say that rule $Y \Rightarrow Mean_J(T_Y)$ contains rule $X \Rightarrow Mean_J(T_X)$, if $Y \subseteq X$. If Y and X contain a single quantitative attribute only (as in 2.1.2) where $Y = (e, a, b)$ then we say that $Y \Rightarrow Mean_J(T_Y)$ contains $X \Rightarrow Mean_J(T_X)$ if $X = (e, c, d)$ with $a \leq c \leq d \leq b$. The intuition is that Y covers a larger set of transactions. Containment with more complex profiles is generalized in the natural way.

Definition 3 *We now define* basic rules, *and* sub-rules*:*

1. *A rule is a* basic rule, *if it is not contained in any other rule.*
2. *A rule* $X \Rightarrow Mean_J(T_X)$ *is a* sub-rule *of* $Y \Rightarrow Mean_J(T_Y)$ *if:*
   (a) $Y \Rightarrow Mean_J(T_Y)$ *contains* $X \Rightarrow Mean_J(T_X)$
   (b) $Mean_J(T_X) \not\approx Mean_J(T_Y - T_X)$

A sub-rule $X \Rightarrow Mean_J(T_X)$ *is a* basic sub-rule *if it is not contained in any other sub-rule of* $Y \Rightarrow Mean_J(T_Y)$.

Note that a basic sub-rule is a basic rule with regards to the database T_Y.

As we have shown, contained rules which are not sub-rules (as their distribution measure is not significantly different) are undesirable. In the above example, the second rule is contained yet does not have a different mean, whereas the third rule is a sub-rule exactly because of its different mean with respect to the super-rule. We therefore wish to find all basic rules, their basic sub-rules, the basic sub-rules of these sub-rules, and so on. This set of rules is what we aim to find in the quantitative association rules problem.

Definition 4 *We recursively define* desired rules *(those which we wish to obtain):*

1. *Any basic rule is* desired.
2. *Any basic sub-rule of a desired rule is also* desired.

In Section 3 we provide algorithms to find all desired rules.

2.3 Generalizing Categorical Association Rules

Our definition of association rules is actually a generalization of the definition of categorical association rules. In the categorical case, the left-hand and right-hand sides are defined by lists of items X and Y. The measures of significance used by [2] are support and confidence. Since the appearance of a set of items is a

Bernoulli random variable, the mean of "Y given X" is exactly the confidence of the rule. The rules defined by [3] are defined in the same way with a different significance measure (a statistical χ^2 test is used).

3 Algorithms for Finding Rules

Efficient algorithms for finding quantitative association rules are provided for two types of rules:

1. Rules $X \Rightarrow Mean_J(T_X)$ where both X and J contain a single quantitative attribute only.
2. Rules $X \Rightarrow M_J(T_X)$ where $X \subseteq E_C$ (only categorical attributes) and $J \subseteq E_Q$ (only quantitative attributes). There is no limit on the number of attributes in X or J. The algorithm is correct for any measure M.

3.1 Finding rules from one Numerical to one Numerical attribute

In this section we introduce an algorithm that finds rules based on the mean distribution measure. Our algorithm finds rules between two given quantitative attributes. This is then applied for every pair, thereby obtaining all rules of this kind. We enable the user to specify a minimum support parameter here. This is not necessary at all for computational reasons as will be shown later. However, rules based on very few data points may be inconsequential, even if they are true. Furthermore, it can be shown that multiple hypothesis testing with a very small minimum support can lead to a high error (see section 5).

Algorithm Motivation. Let i and j be a pair of quantitative attributes. Note that if we sort the database by attribute i, then any above or below-average continuous region of values in j is a rule (provided it passes the necessary statistical test). This is because attribute i is sorted, and therefore any continuous region is a range. However, we must also ensure that the rule is irreducible and maximal.

Our algorithm is based on the following simple idea: if the regions $[a, b]$ and $[b, c]$ are both above or below average, then so too is the region $[a, c]$. It is completely symmetrical to search for above or below average regions, we will therefore refer only to above-average from now on. Further note than when we say "above-average" we mean above the overall mean plus *mindif*.

The Window Procedure. The following data-driven procedure, called "Window", accepts as input an array of values and the average of the values in the array (plus *mindif*). The input array is the array of values of attribute j, sorted by i. We execute a single pass to find all rules from i to j. The procedure works with two windows or regions: A and B. A is an irreducible above-average region (this remains invariant throughout). B is an adjacent region that may be joined to A if $A \cup B$ will also be irreducible.

To begin, we initialize A to the first above-average value in j (this is clearly irreducible) and B is empty. Given A and B, we add the next value in j to B. There are three possibilities at this stage:

1. The average of the B region is above-average: we join B to A (emptying B).

   A is still irreducible: separating $A \cup B$ into A and B or into A_1 and $A_2 \cup B$ obviously leaves two above-average regions. Separating $A \cup B$ into two halves $A \cup B_1$ and B_2 also leaves two above-average regions because if $A \cup B_1$ is not above-average, step (2) below would already have happened, and if B_2 is not above-average then B_1 must be above average and this step would have happened earlier.

2. The average of the region including A and B together is not above-average: the A region is a potential rule (and not the B region). We continue by emptying B and initializing A to be the first above-average value after the region of the potential rule.

   Note that no rule can contain $A \cup B$ at the beginning, as the rule would then be reducible. Furthermore, A certainly has maximal support by definition for if we expand A, either the region will not be above-average or it will be reducible.

3. If neither of the above is true, then we simply continue by adding the next value in j to B.

Window is shown in figure 1. We run Window twice in order to find both above and below average rules. Upon finding a potential rule, we execute a Z-test to determine whether or not we accept the rule or not. If yes, we call Window recursively with the input array as the j values supporting the accepted rule. The input average is the average of the rule (plus/minus *mindif* appropriately). This recursive call finds all sub-rules of the rule and so on. If the rule is not accepted, we simply continue searching for rules in the following regions.

Complexity Analysis. For a given pair of attributes, the time taken for n transactions is $O(n \log n)$ for the sort, plus the complexity of the Window algorithm. The complexity of Window is clearly upper-bound by $O(n)$ times the number of levels of rules (i.e. the number of recursive calls). Since the number of levels is expected to be low (as experience has shown), we effectively maintain linear complexity. Note that minimum support has no effect on the running time, enabling us to find rules with very low support. For k quantitative attributes, the time taken to find all rules of this type is therefore $O(k \cdot n \log n + k^2 \cdot n)$.

```
Input: an array Array and a number a = average + mindif

Window(Array, a)
 current ← index of the beginning of Array
 While (current < end_of_Array)
 {
    current ← next above average value
    Initialize parameters for A and B regions
    While (AVERAGE(A, B) ≥ a) // Weighted average
    {
       Update B to include Array[current]
       current ← current + 1
       if (AVG(B) > a)
          Join A to B and empty B
    }
    If the values in region A pass a Z-test:
       Add the appropriate rule to the set of results
       Call Window(Array[A region], AVG(A) ± mindif)

    current ← first index after the A region
 }
```

Figure 1: Window Procedure for finding "Numerical ⇒ Numerical" Rules

We note that with very large databases, this sort may take considerably longer as it needs to be executed in secondary memory.

3.2 Finding rules from Categorical to Numerical attributes

Due to lack of space in this abstract we provide an outline for this algorithm only. We note that the left-hand side of a rule in this case is essentially a frequent set and this property forms the basis of our algorithm.

Algorithm Outline. The algorithm has three distinct stages:

1. Find all frequent sets of categorical items only, using known algorithms such as Apriori (see [2]).
2. For all quantitative attributes, calculate the distribution measure (mean/variance) for each frequent set, using the hash-tree data structure presented in [2]. One pass over the database is sufficient.
3. Find all non-contained rules and sub-rules. For every frequent set X and quantitative attribute e, it remains to check if $X \Rightarrow Mean_e(T_X)$ and $X \Rightarrow Variance_e(T_X)$ are basic rules or sub-rules or neither. We do this by traversing a lattice of the frequent sets while keeping track of containment relations between sets and the sub-rule hierarchy.

We note that the ideas in sections 3.1 and 3.2 may be combined in order to find rules with profiles containing many categorical and a single numerical attribute. For a given frequent set X, we run Window on T_X. We may run Window in parallel on each frequent set and efficiently achieve the desired result.

4 Experimental Evaluation

4.1 A Rigorous Evaluation

Measuring Success. A major problem confronting data mining researchers is the question of how to measure success. In any evaluation it is necessary to measure both the correctness of the rules obtained and the *interestingness* of those rules to the user. A rule saying that "abortion ⇒ female" is certainly true, but is completely uninteresting.

We deal with the issue of *correctness* with statistical inference techniques. On the other hand, evaluating how interesting the rules obtained are is of great difficulty. We, as computer scientists, are certainly unable to judge whether rules found are of interest to the user or not. As our goal is to help the user, we believe that the only way to measure success is to ask the end user himself. He is not only the most *objective* judge, but the only one qualified to judge at all. He can inform us on whether or not the type of rule found is helpful, if it revealed information new to him and what percentage of the rules found are truly interesting.

Our Evaluation. We therefore tested our concept on a real-world database and had a domain expert perform an in-depth evaluation of the results. The database we mined is from the field of linguistics and was built during a study on the English writing habits of non-native English speakers. The study was conducted by Prof. Joel Walters of the English Department in Bar-Ilan University. Previously Prof. Walters researched the database extensively using standard statistical tools such as SPSS. We presented Prof. Walters with the association rules we discovered, and asked him to evaluate the rules. For each rule, Prof. Walters categorized the rule as: non-interesting, interesting, or very-interesting. Among the interesting and very-interesting rules, he marked if he would have otherwise found them or not (e.g. using SPSS).

Description of the Database. The database is based on a study involving essay writing under different conditions. Each transaction in the database contains data extracted from an essay and the background information on the author. The data extracted from the essay includes part-of-speech measures (as in percentage of words used which are nouns, adjectives etc.) and lexical measures (relating to the level of words used, the level of variation, originality and many others). The database contains 643 transactions and 42 attributes: 15 categorical and 27 quantitative. With approximately 27,000 entries and 42 different factors (making for many hundreds of possible patterns), the database is large and computerized tools are necessary.

	Categorical ⇒ Numerical	Numerical ⇒ Numerical	Overall
Number of Rules:	70	284	354
Not Interesting:	50 or 71%	178 or 63%	228 or 64%
Interesting:	16 or 23% (8)	86 or 30% (62)	102 or 29% (70)
Very Interesting:	4 or 6% (4)	20 or 7% (18)	24 or 7% (22)

Table 1: *Interestingness* classification of rules

The Evaluation Results. The results of the quantitative evaluation by the domain expert are summarized in Table 1 (the numbers in brackets are the number of rules that would not have otherwise been found).

Overall we see that 36% of rules were interesting or very interesting, 26% of all rules would not have been found using the standard hypothesis checking model. 7% of the rules were graded very interesting and would not have otherwise been found. This is *very* high for an automated tool and the result is critical to the usefulness of the method. Users are unlikely to use tools which provide interesting results hidden amongst endless junk.

Rule Complexity. If we further look at the breakdown of interesting rules within the Categorical ⇒ Numerical rules, more than 50% of rules with one categorical attribute in the profile were graded interesting! On the other hand, those with more than one attribute in the profile were judged not-interesting in 86% of the cases. This is most likely due to the difficulty in understanding complex rules and shows that most interesting rules have simple profiles. This supports our claim that our algorithms cover most of the interesting cases.

A Qualitative Evaluation. The strength of our technique can be seen by viewing a number of results judged to be interesting by our evaluator. Due to lack of space, we bring only one example. Some participants in the study were given a source text and were asked to base their essay on it and others were not. We present a suprising rule regarding the effect of these source texts. Our evaluator judged the following rule to be interesting and claimed that he would not have found it using standard statistical tools:

First Language = Russian AND No Source Text ⇒
Use of "the": mean = 3.9% (overall mean = 6.7)

This rule tells us that Russians who were not presented with a source text used the word *the* well below average. It is a known fact to Linguists that the Russian language has no definite article. Therefore, we are not surprised to see that Russians use the word *the* less. However, this was not inferred from the database. Rather, we found that only when the participants had no source text to base on, they fell back on their Russian habit of not using a definite article. On the basis of this rule Prof. Walters found that Russians given a source text used the word *the* 8.1% of the time in contrast to the 3.9% result shown in the rule. His conclusion was that essay writing based on source texts should be used for teaching use of the definite article.

This rule is an important example of where relying only on a priori hypotheses is not enough. We would not have found that Russians use the word *the* less and we would not have guessed that the source texts were the key. As a result, we learnt something new in discovering that having a text to base your writing on can also improve style. This discovery was of great importance to Prof. Walters.

Running Time. We ran our tool on a Pentium-Pro with 128Mb RAM. With minimum support 40, the overall time taken on the *Linguistics* database (643 transactions, 42 attributes) was just 10.1 seconds. Of this time, 0.79 seconds was spent on Window (finding rules: 1 Numerical ⇒ 1 Numerical, not including time to load the data into memory). For a minimum support of 20, the overall time taken was 23.8 seconds, the time spent of Window was 0.81 seconds.

The Statistical Tests. We found that the use of statistical tests to validate the accuracy of potential rules is crucial. In the *Linguistics* database, 29,959 potential rules were discovered by the algorithm, but were rejected due to lack of statistical confidence (a confidence level of 95%). Only 354 rules were accepted. The difference was even more extreme for a minimum support of 20, where we accepted 1,018 rules and rejected 101,449. A person may view 600 rules in a reasonable amount of time. With 30,000 rules however, we have engaged in data explosion rather than data reduction and we cannot be of any help to the user.

Summary. We found many rules determined to be truly interesting and revealing to the user. A very high percentage of these rules were not likely to have been discovered at all without our data mining tool. These two results show that our notion of quantitative associations fulfills the ultimate goals of the data mining concept. Furthermore, our rules are easily understood and interpreted, concise even when they are complex and most of all really do describe exceptional and therefore noteworthy behavior.

4.2 Scalability

We also checked the scalability of our algorithms. For this we used Synthetic Data Sets (created by the IBM Quest Synthetic Data Generation Code). We created a database with 9 attributes, 3 categorical, 6 quantitative. We used a minimum support of 40. The results are depicted in Table 2 below.

Number of transactions	Overall time	Time for Window	Number of rules found
10,000	14.234	0.812	170
20,000	32.204	1.829	272
30,000	56.078	2.844	399
40,000	86.657	3.922	548
50,000	126.08	5.016	650

Table 2: *Scalability* of the algorithm

4.3 A Comparative Evaluation of [7]

Remember that in [7] a quantitative association rule is defined as a rule $X \Rightarrow Y$ with a certain support and confidence, where X and Y contain categorical items or numerical ranges. Their algorithm is based on mapping the problem to the categorical case by way of discretization, finding all association rules and then filtering superfluous nested rules. We now present an evaluation (rendered by our expert) of the rules generated by [7] and examples of some of the problems. We used the *Linguistics* database as our basis for the evaluation.

A Quantitative Evaluation. In order to create a fair comparison of [7] rules to ours, we limited the search for rules with one attribute on each side of the rule. This was so we could choose a relatively high maximum support, low minimum support and low K (otherwise we would encounter extreme computational difficulties, especially with a database of 42 attributes). We also wished to limit the number of rules we found to something that could be realistically evaluated. In order to do this, we chose 10 pairs of attributes uniformly randomly (with the condition that at least one of the attributes was quantitative) and obtained rules from these pairs only. We then evaluated these rules as a sample of the set of all rules of this type.
After some fine tuning we settled on the following input parameters:

Minimum Support	= 40 or 6%
Maximum Support	= 0.7
Completeness Level (K)	= 9
Interest Level (R)	= 1.5
Minimum Confidence	= 0.6

The minimum support was chosen to be the same as for our algorithm and we chose K=9 in order to reduce the number of rules found.

We obtained 81 rules in the output. As the number of possible pairs of attributes equals 756 (for 15 categorical and 27 quantitative attributes) we expect approximately 6,000 rules from this limited type alone. This is a significant problem to anyone who must review the results (note that we obtain 354 rules in total with minsup = 40).
The results of the evaluation are as follows:

Number of rules:	81 (chosen *randomly*)
Not interesting:	80
Interesting:	1
Very interesting:	0

Overall, 1.2% of the rules were judged to be interesting.

Discussion. We will now present two examples demonstrating some of the conceptual drawbacks to the [7] definition. As we have mentioned, range is a weak measure for describing a distribution. The following rule we found illustrates this point:

Lexical Variation $\in$ [27.91, 62.02] $\Rightarrow$
Lex. Originality $\in$ [4.27, 31.19] (sup 42%, conf 68%)

Lexical variation measures the diversity of vocabulary used by the participants. Lexical originality is a measure of how many words the participant used that no other participants in the study used. The above rule tells us that those with average and below lexical variation (the average is 59.2) have high lexical originality (the average is 6.7, which is included in the interval, but the interval extends far to the right). This rule may seem interesting and surprising but in actuality is very misleading. It is true that 68% of the values are between 4.27 and 31.19 but nearly all of those are below 13! In fact, the following rule was also found:

Lexical Variation $\in$ [27.91, 62.02] $\Rightarrow$
Lex. Originality $\in$ [2.91, 12.54] (sup 42%, conf 68%)

We checked the original data and found that only 39 transactions within the Lexical Variation range, have a Lexical Originality value above 12.54.

Another example of rules found to be misleading due to the distribution measure is the following:

English Proficiency = Advanced $\Rightarrow$
Length $\in$ [91, 214] (sup 6%, conf 50%)
English Proficiency = Fluent $\Rightarrow$
Length $\in$ [91, 214] (sup 25%, conf 50%)

The "English Proficiency" categories ranked the student's English skills according to accepted levels. The above pair of rules gives us the feeling that the different proficiencies "Advanced" and "Fluent" are actually very similar regarding their writing ability (as essay length is a known predictor of writing ability). This is a surprising rule and would be very interesting. However, consider the following statistics: the average essay

length in the entire study was 172, for "Advanced" was 108 (way below average, i.e. low proficiency) and for "Fluent" was 222 (a strong sign that this group is well skilled). This example strongly confirms our claim that describing a numerical distribution with a range and probability may be very misleading. Not only may we not find interesting behavior, we may be presented with rules which lead us to erroneous conclusions.

5 Discussion and Future Work

Generalizing Quantitative Association Rules. We introduced a general definition for quantitative association rules in the form of *"Profile ⇒ Significant Distribution"*. However, we developed the definitions and algorithms necessary for rules of two specific cases only. These cases proved to be very useful and are important categories of rules. However, we would like to see a truly general definition of quantitative association rules, combining categorical attributes in the distribution as well. We note the difficulty in expanding the profile to include two quantitative attributes. Firstly, it is easily shown that it is not possible to use one-dimensional rules in order to find *all* two-dimensional rules. Secondly, a conceptual difficulty arises in that the union of two overlapping, above-average rules is not necessarily above-average. This may lead to the undesirable property of many overlapping rules.

Multiple Hypothesis Testing. A general problem in data mining is that of multiple hypothesis testing. In statistical terms, these are called *multiple comparison procedures* and results are likely to be obtained even on random databases. We ran the following experiment in this context. We applied independent random permutations on each column of our test database (*Linguistics*) obtaining a database with no correlations, yet with the same original distributions for each attribute separately. This was done 50 times obtaining 50 random databases. We then ran our algorithm on these databases and counted the number of rules obtained. We found that the random variable of the number of rules has a Poisson distribution with mean 16.7. Thus the probability of obtaining more than 30 rules if correlations do not really exist in the database is less than 0.001. As we found more than 300 rules this is overwhelming evidence that the results are not random. However, we stress that in-depth statistical research is still required in order to provide a strong theoretical background for these types of data-mining procedures.

Other Statistical Tests. An interesting question for future work is that of the effect of the specific statistical tests used. We used the Z-test as it is the most natural mean test, especially as we need not assume anything about the distribution of the values. However, the effects of using other tests and methods, or even a combination of them, is an interesting and important issue for continuing research.

6 Acknowledgments

We would like to thank Ronen Feldman for his invaluable contributions. The in-depth evaluation is due to Joel Walters and we thank him for his great investment of time and effort. Finally, we thank Lawrence Freedman for his enlightening discussions regarding the statistical background necessary for our work.

References

[1] R. Agrawal, T. Imielinski and A. Swami. Mining association rules between sets of items in large databases. *Proc. of the 1993 ACM SIGMOD Intl. Conference on Management of Data*, pp 207-216.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. *Proc. of the 20th Intl. Conference on VLDB*, 1994.

[3] S. Brin, R. Motwani and C. Silverstein. Beyond Market Baskets: Generalizing Association Rules to Correlations. *Proc. of the 1997 ACM SIGMOD Conference on Management of Data*, 1997.

[4] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama. Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms and Visualisation. *Proc. of the 1996 ACM SIGMOD Conference on Management of Data.*

[5] Lindgren, Bernard W. *Statistical Theory.* Macmillan Publishing Co., Inc. New York, 1976.

[6] H. Mannila, H. Toivonen and A. I. Verkamo. Efficient Algorithms for discovering association rules. *KDD-94: AAAI Workshop on Knowledge Discovery in Databases*, pp 181-192, 1994.

[7] R. Srikant, R. Agrawal. Mining Quantitative Association Rules in Large Relational Tables. *Proc. of the ACM SIGMOD Conference on Management of Data*, 1996.

[8] H. Toivonen. Sampling Large Databases for Association Rules. *Proc. of the 22nd VLDB Conference*, 1996.

[9] K. Yoda, T. Fukuda, Y. Morimoto, S. Morishita, T. Tokuyama. Computing Optimized Rectilinear Regions for Association Rules. *Proc. of KDD '97*, August 1997.

[10] Z. Zhang, Y. Lu and B. Zhang. An Effective Partitioning-Combining Algorithm for Discovering Quantitative Association Rules. *Proc. of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1997.