SYMBOLIC DATA ANALYSIS: DEFINITIONS AND EXAMPLES

L. Billard
Department of Statistics
University of Georgia
Athens, GA 30602 -1952 USA

E. Diday
CEREMADE
Universite de Paris 9 Dauphine
75775 Paris Cedex 16 France

Abstract

With the advent of computers, large, very large datasets have become routine. What is not so routine is how to analyse these data and/or how to glean useful information from within their massive confines. One approach is to summarize large data sets in such a way that the resulting summary dataset is of a manageable size. One consequence of this is that the data may no longer be formatted as single values such as is the case for classical data, but may be represented by lists, intervals, distributions and the like. These summarized data are examples of symbolic data. This paper looks at the concept of symbolic data in general, and then attempts to review the methods currently available to analyse such data. It quickly becomes clear that the range of methodologies available draws analogies with developments prior to 1900 which formed a foundation for the inferential statistics of the 1900's, methods that are largely limited to small (by comparison) data sets and limited to classical data formats. The scarcity of available methodologies for symbolic data also becomes clear and so draws attention to an enormous need for the development of a vast catalogue (so to speak) of new symbolic methodologies along with rigorous mathematical foundational work for these methods.

1 Introduction

With the advent of computers, large, very large datasets have become routine. What is not so routine is how to analyse the attendant data and/or how to glean useful information from within their massive confines. It is evident however that, even in those situations where in theory available methodology might seem to apply, routine use of such statistical techniques is often inappropriate. The reasons are many. Broadly, one reason surrounds the issue of whether or not the data set really is a sample from some populations since oftentimes the data constitute the "whole", as, e.g., in the record of all credit transactions of all card users. A related question pertains to whether the data at a specified point in time can be viewed

as being from the same population at another time, as, e.g., will the credit card dataset have the same pattern the next "week" or when the next week's transactions are added to the data already collected?

Another major broad reason that known techniques fail revolves around the issue of the shear size of the data set. For example, suppose there are n observations with p variables associated with each individual. Trying to invert an $n \times n$ state matrix $\boldsymbol{X}$ when n is measured in the hundreds of thousands or more and p is a hundred or more, whilst theoretically possible, will be computationally heavy. Even as computer capabilities expand (e.g., to invert larger and larger matrices in a reasonable time), these expansions also have a consequence that even larger data sets will be generated. Therefore, while traditional methods have served well on the smaller data sets that dominated in the past, it now behooves us as data analysts to develop procedures that work well on the large modern datasets, procedures that will inform us of the underlying information (or knowledge) inherent in the data.

One approach is to summarize large data sets in such a way that the resulting summary data set is of a manageable size. Thus, in the credit card example instead of hundreds as specific transactions for each person (or credit card) over time, a summary of the transactions per card (or, per unit time such as a week) can be made. One such summary format could be a range of transactions by dollars spent (e.g., \$10 - \$982); or, the summary could be by type of purchase (e.g., gas, clothes, food, ...); or, the summary could be by type and expenditure (e.g., {gas, \$10 - \$30}, {food, \$15 - \$95}, ...); or, etc. In each of these examples, the data are no longer single values as in traditional data such as, in this example, \$1524 as the total credit card expenditure, or 37 as the total number of transactions, or etc., per person per unit time. Instead, the summarized data constitute ranges, lists, etc., and are therefore examples of symbolic data. In particular, symbolic data have their own internal structure (not present, nor possible, in classical data) and as such should thence be analysed using symbolic data analysis techniques.

While the summarization of very large data sets can produce smaller data sets consisting of symbolic data, symbolic data are distinctive in their own right on any sized data sets small or large. For example, it is not unreasonable to have data consisting of variables each recorded in a range such as pulse rate (e.g., $\{60, 72\}$), systolic blood pressure (e.g., $\{120, 130\}$) and diastolic blood pressure (e.g., $\{85, 90\}$) for each of $n = 10$ patients (or, for $n = 10$ million patients). Or, we may have $n = 20$ students characterized by a histogram or distribution of their marks for each of several variables mathematics, physics, statistics, ..., say. Birds may be characterized by colors e.g., Bird 1 = {black}, Bird 2 = {yellow, blue}, Bird 3 = {half yellow, half red}, ... That is, the variable 'color' takes not just one possible color for any one bird, but could be a list of all colors or a list with corresponding proportion of each color for that bird. On the other hand, the data point {black} may

indicate a collection of birds all of whom are black; and the point {yellow (.4), red (.6)} may be a collection of birds all of which are 40% yellow and 60% red in color, or a collection of which 40% are entirely yellow and 60% are entirely red, and so on. There are endless examples. In a different direction, we may not have a specific bird(s), but are interested in the *concept* of a black bird or of a yellow and red bird. Likewise, we can formalize an engineering company as having a knowledge base consisting of the experiences of its employees. Such experiences are more aptly described as concepts rather than as standard data, and as such are also examples of symbolic data. For small symbolic data sets, the question is how the analysis proceeds. For large data sets, the first question is the approach adopted to summarize the data into a (necessarily) smaller data set. Some summarization methods necessarily involve symbolic data and symbolic analysis in some format (while some need not). Buried behind any summarization is the notion of a symbolic concept, with any one aggregation being tied necessarily to the concept relating to a specific aim of an ensuing analysis.

In this work, we attempt to review concepts and methods developed variously under the headings of symbolic data analysis, or the like. In reality, these methods so far have tended to be limited to developing methodologies to organize the data into meaningful and manageable formats, somewhat akin to the developments leading to frequency histograms and other basic descriptive statistics efforts prior to 1900, which themselves formed a foundation for the inferential statistics of the 1900's. A brief review of existing symbolic statistical methods is included herein. An extensive coverage of earlier results can be found in Bock and Diday (2000). What quickly becomes clear is that thus far very little statistical methodology has been developed for the resulting symbolic data formats. However, in a different sense, the fundamental exploratory data analyses of Tukey and his colleagues (see, e.g., Tukey, 1977) presages much of what is currently being developed.

Exploratory data analysis, data mining, knowledge discovery in databases, statistics, symbolic data, even fuzzy data, and the like, are becoming everyday terms. Symbolic data analysis extends the ideas in traditional exploratory data analysis to more general and more complex data. Siebes (1998) attempts to identify data mining as the step in which patterns in the data are discovered automatically (using computational algorithms, e.g.), while knowledge discovery covers not only the data mining stage but also preprocessing steps (such as cleaning the data) and post-processing steps (such as the interpretation of the results). Obviously, it is this post-processing stage which has been a traditional role of the statistician. Elder and Pregibon (1996) offer a statistical perspective on knowledge discovery in data bases. Hand et al. (2000) defines data mining "as the secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners." The size of the database is such that classical exploratory data

analyses are often inadequate. Since some problems in data mining and knowledge discovery in databases lead naturally to symbolic data formats, symbolic data analyses have a role to play here also. The engagement of cross-disciplinary teams in handling large data sets (such as computer scientists and statisticians) is however becoming essential.

A distinction should also be drawn with fuzzy data and compositional data. Fuzzy data may be represented as the degree to which a given value may hold; see, e.g., Bandemer and Nather (1992), and Viertl (1996). Compositional data (Aitchison, 1986, 1992, 1997) are vectors of nonnegative real components with a constant sum; probability measures or histogram data are a special case with sum equal to one. These types of data can be written as symbolic data by taking into account the variation inside a class of units described by such data and by then using this class as a new unit.

The purpose of this paper is to review concepts of symbolic data and procedures of their analysis as currently available in the literature. Therefore, symbolic data, sometimes called "atoms of knowledge" so to speak, are defined and contrasted with classical data in Section 2. The construction of classes of symbolic objects, a necessary precursor to statistical analyses when the size of the original data set is too large for classical analyses, or where knowledge (in the form of classes, concepts, taxonomies and so forth) are given as input instead of standard data is discussed in Section 3. In Section 4, we briefly describe available methods of symbolic data analysis, and then discuss some of these in more detail in subsequent sections. What becomes apparent is the inevitability of an increasing prevalence of symbolic data, and hence the attendant need to develop statistical methodologies to analyse such data. It will also be apparent that few methods currently exist, and even for those that do exist the need remains to establish mathematical underpinning and rigor including statistical properties of the results of these procedures. Typically, point estimators have been developed, but there are still essentially no results governing their properties such as standard errors and distribution theory. These remain as outstanding problems.

2 Symbolic Data Sets

A data set may from its outset be structured as a symbolic data set. Alternatively, it may be structured as a classical data set but will become organized as symbolic data in order to establish it in a more manageable fashion, especially when initially it is very large in size. In this section, we present examples of both classical and symbolic data. Also, we introduce notation describing symbolic data sets for analysis. This process includes those situations, e.g., when two or more data sets are being merged, or when different features of the data are to be highlighted.

Suppose we have a data set consisting of the medical records of individuals in a country. Suppose for each individual, there will be a record of geographical location variables, such as

region (north, north-east, south, ...), city (Boston, Atlanta, ...), urban/rural (Yes, No), and so on. There will be demographic variables such as gender, marital status, age, information on parents (alive still, or not) siblings, number of children, employer, health provider, etc. Basic medical variables could include weight, pulse rate, blood pressure, etc. Other health variables (for which the list of possible variables is endless) would include incidences of certain ailments and diseases; likewise, for a given incidence or prognosis, treatments and other related variables associated with that disease are recorded. A typical such data set may follow the lines of Table 1.

Let p be the number of variables for each individual $i \in \Omega = \{1, \ldots, n\}$, where clearly p and n can be large, or even extremely large; and let Y_j, $j = 1, \ldots, p$, represent the jth variable. Let $Y_j = x_{ij}$ be the particular value assumed by the variable Y_j for the ith individual in the classical setting, and write $\boldsymbol{X} = (x_{ij})$ as the $n \times p$ matrix of the entire data set. Let the domain of Y_j be $\mathcal{Y}_j$; so $\boldsymbol{X} = (Y_1, \ldots, Y_p)$ takes values in $\mathcal{X} = \times_{j=1}^{p} \mathcal{Y}_j$. [Since the presence or absence of missing values is of no importance for the present, let us assume all values exist, even though this is most unlikely for large data sets.]

Variables can be quantitative, e.g., age with $\mathcal{Y}_{\text{age}} = \{x \geq 0\} = \mathcal{Y}_+$ as a continuous random variable; or with $\mathcal{Y}_{\text{age}} = \{0, 1, 2, \ldots\} = \mathcal{N}_0$, as a discrete random variable. Variables can be categorical, e.g., city with $\mathcal{Y}_{\text{city}} = \{\text{Atlanta, Boston, ...}\}$ or coded $\mathcal{Y}_{\text{city}} = \{1, 2, \ldots\}$, respectively. Disease variables can be recorded as categories (coded or not) of a single variable with domain $\mathcal{Y} = \{\text{heart, stroke, cancer, cirrhosis, ...}\}$, or, as is more likely, as an indicator variable, e.g., $Y = \text{cancer}$ with domain $\mathcal{Y} = \{\text{No, Yes}\}$ or $\{0, 1\}$ or with other coded levels indicating stages of disease. Likewise, for a recording of the many possible types of cancers, each type may be represented by a variable Y, or may be represented by a category of the cancer variable.

The precise nature of the description of the variables is not critical. What is crucial in the classical setting is that for each x_{ij} in $\boldsymbol{X}$, there is precisely one possible realized value. That is, e.g., an individual's $Y_{\text{age}} = 24$, say, or $Y_{\text{city}} = \text{Boston}$, $Y_{\text{cancer}} = \text{Yes}$, $Y_{\text{pulse}} = 64$, and so on. Thus, a classical data point is a single point in the p-dimensional space $\mathcal{X}$.

In contrast, a symbolic data point can be a hypercube in p-dimensional space or Cartesian product of distributions. Entries in a symbolic data set (denoted by ξ_{ij}) are not restricted to a single specific value. Thus, age could be recorded as being in an interval, e.g., $[0, 10), [10, 20), [20, 30), \ldots$. This could occur when the data point represents the age of a family or group of individuals whose ages collectively fall in an interval (such as [20, 30) years, say); or the data may correspond to a single individual whose precise age is unknown other than it is known to be within an interval range, or whose age has varied over time in the course of the experiment which generated the data; or combinations and variations thereof, producing interval-ranged data. In a different direction, it may not be possible to

measure some characteristic accurately as a single value, e.g., pulse rate at 64, but rather measures the variable as an $(x \pm \delta)$ value, e.g., pulse rate is (64 ± 1). A person's weight may fluctuate between (130, 135) over a weekly period. An individual may have ≤ 2, or > 2 siblings (or children, or ...). The blood pressure variable may be recorded by its [low, high] values, e.g, $\xi_{ij} = [78, 120]$. These variables are interval-valued symbolic variables.

A different type of variable would be a cancer variable which may have a domain $\mathcal{Y} =$ {lung, bone, breast, liver, lymphoma, prostate, ...} listing all possible cancers with a specific individual having the particular values $\xi_{ij} =$ {lung, liver}, for example. In another example, suppose the variable Y_j represents type of automobile owned (say) by a household, with domain $\{\mathcal{Y}_j =$ {Chevrolet, Ford, Toyota, Volvo, ...}. A particular household i may have the value $\xi_{ij} =$ {Toyota, Volvo}. Such variables are called multi-valued variables.

A third type of symbolic variable is a modal variable. Modal variables are multi-state variables with a frequency, probability, or weight attached to each of the specific values in the data. I.e., the modal variable Y is a mapping

$$Y(i) = \{U(i), \pi_i\} \quad \text{for} \quad i \in \Omega$$

where π_i is a nonnegative measure or a distribution on the domain $\mathcal{Y}$ of possible observation values and $U(i) \subseteq \mathcal{Y}$ is the support of π_i. For example, if three of an individual's siblings are diabetic and one isn't, then the variable describing propensity to diabetes could take the particular value $\xi_{ij} =$ {3/4 diabetes, 1/4 nondiabeties}. More generally, ξ_{ij} may be a histogram, an empirical distribution function, a probability distribution, a model, or so on. Indeed, Schweitzer (1984) opined that "distributions are the numbers of the future". Whilst in this example the weights (3/4, 1/4) might represent relative frequencies, other kinds of weights such as "capacities", "credibilities", "necessities", "possibilities", etc. may be used. Here, we define "capacity" in the sense of Choquet (1954) as the probability that at least one individual in the class has a certain Y value (e.g., is diabetic); and "credibility" is defined in the sense of Schafer (1976) as the probability every individual in the class has that characteristic (see, Diday, 1995).

In general then, unlike classical data for which each data point consists of a single (categorical or quantitative) value, symbolic data can contain internal variation and can be structured. It is the presence of this internal variation which necessitates the need for new techniques for analysis which in general will differ from those for classical data. Note however that classical data represent a special case; e.g., the classical point $x = a$ is equivalent to the symbolic interval $\xi = [a, a]$.

Notationally, we have a basic set of objects, which are elements or entities, $E = \{1, \ldots, N\}$ called the object set. This object set can represent a universe of individuals $E = \Omega$ (as above) in which case $N = n$; or if $N \leq n$, any one object set is a subset of Ω. Also, as frequently occurs in symbolic analyses, the objects u in E are classes $C_1, \ldots, C_m$

of individuals in Ω, with $E = \{C_1, \ldots, C_m\}$, and $N = m$. Thus, e.g., class C_1 may consist of all those individuals in Ω who have had cancer. Each object $u \in E$ is described by p symbolic variables Y_j, $j = 1, \ldots, p$, with domain $\mathcal{Y}_j$, and with $\mathcal{Y}_j$ being a mapping from the object set E to a range $\mathcal{Y}_j$ which depends on the type of variable Y_j is. Thus, if Y_j is a classical quantitative variable, the domain $\mathcal{B}_j$ is a subset of the real line $\Re$, i.e., $\mathcal{B}_j \subseteq \Re$; if Y_j is an interval variable, $\mathcal{B}_j = \{[\alpha, \beta],\ -\infty < \alpha, \beta < \infty\}$; if Y_j is categorical (nominal, ordinal, subsets of a finite domain Y_j), then $\mathcal{B}_j = \{B | B \subseteq \{(\text{list of cancers, e.g.})\}\}$; and if Y_j is a modal variable, $\mathcal{B}_j = M(\mathcal{Y}_j)$ where $M(\mathcal{Y})$ is family of all nonnegative measures on $\mathcal{Y}$.

Then, the symbolic data for the object set E are represented by the $N \times p$ matrix $\boldsymbol{X} = (\xi_{uj})$ where $\xi_{uj} = Y_j(u) \in \mathcal{B}_j$ is the observed symbolic value for the variable Y_j, $j = 1, \ldots, p$, for the object $u \in E$. The row x'_u of $\boldsymbol{X}$ is called the symbolic description of the object u. Thus, for the data in Table 2, the first row

$x'_1 = \{[20, 30], [79, 120]$, Boston, {Brain tumor}, {Male}, [170, 180]$\}$

represents a male in his 20's who has a brain tumor, a blood pressure of 120/79, weighs between 170 and 180 pounds and lives in Boston. The object u associated with this x'_u may be a specific male individual followed over a ten-year period whose weight has fluctuated between 170 and 180 pounds over that interval, or, u could be a collection of individuals whose ages range from 20 to 30 and who have the characteristics described by x'_u. The data x'_4 in Table 2 may represent the same individual as that represented by the $i = 4$th individual of Table 1 but where it is known only that she has either breast cancer (with probability p) or lung cancer (with probability $1-p$) but it is not known which. On the other hand, it could represent the set of 47 year old women from El Paso of whom a proportion p have either lung cancer and proportion $(1 - p)$ have breast cancer; or it could represent individuals who have both lung and breast cancer; and so on. (At some stage, whether the variable (Type of Cancer here) is categorical, a list, modal or whatever, would have to be explicitly defined.)

Another issue relates to dependent variables, which for symbolic data implies logical dependence, hierarchical dependence, taxonomic, or stochastic dependence. Logical dependence is as the word implies, as in the example, if [age $\leq$ 10], then [# children = 0]. Hierarchical dependence occurs when the outcome of one variable (e.g., Y_2 = treatment for cancer, say) with $Y_2 = \{$chemo, radiation, ...$\}$ depends on the actual outcome realized for another variables (e.g., Y_1 = Has cancer with $Y_1 = \{$No, Yes$\}$, say). If Y_1 has the value {Yes}, then $Y_2 = \{$chemotherapy, say$\}$; while if $Y_1 = \{$No$\}$ then clearly Y_2 is not applicable. [We assume for illustrative purposes here that the individual does not have chemotherapy treatment for some other reason.] In these cases, the non-applicable variable Z is defined with domain $Z = \{NA\}$. Such variables are also called mother $(Y_1)-$ daughter (Y_2) vari-

ables. Other variables may exhibit a taxonomic dependence; e.g., Y_1 = region and Y_2 = city can take values, if Y_1 = NorthEast then Y_2 = Boston, or if Y_1 = South then Y_2 = Atlanta, say.

3 Classes and Their Construction; Symbolic Objects

At the outset, our symbolic data set may already be sufficiently small in size that an appropriate symbolic statistical analysis can proceed directly. An example is the data of Table 11 used to illustrate a symbolic principal component analysis. More generally however and almost inevitably before any real (symbolic) data analysis can be conducted especially for large data sets, there will need to be implemented various degrees of data manipulation to organize the information into classes appropriate to specific questions at hand. In some instances, the objects in E (or Ω) are already aggregated into classes, though even here certain questions may require a reorganization into a different classification of classes regardless of whether the data set is small or large. For example, one set of classes $C_1, \ldots, C_m$ may represent individuals categorized according to m different types of primary diseases; while another analysis may necessitate a class structure by cities, gender, age, gender and age, or etc. Another earlier stage is when initially the data are separately recorded as for classical statistical and computer science databases for each individual $i \in \Omega = \{1, \ldots, n\}$, with n extremely large; likewise for very large symbolic databases. This stage of the symbolic data analysis then corresponds to the aggregation of these n objects into m classes where m is much smaller, and is designed so as to elicit more manageable formats prior to any statistical analysis. Note that this construction may, but need not, be distinct from classes that are obtained from a clustering procedure. Note also that the m aggregated classes may represent m patterns elicited from a data mining procedure.

This leads us to the concept of a symbolic object developed in a series of papers by Diday and his colleagues (e.g., Diday, 1987, 1989, 1990; Bock and Diday, 2000; and Stephan et al., 2000). We introduce this here first through some motivating examples; and then at the end of this section, a more rigorous definition is presented.

Some Examples

Suppose we are interested in the concept "Northeasterner". Thus, we have a description d representing the particular values {Boston, ..., other N-E cities, ...} in the domain $\mathcal{Y}_{\text{city}}$; and we have a relation R (here $\in$)linking the variable Y_{city} with the particular description of interest. We write this as $[Y_{\text{city}} \in \{\text{Boston, ..., other N-E cities, ...}\}] = a$, say. Then, each individual i in $\Omega = \{1, \ldots, n\}$ is either a Northeasterner or is not. That is, a is a mapping from $\Omega \to \{0, 1\}$, where for an individual i who lives in the Northeast, $a(i) = 1$; otherwise, $a(i) = 0$, $i \in \Omega$. Thus, if an individual i lives in Boston (i.e., $Y_{\text{city}}(i) = \text{Boston}$), then we

have $a(i)$ =[Boston $\subseteq$ {Boston, ..., other N-E cities, ...}] = 1.

The set of all $i \in \Omega$ for whom $a(i) = 1$, is called the extent of a in Ω. The triple $s = (a, R, d)$ is a symbolic object where R is a relation between the description $Y(i)$ of the (silent) variable Y and a description d and a is a mapping from Ω to L which depends on R and d. (In the Northeasterner example, $L = \{0, 1\}$). The description d can be an intentional description; e.g., as the name suggests, we intend to find the set of individuals in Ω who live in the "Northeast". Thus, the concept "Northeasterner" is somewhat akin to the classical concept of population; and the extent in Ω corresponds to the sample of individuals from the Northeast in the actual data set. Recall however that Ω may already be the "population" or it may be a "sample" in the classical statistical sense of sampling, as noted in Section 2.

Symbolic objects play a role in one of three major ways within the scope of symbolic data analyses. First, a symbolic object may represent a concept by its intent (e.g., its description and a way for calculating its extent) and can be used as the input of a symbolic data analysis. Thus, the concept "Northeasterner" can be represented by a symbolic object whose intent is defined by a characteristic description and a way to find its extent which is the set of people who live in the Northeast. A set of such regions and their associated symbolic objects can constitute the input of a symbolic data analysis. Secondly, it can be used as output from a symbolic data analysis as when a clustering analysis suggests Northeasterners belong to a particular cluster where the cluster itself can be considered as a concept and be represented by a symbolic object. The third situation is when we have a new individual (i') who has description d', and we want to know if this individual (i') matches the symbolic object whose description is d; that is, we compare d and d' by R to give $[d'Rd] \in L = \{0, 1\}$, where $[d'Rd] = 1$ means that there is a connection between d' and d. This "new" individual may be an "old" individual but with updated data; or it may be a new individual being added to the data base who may or may not "fit into" one of the classes of symbolic objects already present, (e.g., should this person be provided with specific insurance coverage?).

In the context of the aggregation of our data into a smaller number of classes, were we to aggregate the individuals in Ω by city, i.e., by the value of the variable Y_{city}, then the respective classes C_u, $u \in \{1, \ldots, m\}$ comprise those individuals in Ω which are in the extent of the corresponding mapping a_u, say. Subsequent statistical analysis can take either of two broad directions. Either, we analyse, separately for each class, the classical or symbolic data for the individuals in C_u as a sample of n_u observations as appropriate; or, we summarize the data for each class to give a new data set with one "observation" per class. In this latter case, the data set values will be symbolic data regardless of whether the original values were classical or symbolic data. For example, even though each individual

in Ω is recorded as having or not having had cancer (Y_{cancer} = No, Yes), i.e., as a classical data value, this variable when related to the class for city (say) will become, e.g., {Yes (.1), No (.9)}, i.e., 10% have had cancer and 90% have not. Thus, the variable Y_{cancer} is now a modal valued variable.

Likewise, a class that is constructed as the extent of a symbolic object, is typically described by a symbolic data set. For example, suppose our interest lies with "those who live in Boston", i.e., $a = [Y_{\text{city}} = \text{Boston}]$; and suppose the variable Y_{child} is the number of children each individual $i \in \Omega$ has with possible values $\{0, 1, 2, \geq 3\}$. Suppose the data value for each i is a classical value. (The adjustment for a symbolic data value such as individual i has 1 or 2 children, i.e., $\xi_i = \{1, 2\}$, readily follows). Then, the object representing all those who live in Boston will now have the symbolic variable Y_{child} with particular value

$$Y_{\text{child}} = \{(0, f_0), (1, f_1), (2, f_2), (\geq 3, f_3)\},$$

where f_i, $i = 0, 1, 2, \geq 3$, is the relative frequency of individuals in this class who have i children.

A special case of a symbolic object is an assertion. Assertions, also called queries, are particularly important when aggregating individuals into classes from an initial (relational) database. Let us denote by $z = (z_1, \ldots, z_p)$ the required description of interest of an individual or of a concept w. Here, z_j can be a classical single-valued entity x_j or a symbolic entity ξ_j. That is, while an x_j represents a realized classical data value and ξ_j represents a realized symbolic data value, z_j is a value being specifically sought or specified. Thus, for example, suppose we are interested in the symbolic object representing those who live in the Northeast. Then, z_{city} is the set of Northeastern cities. We formulate this as the assertion

$$a = [Y_{\text{city}} \in \{\text{Boston, ..., other N-E cities, ...}\}] \tag{1}$$

where a is mapping from Ω to $\{0, 1\}$ such that, for individual or object w, $a(w) = 1$ if $Y_{\text{city}}(w) \in \{\text{Boston, ..., other N-E cities, ...}\}$.

In general, an assertion takes the form

$$a = [Y_{j_1} R_{j_1} z_{j_1}] \wedge [Y_{j_2} R_{j_2} z_{j_2}] \wedge \ldots \wedge [Y_{j_v} R_{j_v} z_{j_v}] \tag{2}$$

for $1 \leq j_1, \ldots, j_v \leq p$, where '$\wedge$' indicates the logical multiplicative 'and', and R represents the specified relationship between the symbolic variable Y_j and description value z_j. For each individual $i \in \Omega, a(i) = 1$ (or 0) when the assertion is true (or not) for that individual. More precisely, an assertion is a conjunction of v events $[Y_k R_k z_k]$, $k = 1, \ldots, v$.

For example, the assertions

$$a = [Y_{\text{cancer}} = \text{Yes}], \quad a = [Y_{\text{age}} \geq 60] \tag{3}$$

represent all individuals with cancer, and all individuals aged 60 and over, respectively. The assertion

$$a = [Y_{\text{cancer}} = \text{Yes}] \wedge [Y_{\text{age}} \geq 60] \tag{4}$$

represents all cancer patients who are 60 or more years old; while the assertion

$$a = [Y_{\text{age}} < 20] \wedge [Y_{\text{age}} > 70] \tag{5}$$

seeks all individuals under 20 and over 70 years old. In each case, we are dealing with a concept "those aged over 60", "those over 60 with cancer", etc.; and a maps the particular individuals present in Ω onto the space $\{0, 1\}$.

If instead of recording the cancer variable as a categorical {Yes, No} variable, it were recorded as a {lung, liver, breast, ...} variable, the assertion

$$a = [Y_{\text{cancer}} \in \{\text{lung, liver}\}] \tag{6}$$

is describing the class of individuals who have either lung cancer or liver cancer or both. Likewise, an assertion can take the form

$$a = [Y_{\text{age}} \subseteq [20, 30]] \wedge [Y_{\text{city}} \in \{\text{Boston}\}]; \tag{7}$$

that is, this assertion describes those who live in Boston and are in the 20s age-wise.

The relations R can take any of the forms $=, \neq, \in, \leq, \subseteq$, etc., or can be a matching relationship (such as a comparison of probability distributions), or a structured sequence, and so on. They form the link between the symbolic variable Y and the specific description z of interest. The domain of the symbolic object can be written as,

$$D = D_1 \times \ldots \times D_p \subseteq \mathcal{X} = \times_{j=1}^{p} \mathcal{Y}_j.$$

where $D_j \subseteq \mathcal{Y}_j$. The p-tuple $(D_1, \ldots, D_p)$ of sets is called a description system, and each subset D is a description set consisting of description vectors $z = (z_1, \ldots, z_p)$; while a combination of elements $z_j \in \mathcal{Y}_j$ and sets $D_j \subseteq \mathcal{Y}_j$ as in (7) above for example is simply a description. When there are constraints on any of the variables (such as when logical dependencies exist), then the space D has a "hole" in it corresponding to those values which match the constraints. The totality of all descriptions D is the description space $\mathcal{D}$. Hence, the assertion can be written as

$$a = [Y \in D] \equiv \bigwedge_{k=1}^{v} [Y_{j_k} R_{j_k} z_{j_k}] = [YRz]$$

where $R = \bigwedge_{k=1}^{v} R_{j_k}$ is called the product relation.

Note that implicitly, if an assertion does not involve a particular variable Y_w, say, then the domain relating to that variable remains unchanged at $\mathcal{Y}_w$. For example, if $p = 3$, and

Y_1 = city, Y_2 = age and Y_3 = weight, then the assertion $a = [Y_1 \in \{\text{Boston, Atlanta}\}]$ has domain $D = \{\text{Boston, Atlanta}\} \times \mathcal{Y}_2 \times \mathcal{Y}_3$ and seeks all those in Boston and Atlanta only (but regardless of age and weight).

Formal Definitions

Let us now formally define the following concepts.

Definition: When an assertion is asked of any particular object $i \in \Omega$, it assumes a value true ($a = 1$) if that assertion holds for that object, or false ($a = 0$) if not. We write

$$a(i) = [Y(i) \in D] = \begin{cases} 1, & Y(i) \in D, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The function $a(i)$ is called a **truth function** and represents a mapping of Ω onto $\{0, 1\}$. The set of all $i \in \Omega$ for which the assertion holds is called the **extension** in Ω, denoted by $Ext(a)$ or Q,

$$Ext(a) = Q = \{i \in \Omega | Y(i) \in D\} = \{i \in \Omega | a(i) = 1\}. \tag{9}$$

Formally, the mapping $a : \Omega \to \{0, 1\}$ is called an **extension mapping**.

Typically, a class would be identified by the symbolic object that described it. For example, the assertion (7) corresponds to the symbolic object "20-30 year olds living in Boston". The extension of this assertion, $Ext(a)$, produces the class consisting of all individuals $i \in \Omega$ which match this description a, i.e., those i for whom $a(i) = 1$.

A class may be constructed, as in the foregoing, by seeking the extension of an assertion a. Alternatively, it may be that it is desired to find all those (other) individuals in Ω who match the description $Y_j(u)$ of a given individual $u \in \Omega$. Thus in this case, the assertion is

$$a(i) = \bigwedge_{j=1}^{p} [Y_j = Y_j(u)]$$

where now $z_j = Y_j(u)$, and has extension

$$Ext(a_i) = \{i \in \Omega | Y_j(i) = Y_j(u), \quad j = 1, \ldots, p\}.$$

More generally, an assertion may hold with some intermediate value (such as a probability), i.e., $0 \leq a(i) \leq 1$, representing a degree of matching of an object i with an assertion a. Flexible matching is also possible. In these cases, the mapping a is onto the interval [0, 1], i.e., $a : \Omega \to [0, 1]$. Then, the extension of a has level α, $0 \leq \alpha \leq 1$, where now

$$Ext_\alpha(a) = Q_\alpha = \{i \in \Omega | a(i) \geq \alpha\}.$$

Diday and Emilion (1996) and Diday et al. (1996) study modal symbolic objects and also provide some of its theoretical underpinning; see also Section 10.

We now have the formal definition of a symbolic object as follows.

Definition: A **symbolic object** is the triple $s = (a, R, d)$ where a is a mapping $a : \Omega \rightarrow L$ which maps individuals $i \in \Omega$ onto the space L depending on the relation R between descriptions and the description d. When $L = \{0, 1\}$, i.e., when a is a binary mapping, then s is a **Boolean symbolic object**. When $L = [0, 1]$, then s is a **modal symbolic object**.

That is, a symbolic object is a mathematical model of a concept (see, Diday, 1995). If $[a, R, d] \in \{0, 1\}$, then s is a Boolean symbolic object and if $[a, R, d] \in [0, 1]$, then s is a modal symbolic object. For example, in (7) above, we have the relations $R = (\subseteq, \in)$ and the description $d = ([20, 30], \{\text{Boston}\})$. The intent is to find "20-30 year olds who live in Boston". The extent consists of all individuals in Ω who match this description, i.e., those i for whom $a(i) = 1$.

Whilst recognition of the need to develop methods for analyzing symbolic data and tools to describe symbol objects is relatively new, the idea of considering higher level units as concepts is in fact ancient. Aristotle Organon in the 4th century BC (Aristotle, IVBC, 1994) clearly distinguishes first order individuals (such as **the** horse or **the** man) which represented units in the world (the statistical population) from second order individuals (such as **a** horse or **a** man) represented as units in a class of individuals. Later, Arnault and Nicole (1662) defined a concept by notions of an intent and an extent (whose meanings in 1662 match those herein) as: "*Now, in these universal ideas there are two things which is important to keep quite distinct: comprehension and extension (for "intent" and "extent"). I call the comprehension of an idea the attributes which it contains and which cannot be taken away from it without destroying it; thus the comprehension of the idea of a triangle includes, to a superficial extent, figure, three lines, three angles, the equality of these three angles to two right angles etc. I call the extension of an idea the subjects to which it applies, which are also called the inferiors of a universal term, that being called superior to them. Thus the idea of triangle in general extends to all different kinds of triangle*".

Finally, computational implementation of the generation of appropriate classes can be executed by queries (assertions) used in search engines such as the exhaustive search, genetic, or hill climbing algorithms, and/or by the use of available software such as the standard query language (SQL) package or other "object oriented" languages such as C^{++} or JAVA. Stephan et al. (2000) provide some detailed examples illustrating the use of SQL. Another package specifically written for symbolic data is the SODAS (Symbolic Official Data Analysis System) software. Gettler-Summa (1999, 2000) developed the MGS (marking and generalization by symbolic descriptions algorithm) for building symbolic descriptions starting with classical nominal data. Thus, the outputs (ready for symbolic data analysis) are symbolic objects which have modal or multi-valued variables and which identify logical

links between the variables. Csernel (1997), inspired by Codd's (1972) methods of dealing with relational databases with functional dependencies between variables, developed a normalization of Boolean symbolic objects taking into account constraints on the variables.

While we have not addressed it explicitly, the concepts described in this section can also be applied to merged or linked data sets. For example, suppose we also had a data set consisting of meteorological and environmental variables (measured as classical or symbolic data) for cities in the U.S. Then, if the data are merged by city, it is easy to obtain relevant meteorological and environmental data for each individual identified in Table 1. Thus, for example, questions relating to environmental measures and cancer incidences could be considered.

4 Symbolic Data Analyses

Given a symbolic data set, the next step is to conduct statistical analyses as appropriate. As for classical data, the possibilities are endless. Unlike classical data for which a century of effort has produced a considerable library of analytical/statistical methodologies, symbolic statistical analyses are new and available methodologies are still only few in number. In the following sections, we shall review some of these as they pertain to symbolic data.

Therefore, in Sections 5 and 6, we consider descriptive statistics. Then, we review principal component methods for interval-valued data in Section 7. Clustering methods are treated in Section 8. Attention will be restricted to the more fully developed criterion-based divisive clustering approach for categorical and interval-valued variables. A summary of other methods limited largely to specific special cases will also be provided. Much of the current progress in symbolic data analyses revolve around cluster-related questions. Theoretical underpining where it exists is covered in Section 9.

Except for the specific distance measures developed for symbolic clustering analysis (see Section 8), we shall not attempt herein to review the literature on symbolic similarity and dissimilarity measures. Classical measures include the Minkowski or L_q distance (with $q = 2$ being the Euclidean distance measure), Hamming distance, Mahalanobis distance, and Gower-Legendre family, plus the large class of dissimilarity measures for probability distributions from classical probability and statistical theory, e.g., the familiar chi-square measure, Bhattacharyya distance and Chernoff's distance, to name but a very few. Symbolic analogues have been proposed for specific cases. Gowda and Diday (1991) first introduced a basic dissimilarity measure for Boolean symbolic objects. Later, Ichino and Yaguchi (1994), using Cartesian operators, developed a symbolic generalized Minkowski distance of order $q \geq 1$. This was extended by De Carvalho (1994, 1998) to include Boolean symbolic objects constrained by logical dependencies between variables. Matching of Boolean symbolic objects is considered by Esposito et al. (1991). These measures invoke the concepts of

Cartesian join, and meet. Some are developed along the lines of distance functions (as described in Section 8; see, e.g., Chavent, 1998). Many are limited to univariate cases. Since such measures vary widely and the choices of what to use are generally dependent on the task at hand, we shall not expand on their details further. In a different direction, Morineau et al. (1994), Loustaunau et. al. (1997) and Gettler-Summa and Pardoux (2000) consider three-way tables. A more detailed and exhaustive description of symbolic methodologies in general up to its publication can be found in Bock and Diday (2000). Billard and Diday (2002b) also has an expanded coverage providing some additional illustrations.

5 Descriptive Univariate Statistics

5.1 Some preliminaries

Basic descriptive statistics include frequency histograms, and sample means and variances. We consider symbolic data analogues of these statistics, for multi-valued and interval-valued variables with rules and modal variables. As we develop these statistics, let us bear in mind the following example which illustrates the need to distinguish between the levels (e.g., structure) in symbolic data when constructing (e.g.) histograms compared with that for a standard histogram.

Suppose an isolated island contains one thousand penguins and one thousand ostriches both non-flying species of birds, and four thousand pigeons which is a flying bird species. Suppose we are interested in the variable "flying" which here takes two possible values "Yes" or "No". Then, a standard histogram based on birds is such that the frequency for "Yes" is two times higher than that for "No"; see Figure 1(a). In contrast, if we consider the individuals to be the species, then since there are two non-flying species and one flying species, the frequency for "No" is now two times higher than it is for "Yes"; see Figure 1(b). Notice that a histogram of species is just a histogram. What this example shows is that depending on the level of individuals (here birds) and on the level of concept (here species) we obtain completely different histograms. Species here actually has within it another level of information corresponding to the type of bird and the frequency of each. That is, "species" itself contains structure at another level from the variable species as used to produce the histogram of Figure 1(b). A symbolic histogram (developed later in this section) takes this structure into account.

We consider univariate statistics (for the $p \geq 1$ variate data) in this section and bivariate statistics in Section 6. For integer-valued and interval-valued variables, we follow the approach adopted by Bertrand and Goupil (2000). DeCarvalho (1994, 1995) and Chouakria et al. (1998) have used different but equivalent methods to find the histogram and interval probabilities (see equation (26) below) for interval-valued variables.

Before describing these quantities, we first need to introduce the concept of virtual extensions. Recall from Section 3 that the symbolic description of an object $u \in E$ (or equivalently, $i \in \Omega$) was given by the description vector $d_u = (\xi_{u1}, \ldots, \xi_{up})$, $u = 1, \ldots, m$, or more generally, $d \in (D_1, \ldots, D_p)$ in the space $D = \times_{j=1}^{p} D_j$, where in any particular case the realization of Y_j may be an x_j as for classical data or an ξ_j of symbolic data. *Individual descriptions*, denoted by x, are those for which each D_j is a set of one value only, i.e., $x \equiv d = (\{x_1\}, \ldots, \{x_p\})$, $x \in \mathcal{X} = \times_{j=1}^{p} \mathcal{Y}_j$.

The calculation of the symbolic frequency histogram involves a count of the number of individual descriptions that match certain implicit logical dependencies in the data. A logical dependency can be represented by a rule v,

$$v : [x \in A] \Rightarrow [x \in B] \tag{10}$$

for $A \subseteq D$, $B \subseteq D$, and $x \in \mathcal{X}$ and where v is a mapping of $\mathcal{X}$ onto $\{0, 1\}$ with $v(x) = 0(1)$ if the rule is not (is) satisfied by x. For example, suppose $x = (x_1, x_2) = (10, 0)$ is an individual description of $Y_1 =$ age and $Y_2 =$ number of children for $i \in \Omega$, and suppose $A = \{\text{age} \leq 12\}$ and $B = \{0\}$. Then, the rule that an individual whose age is less than 12 implies they have had no children is logically true, whereas an individual whose age is under 12 but has had 2 children is not logically true. It follows that an individual description vector x satisfies the rule v if and only if $x \in A \cap B$ or $x \notin A$. This formulation of the logical dependency rule is sufficient for the purposes of establishing basic descriptive statistics. Verde and DeCarvalho (1998) discuss a variety of related rules (such as logical equivalence, logical implication, multiple dependencies, hierarchical dependencies, and so on). We have the following formal definition.

Definition: The **virtual description** of the description vector d as the set of all individual description vectors x that satisfy all the (logical dependency) rules v in $\mathcal{X}$. We write this as, for V_x the set of all rules v operating on x,

$$vir(d) = \{x \in D; v(x) = 1, \quad \text{for all } v \text{ in } V_x\}. \tag{11}$$

5.2 Multi-valued Variables - Univariate Statistics

Suppose we want to find the frequency distribution for the particular multivalued symbolic variable $Y_j \equiv Z$ which takes possible particular values $\xi \in \mathcal{Z}$. These can be categorical values (e.g., types of cancer), or any form of discrete random variable. We define the observed frequency of ξ as

$$O_Z(\xi) = \sum_{u \in E} \pi_Z(\xi; u) \tag{12}$$

where the summation is over $u \in E = \{1, \ldots, m\}$ and where

$$\pi_Z(\xi; u) = \frac{|\{x \in vir(d_u) | x_Z = \xi\}|}{|vir(d_u)|} \tag{13}$$

is the percentage of the individual description vectors in $vir(d_u)$ such that $x_Z = \xi$, and where $|A|$ is the number of individual descriptions in the space A. In the summation in (12), any u for which $vir(d_u)$ is empty is ignored. We note that this observed frequency is a positive real number and not necessarily an integer as for classical data. In the classical case, $|vir(d_u)| = 1$ and so is a special case of (13). We can easily show that

$$\sum_{\xi \in Z} O_Z(\xi) = m' \tag{14}$$

where $m' = (m - m_0)$ with m_0 being the number of u for which $|vir(d_u)| = 0$.

For a multi-valued symbolic variable Z, taking values $\xi \in \mathcal{Z}$, the **empirical frequency distribution** is the set of pairs $[\xi, O_Z(\xi)]$ for $\xi \in \mathcal{Z}$, and the **relative frequency distribution** or **frequency histogram** is the set

$$[\xi, (m')^{-1} O_Z(\xi)]. \tag{15}$$

The following definitions follow readily.

The **empirical distribution function** of Z is given by

$$F_Z(\xi) = \frac{1}{m'} \sum_{\xi_k \leq \xi} O_Z(\xi_k). \tag{16}$$

When the possible values ξ for the multi-valued symbolic variable $Y_j = Z$ are quantitative, we can define a symbolic mean, variance, and median, as follows.
The **symbolic sample** mean is

$$\bar{Z} = \frac{1}{m'} \sum_{\xi_k} \xi_k O_Z(\xi_k); \tag{17}$$

the **symbolic sample variance** is

$$S_Z^2 = \frac{1}{m'} \sum_{\xi_k} O_Z(\xi_k)[\xi_k - \bar{Z}]^2; \tag{18}$$

and, the **symbolic median** is the value ξ for which

$$F_Z(\xi) \geq 1/2, \qquad F_Z(\xi^- \leq 1/2). \tag{19}$$

An Example

To illustrate, suppose a large data set (consisting of patients served through a particular HMO in Boston, say) was aggregated in such a way that it produced the data of Table 3 representing the outcomes relative to the presence of cancer Y_1 with $\mathcal{Y}_1 = \{\text{No}= 0, \text{Yes} = 1\}$ and Number of cancer related treatments Y_2 with $\mathcal{Y}_2 = \{0, 1, 2, 3\}$ on $m = 9$ objects. Thus, e.g., $d_1 = (\{0, 1\}, \{2\})$ is the description vector for the object in row 1. Hence,

for the individuals represented by this description, the observation $Y_1 = \{0, 1\}$ tells us that either some individuals have cancer and some do not or we do not know the precise Yes/No diagnosis for the individuals classified here, while the observation $Y_2 = \{2\}$ tells us all individuals represented by d_1 have had two cancer related treatments. In contrast, $d_7 = (\{1\}, \{2, 3\})$ represents individuals all of whom have had a cancer diagnosis ($Y_1 = 1$) and who have had either 2 or 3 treatments, $Y_2 = \{2, 3\}$. Suppose further there is a logical dependency

$$v : y_1 \in \{0\} \Rightarrow y_2 = \{0\}, \tag{20}$$

i.e., if no cancer has been diagnosed, then there must have been no cancer treatments. Notice that $y_1 \in \{0\} \Rightarrow A = \{(0, 0), (0, 1), (0, 2)(0, 3)\}$ and $y_2 \in \{0\} \Rightarrow B = \{(0, 0), (1, 0), (2, 0), (3, 0)\}$. From (20), it follows that an individual description x which satisfies this rule is $x \in A \cap B = \{(0, 0)\}$ or $x \notin A$, i.e., $x \in \{(1, 0), (1, 1), (1, 2), (1, 3)\}$. Let all possible cases that satisfy the rule be represented by $C = \{(0, 0), (1, 0), \ldots, (1, 3)\}$. We apply (20) to each d_u, $u = 1, \ldots, m$, in the data to find the virtual extensions $vir(d_u)$. Thus, for the first description d_1, we have

$$vir(d_1) = \{x \in \{0, 1\} \times \{2\} : v(x) = 1\}.$$

The individual descriptions $x \in \{0, 1\} \times \{2\}$ are (0,2) and (1,2), of which only one, $x = (1, 2)$, is also in the space C. Therefore, $vir(d_1) = \{(1, 2)\}$.

Clearly then, this operation is mathematically cleaning the data (so to speak) by identifying only those values which make logical sense (by satisfying the logical dependency rule of equation (20)). Thus, the data values $(Y_1, Y_2) = (0, 2)$ which record that both no cancer was present and there were two cancer related treatments are identified as erroneous data values (under the prevailing circumstances as specified by the rule v) and so are not used in this analyis to calculate the mean. [While not attempting to do so here, this identification does not preclude inclusion of other procedures which might subsequently be engaged to imput what these values might have been.] For small data sets, it may be possible to "correct" the data visually (or some such variation thereof). For very large data sets, this is not always possible; hence a logical dependency rule to do so mathematically/computationally is essential.

Similarly, for the second description d_2, we can obtain

$$vir(d_2) = \{x \in \{0, 1\} \times \{0, 1\} : v(x) = 1\}.$$

Here, the individual description vectors $x \in \{0, 1\} \times \{0, 1\}$ are (0,0), (0,1), (1,0) and (1,1) of which $x = (0, 0)$, $x = (1, 0)$ and $x = (1, 1)$ are in C. Hence, $vir(d_2) = \{(0, 0), (1, 0), (1, 1)\}$. The virtual extensions for all d_u, $u = 1, \ldots, 9$, are shown in Table 3. Notice that $vir(d_5) = \phi$ is the null set, since the data value d_5 cannot be logically true in the presence of the rule v.

We can now find the frequency distribution. Suppose we first find this distribution for Y_1. By definition (12), we have observed frequencies

$$\begin{aligned} O_{Y_1}(0) &= \sum_{u \in E'} \frac{|\{x \in vir(d_u) | x_{Y_1} = 0\}|}{|vir(d_u)|} \\ &= \frac{0}{1} + \frac{1}{3} + \frac{0}{1} + \frac{0}{2} + \frac{1}{1} + \frac{0}{2} + \frac{0}{2} + \frac{0}{2} = 4/3, \end{aligned}$$

and, likewise, $O_{Y_1}(1) = 20/3$, where $E' = E - (u = 5)$ and so $|E'| = 8 = m'$. Therefore, the relative frequency distribution for Y_1 is, from equation (15),

$$\text{Rel freq } (Y_1) : [(0, O_{Y_1}(0)/m'), \quad (1, O_{Y_1}(1)/m')] = [(0, 1/6), (1, 5/6)].$$

Similarly, we have observed frequencies for the possible Y_2 values $\xi = 0, 1, 2, 3$, respectively, as

$$O_{Y_2}(0) = \sum_{u \in E'} \frac{|\{x \in vir(d_u) | x_{Y_2} = 0\}|}{|vir(d_u)|}$$

$$= \frac{0}{1} + \frac{2}{3} + \frac{0}{1} + \frac{0}{2} + \frac{1}{1} + \frac{0}{2} + \frac{0}{2} + \frac{0}{2} = 5/3,$$

$$O_{Y_2}(1) = 4/3, \qquad O_{Y_2}(2) = 2.5, \qquad O_{Y_2}(3) = 2.5;$$

and hence, the relative frequency for Y_2 is, from equation (15),

$$\text{Rel freq } (Y_2) : [(0, 5/24), (1, 1/6), (2, 5/16), (3, 5/16)].$$

The empirical distribution function for Y_2 is, from equation (16),

$$F_{Y_2}(\xi) = \begin{cases} 5/24, & \xi < 1, \\ 3/8, & 1 \le \xi < 2, \\ 11/16, & 2 \le \xi < 3, \\ 1, & \xi \ge 3. \end{cases}$$

From equation (17), we have that the symbolic sample mean of Y_1 and Y_2, respectively, is $\bar{Y}_1 = 5/6 = 0.833$ and $\bar{Y}_2 = 83/48 = 1.729$; from equation (18), the symbolic sample variance of Y_1 and Y_2, respectively, is $S_1^2 = 0.1239$ and $S_2^2 = 1.1975$ respectively; and the median of Y_2 is 2.

Finally, we observe that we can calculate weighted frequencies, weighted means and weighted variances by replacing (12) by

$$O_Z(\xi) = \sum_{u \in E} w_u \pi_Z(\xi, u) \tag{21}$$

with $w_u \ge 0$ and $\Sigma w_u = 1$. For example, if the objects $u \in E$ are classes C_u comprised of individuals from the set of individuals $\Omega = \{1, \ldots, n\}$, i.e., $C_u \subseteq \Omega$, a possible weight is

$w_u = |C_u|/|\Omega|$ corresponding to the relative sizes of the class $C_u, u = 1, \ldots, m$. Or, more generally, if we consider each individual description vector x as an elementary unit of the object u, we can use weights, for $u \in E$,

$$w_u = \frac{|vir(d_u)|}{\sum_{u \in E} |vir(d_u)|}. \tag{22}$$

For example, if the u of Table 3 represent classes C_u of size $|C_u|$, with $|\Omega| = 1000$, as shown in Table 3, then, if we use the weights $w_u = |C_u|/|\Omega|$, we can show that

$$O_{Y_1}(0) = \frac{1}{1000}[128 \times \frac{0}{1} + 75 \times \frac{1}{3} + 249 \times \frac{0}{1} + \ldots + 121 \times \frac{0}{2}] = 0.229;$$

likewise,

$$O_{Y_1}(1) = 0.771.$$

Hence, the relative frequency of Y_1 is

$$\text{Rel freq } (Y_1) : [(0, 0.229), (1, 0.771)].$$

Likewise,

$$O_{Y_2}(0) = 0.2540, \quad O_{Y_2}(1) = 0.0970, \quad O_{Y_2}(2) = 0.2395, \quad O_{Y_2}(3) = 0.4095;$$

and hence the relative frequency of Y_2 is

$$\text{Rel freq } (Y_2) : [(0, 0.2540), (1, 0.0970), (2, 0.2395), (3, 0.4095)].$$

The empirical weighted distribution function for Y_2 becomes

$$F_{Y_2}(\xi) = \begin{cases} 0.2540, & \xi < 1, \\ 0.3510, & 1 \leq \xi < 2, \\ 0.5905, & 2 \leq \xi < 3, \\ 1, & \xi \geq 3. \end{cases}$$

Similarly, we can show that the symbolic weighted sample mean of Y_1 is $\bar{Y}_1 = 0.7710$ and of Y_2 is $\bar{Y}_2 = 1.8045$, and the symbolic weighted sample variance of Y_1 is $\xi_1^2 = 0.176$ and of Y_2 is $S_2^2 = 1.4843$.

5.3. Interval-valued Variables - Univariate Statistics

The corresponding description statistics for interval-valued variables are obtained analogously to those for multi-valued variables; see Bertrand and Goupil (2000). Let us suppose we are interested in the particular variable $Y_j \equiv Z$, and suppose the observation value for

object u is the interval $Z(u) = [a_u, b_u]$, for $u \in E = \{1, \ldots, m\}$. The individual description vectors $x \in vir(d_u)$ are assumed to be uniformly distributed over the interval $Z(u)$. Therefore, it follows that, for each ξ,

$$P\{x \leq \xi | x \in vir(d_u)\} = \begin{cases} 0, & \xi < a_u, \\ \frac{\xi - a_u}{b_u - a_u}, & a_u \leq \xi < b_u, \\ 1, & \xi \geq b_u. \end{cases} \tag{23}$$

The individual description vector x takes values globally in $\bigcup_{u \in E} vir(d_u)$. Further, it is assumed that each object is equally likely to be observed with probability $1/m$. Therefore, the empirical distribution function, $F_Z(\xi)$, is the distribution function of a mixture of m uniform distributions $\{Z(u), u = 1, \ldots, m\}$. Therefore, from (23),

$$\begin{aligned} F_Z(\xi) &= \frac{1}{m} \sum_{u \in E} P\{x \leq \xi | x \in vir(d_u)\} \\ &= \frac{1}{m} \left\{ \sum_{\xi \in Z(u)} \left(\frac{\xi - a_u}{b_u - a_u} \right) + |(u | \xi \geq b_u)| \right\}. \end{aligned}$$

Hence, by taking the derivative with respect to ξ, we obtain the **empirical density function** of Z as

$$f(\xi) = \frac{1}{m} \sum_{u : \xi \in Z(u)} \left(\frac{1}{b_u - a_u} \right). \tag{24}$$

Notice that the summation in (24) is only over those objects u for which $\xi \in Z(u)$. We may write (24) in the alternative form

$$f(\xi) = \frac{1}{m} \sum_{u \in E} \frac{I_u(\xi)}{||Z(u)||}, \quad \xi \in \Re, \tag{25}$$

where $I_u(.)$ is the indicator function that ξ is or is not in the interval $Z(u)$ and where $||Z(u)||$ is the length of that interval. Note that the summation in (25) is only over those objects u for which $\xi \in Z(u)$. The analogy with (15) and (16) is apparent.

To construct a histogram, let $I = [\min_{u \in E} a_u, \max_{u \in E} b_u]$ be the interval which spans all the observed values of Z in $\mathcal{X}$, and suppose we partition I into r subintervals $I_g = [\xi_{g-1}, \xi_g)$, $g = 1, \ldots, r-1$, and $I_r = [\xi_{r-1}, \xi_r]$. Then, the **histogram** for Z is the graphical representation of the frequency distribution $\{(I_g, p_g), g = 1, \ldots, r\}$ where

$$p_g = \frac{1}{m} \sum_{u \in E} \frac{||Z(u) \cap I_g||}{||Z(u)||}, \tag{26}$$

i.e., p_g is the probability an arbitrary individual description vector x lies in the interval I_g. If we want to plot the histogram with height f_g on the interval i_g, so that the "area" is p_g, then

$$p_g = (\xi_g - \xi_{g-1}) \times f_g. \tag{27}$$

Bertrand and Goupil (2000) indicate that, using the law of large numbers, the true limiting distribution of Z as $m \to \infty$ is only approximated by the exact distribution $f(\xi)$ in (25) since this depends on the veracity of the uniform distribution within each interval assumption. Mathematical underpining for the histogram has been developed in Diday (1995) using the strong law of large numbers and the concepts of t-norms and t-conorms as developed by Schweizer and Sklar (1983).

The **symbolic sample mean**, for an interval-valued variable Z, is given by

$$\bar{Z} = \frac{1}{2m} \sum_{u \in E} (b_u + a_u), \tag{28}$$

To verify (28), we recall the empirical mean $\bar{Z}$ in terms of the empirical density function is

$$\bar{Z} = \int_{-\infty}^{\infty} \xi f(\xi) d\xi.$$

Substituting from (25), we have

$$\begin{aligned} \bar{Z} &= \frac{1}{m} \sum_{u \in E} \int_{-\infty}^{\infty} \frac{I_u(\xi)}{||Z(u)||} \xi d\xi \\ &= \frac{1}{m} \sum_{u \in E} \frac{1}{b_u - a_u} \int_{\xi \in Z(u)} \xi d\xi \\ &= \frac{1}{2m} \sum_{u \in E} \frac{b_u^2 - a_u^2}{b_u - a_u} \\ &= \frac{1}{m} \sum_{u \in E} (b_u + a_u)/2, \end{aligned}$$

as required.

Similarly, we can derive the **symbolic sample variance** given by

$$S^2 = \frac{1}{3m} \sum_{u \in E} (b_u^2 + b_u a_u + a_u^2) - \frac{1}{4m^2} [\sum_{u \in E} (b_u + a_u)]^2. \tag{29}$$

As for multi-valued variables, if an object u has some internal inconsistency relative to a logical rule, i.e., if u is such that $|vir(d_u)| = 0$, then the summation in (28) and (29) is over only those u for which $|vir(d_u)| \neq 0$, i.e., over $u \in E'$, and m is replaced by m' (equal to the number of objects u in E'). In the sequel, it will be understood that m and E refer to those u for which these rules hold.

An Example

To illustrate, consider the data from Raju (1997) shown in Table 4, in which the pulse rate (Y_1), systolic blood pressure (Y_2) and diastolic blood pressure (Y_3) are recorded as an

interval for each of the $u = 1, \ldots, 10$ patients. Let us take the data for pulse rate (Y_1). The complete data set spans an interval $I = [44, 112]$ where

$$\min_{u \in E} a_u = 44, \quad \max_{u \in E} b_u = 112.$$

Suppose we want to construct a histogram on the $r = 8$ intervals $I_1 = [40, 50), \ldots, I_8 = [110, 120]$. Using equation (26), we can calculate the probability p_g that an arbitrary individual description vector x lies in the interval I_g, $g = 1, \ldots, 8$. For example, when $g = 4$, the probability that an x lies in the interval $I_4 = [70, 80)$ is

$$p_4 = \frac{1}{10}\{0 + \frac{2}{12} + \frac{10}{34} + \frac{10}{42} + \frac{2}{18} + \frac{10}{30} + \frac{8}{28} + \frac{4}{22} + 0 + 0\} = .1611.$$

Hence, from equation (27), we can calculate the height f_g of the plotted histogram for that interval as

$$f_g = p_g/(\xi_g - \xi_{g-1})$$

i.e.,

$$f_4 = (.1611)/10 = .01611.$$

A summary of the calculated values of p_g for each I_g is given in Table 5, and the plot of the histogram is shown in Figure 2. Table 5 also provides the intervals and probabilities for the interval valued variable Y_2 which represents the systolic blood pressure and for Y_3 which represents the diastolic blood pressure for the same ten patients.

Using the equations (28) and (29), we find the symbolic sample mean and variance respectively. Thus, the mean pulse rate is $\bar{Y}_1 = 79.1$ with variance $S_1^2 = 215.86$; the mean systolic blood pressure is $\bar{Y}_2 = 131.3$ with variance $S_2^2 = 624.18$; and the mean diastolic blood pressure is $\bar{Y}_3 = 84.6$ with variance $S_3^2 = 229.64$.

5.4 Multi-valued Modal Variables

There are many types of modal-valued variables. We consider very briefly two types only, one each of multi-valued and interval valued variables in this subsection and the next (*5.5*), respectively.

Let us suppose we have data from a categorical variable Y_j taking possible values ξ_{jk}, with relative frequencies p_k, $k = 1, \ldots, s$, respectively, with $\Sigma p_k = 1$. Suppose we are interested in the particular symbolic random variable $Y_j \equiv Z$, in the presence of the dependency rule v. Then, we define the observed frequency that $Z = \xi_k$, $k = 1, \ldots, s$, as

$$O_Z(\xi_k) = \sum_{u \in E} \pi_Z(\xi_k; u) \tag{30}$$

where the summation is over all $u \in E$ and where

$$\pi_Z(\xi_k; u) \quad = \quad P(Z = \xi_k | v(x) = 1, u)$$

$$= \frac{\sum_x P(x = \xi_k | v(x) = 1, u)}{\sum_x \sum_{k=1}^{s} P(x = \xi_k | v(x) = 1, u)} \tag{31}$$

where, for each u, $P(x = \xi_k | v(x) = 1, u)$ is the probability that a particular description x ($\equiv x_j$) has the value ξ_k and that the logical dependency rule v holds. If for a specific object u, there are no description vectors satisfying the rule v, i.e., if the denominator of (31) is zero, then that u is omitted in the summation in (30). We note that

$$\sum_{k=1}^{s} O_Z(\xi_k) = m. \tag{32}$$

An example

Suppose households are recorded as having one of the possible central heating fuel types $\mathcal{Y}_1$ taking values in $\mathcal{Y}_1 = \{$gas, solid fuel, electricity, other$\}$ and suppose Y_2 is an indicator variable taking values in $\mathcal{Y}_1 = \{$No, Yes$\}$ depending on whether a household does not (does) have central heating installed. For illustrative convenience, it is assumed the values for $\mathcal{Y}_1$ are conditional on there being central heating present in the household. Suppose that aggregation by geographical region produced the data of Table 6. The original (classical) data set consisted of census data from the UK Office for National Statistics on 34 demographic-socio-economic variables on individual households in 374 parts of the country. The data shown in Table 6 are those obtained for two specific variables after aggregation into 25 regions; and were obtained by applying the Symbolic Official Data Analysis System software to the original classical data set. Thus, region one represented by the object $u = 1$ is such that 87% of the households with central heating are fueled by gas, 7% by solids, 5% by electricity and 1% by some other type of fuel; and that 9% of households did not have central heating while 91% did. Let us suppose interest centers only on the variable Y_1 (without regard to any other variable) and that there is no rule v to be satisfied. Then, it is readily seen that the observed frequency that $Z = \xi_1 =$ gas is

$$O_{Y_1}(\xi_1) = (0.87 + \ldots + 0.43 + 0.00) = 18.05,$$

and likewise, for $\xi_2 =$ solid, $\xi_3 =$ electricity, and $\xi_4 =$ other fuels, we have, respectively,

$$O_{Y_1}(\xi_2) = 1.67, \quad O_{Y_1}(\xi_3) = 3.51, \quad O_{Y_1}(\xi_4) = 1.77.$$

Hence, the relative frequencies are

$$\text{Rel freq } (Y_1) : [(\xi_1, 0.722), (\xi_2, 0.067), (\xi_3, 0.140), (\xi_4, 0.071)].$$

Similarly, we can show that for the Y_2 variable,

$$\text{Rel freq } (Y_2) : [(\text{No}, 0.156), (\text{Yes}, 0.844)].$$

Suppose however there is the logical rule that a household must have one of ξ_k, $k = 1, \ldots, 4$, if it has central heating. For convenience, let us denote $Y_1 = \xi_0$ if there is no fuel type used. (That is, $Y_1 = \xi_0$ corresponds to $Y_2 = \text{No}$). Then, this logical rule can be written as the pair $v = (v_1, v_2)$

$$v = \begin{cases} v_1 : y_2 \in \{\text{No}\} \Rightarrow y_1 \in \{\xi_0\}, \\ v_2 : y_2 \in \{\text{Yes}\} \Rightarrow y_1 \in \{\xi_1, \ldots, \xi_4\}. \end{cases} \tag{33}$$

Let us suppose the relative frequencies for both Y_1 and Y_2 pertain as in Table 7 (that is, with the ξ_0 values having zero relative frequencies) except that the values for the original object $u = 25$, are replaced by the new relative frequencies for Y_1 of (.25, .00, .41, .14, .20), respectively. We will refer to this as object $u = 25^*$.

Let us consider Y_1 and $u = 25^*$. In the presence of the rule v, the particular descriptions $x = (Y_1, Y_2) \in \{(\xi_0, \text{No}),\ \xi_1,\ \text{Yes}),\ (\xi_2,\ \text{Yes}),\ (\xi_3,\ \text{Yes}),\ (\xi_4,\ \text{Yes})\}$ can occur. Thus, the relative frequencies for each of the possible ξ_k values have to be adjusted.

Table 7 displays the apparent relative frequencies for each of the possible description vectors before any rules have been invoked. Those values indicated by a $'+'$ are those which are invalidated after invoking the rule v. Let us consider the relative frequency for ξ_2 (solid fuels). Under the rule v, the adjusted relative frequency for object $u = 25^*$ is

$$\pi_{Y_1}(\xi_2; 25^*) = (.3731)/[.0225 + \ldots + .1820] = .5292.$$

Hence, the observed frequency for ξ_2 over all objects becomes under v

$$O_{Y_1}(\xi_2) = .0637 + \ldots + .0000 + .5292 = 1.5902.$$

Similarly, we calculate under v that

$$O_{Y_1}(\xi_0) = 3.8519, O_{Y_1}(\xi_1) = 15.2229, O_{Y_1}(\xi_3) = 2.9761, O_{Y_1}(\xi_4) = 1.3589.$$

It is readily seen that the summation of (32) holds, with $m = 25$. Hence, the relative frequencies for Y_1 are

$$\text{Rel freq}(Y_1) : [(\xi_0, 0.154), (\xi_1, 0.609), (\xi_2, 0.064), (\xi_3, 0.119), (\xi_4, 0.054)].$$

Likewise, we can show that the relative frequencies for Y_2 are

$$\text{Rel freq}(Y_2) : [(\text{No}, 0.154), (\text{Yes}, 0.846)].$$

Therefore, over all regions, 60.9% of households use gas, 6.4% use solid fuel, 11.9% use electricity and 5.4% use other fuels to operate their central heating systems, while 15.4% do not have central heating.

For the original data values for $u = 25$, we can obtain the respective relative frequencies, respectively, as

$$\text{Rel freq}(Y_1) : [(\xi_0, 0.156), (\xi_1, 0.609), (\xi_2, 0.057), (\xi_3, 0.117), (\xi_4, 0.060)];$$

and

$$\text{Rel freq}(Y_2) : [(\text{No}, 0.156), (\text{Yes}, 0.844)].$$

5.5 Interval-valued Modal Variables

Let us suppose the random variable of interest, $Z \equiv Y$ for object u, $u = 1, \ldots, m$, takes values on the intervals $\xi_{uk} = [a_{uk}, b_{uk})$ with probabilities p_{uk}, $k = 1, \ldots, s_u$. One manifestation of such data would be when (alone, or after aggregation) an object u assumes a histogram as its observed value of Z. An example of such a data set would be that of Table 8, in which the data (such as might be obtained from Table 1 after aggregation) display the histogram of weight of women by age-groups (those aged in their 20s, those in their 30s, ..., those in their 80s and above). Thus, for example, we observe that for women in their thirties (object $u = 2$), 40% weigh between 116 and 124 pounds. Figure 3 displays the histograms for the respective age groups.

By analogy with ordinary interval-valued variables (see Section *5.3*), it is assumed that within each interval $[a_{uk}, b_{uk})$, each individual description vector $x \in vir(du)$ is uniformly distributed across that interval. Therefore, for each ξ_k,

$$P\{x \leq \xi_k | x \in vir(d_u)\} = \begin{cases} 0, & \xi_k < a_{uk}, \\ \frac{\xi_k - a_{uk}}{b_{uk} - a_{uk}}, & a_{uk} \leq \xi_k < b_{uk}, \\ 1, & k \geq b_{uk}. \end{cases} \tag{34}$$

A histogram of all the observed histograms can be constructed as follows. Let $I = [\min_{k,u \in E} a_{ku}, \max_{k,u \in E} b_{ku}]$ be the interval which spans all the observed values of Z in $\mathcal{X}$, and let I be partitioned into r subintervals $I_g = [\xi_{g-1}, \xi_g)$, $g = 1, \ldots, r-1$, and $I_r = [\xi_{r-1}, \xi_r)$. Then, the observed frequency for the interval I_g is

$$O_Z(g) = \sum_{u \in E} \pi_Z(g; u) \tag{35}$$

where

$$\pi_Z(g; u) = \sum_{k \in Z(g)} \frac{||Z(k; u) \cap I_g||}{||Z(k; u)||} p_{uk} \tag{36}$$

where $Z(k; u)$ is the interval $[a_{uk}, b_{uk})$, and where the set $Z(g)$ represents all those intervals $Z(k; u)$ which overlap with I_g, for a given u. Thus, each term in the summation in (36) represents that portion of the interval $Z(k; u)$ which is spanned by I_g and hence that

proportion of its observed relative frequency (p_{uk}) which pertains to the overall histogram interval I_g.

It follows that

$$\sum_{g=1}^{r} O_Z(g) = m. \tag{37}$$

Hence, the relative frequency for the interval I_g is

$$p_g = O_Z(g)/m. \tag{38}$$

The set of values $\{(p_g, I_g), g = 1, \ldots, r\}$ together represents the relative frequency histogram for the combined set of observed histograms.

The **empirical density function** can be derived from

$$f(\xi) = \frac{1}{m} \sum_{u \in E} \sum_{k=1}^{s_u} \frac{I_{uk}(\xi)}{||Z(u;k)||} p_{uk}, \quad \xi \in \Re, \tag{39}$$

where $I_{uk}(\cdot)$ is the indicator function that ξ is in the interval $Z(u;k)$.

The **symbolic sample mean** becomes

$$\bar{Z} = \frac{1}{2m} \sum_{u \in E} \{\sum_{k=1}^{s_u} (b_{uk} + a_{uk}) p_{uk}\}, \tag{40}$$

and the **symbolic sample variance** is

$$S^2 = \frac{1}{3m} \sum_{u \in E} \{\sum_{k=1}^{s_u} (b_{uk}^2 + b_{uk}a_{uk} + a_{uk}^2) p_{uk}\} - \frac{1}{4m^2} \{\sum_{u \in E} \sum_{k=1}^{s_u} (b_{uk} + a_{uk}) p_{uk}\}^2. \tag{41}$$

To illustrate, take the data of Table 8; and let us construct the histogram on the $r = 10$ intervals $[60 - 75), [75 - 90), \ldots, [195 - 210]$. Take the $g = 5$th interval $I_5 = [120 - 135)$. Then, from Table 8, it follows, from (36), that

$$\pi(5;1) = (\frac{132 - 130}{132 - 120})(.24) + (\frac{135 - 132}{144 - 132})(.06) = 0.255;$$

and likewise,

$$\pi(5;2) = 0.5300, \quad \pi(5;3) = 0.1533, \quad \pi(5;4) = 0.1800,$$

$$\pi(5;5) = 0.0364, \quad \pi(5;6) = 0.0000, \quad \pi(5;7) = 0.1200.$$

Hence, from (35), the "observed relative frequency" is

$$O(g = 5) = 1.2747;$$

and from (38), the relative frequency is

$$p_5 = 0.1821.$$

The complete set of histogram values $\{p_g, I_g, g = 1, \ldots, 10\}$ is given in Table 9; and displayed in Figure 4. Likewise, from (40) and (41), $\bar{Z} = 143.9$ and $S^2 = 447.5$, respectively.

It is implicit in the formula (35)-(40) that the rules v hold. Suppose the data of Table 8 represented (instead of weights) costs (in \$s) involved for a certain procedure at seven different hospitals. Suppose further there is a minimum charge of \$100. This translates into a rule

$$v : \{Y < 100\} \Rightarrow \{p_k = 0\}. \tag{42}$$

The data for all hospitals (objects) except the first satisfy this rule. However, some values for the first hospital do not. Hence, there needs to be an adjustment to the overall relative frequencies to accommodate this rule. Let us assume the histogram now spans $r = 8$ intervals $[100, 105), \ldots, [195, 210]$. The relative frequencies for these data after taking into account the rule (42) are shown in column (d) of Table 9.

6 Descriptive Bivariate Statistics

Many of the principles developed for the univariate case can be expanded to a general p-variate case, $p > 1$. In particular, this permits derivation of dependence measures. Thus, for example, calculation of the covariance matrix aids the development of methodologies in, for example, principal component analysis, discriminant analysis and cluster analysis. Recently, Billard and Diday (2000, 2002) extended these methods to fit multiple linear regression models to interval-valued data and histogram data, respectively. Different but related ideas can be extended to enable the derivation of resemblance measures such as Euclidean distances, Minkowski or L_q distances, and Mahalanobis distances. $p > 1$. We restrict attention to two variables, and for the sake of discussion, let us suppose we are interested in the specific variables Z_1 and Z_2 over the space $Z = Z_1 \times Z_2$. We consider multi-valued, interval-valued, and histogram-valued variables, in turn.

6.1 Multi-valued Variables

Some Definitions

We first consider multi-valued variables. Analogously to the definition of the observed frequency of specific values for a single variable given in (12), we define the observed frequency that $(Z_1 = \xi_1, Z_2 = \xi_2)$ by

$$O_{Z_1,Z_2}(\xi_1, \xi_2) = \sum_{u \in E} \pi_{Z_1,Z_2}(\xi_1, \xi_2; u) \tag{43}$$

where

$$\pi_{Z_1,Z_2}(\xi_1, \xi_2; u) = \frac{|\{x \in vir(d_u) | x_{Z_1} = \xi_1, Z_2 = \xi_2\}|}{|vir(d_u)|} \tag{44}$$

is the percentage of the individual description vectors $\boldsymbol{x} = (x_1, x_2)$ in $vir(d_u)$ for which $(Z_1 = \xi_1, Z_2 = \xi_2)$; note that $\pi(\cdot)$ is a real number on $\Re$ in contrast to its being a positive integer for classical data. We can show that

$$\sum_{\xi_1 \in Z_1, \xi_2 \in Z_2} O_{Z_1,Z_2}(\xi_1, \xi_2) = m.$$

Then, we define the **empirical joint frequency distribution** of Z_1 and Z_2 as the set of pairs $[\boldsymbol{\xi}, O_{Z_1 Z_2}(\boldsymbol{\xi})]$ where $\boldsymbol{\xi} = (\xi_1, \xi_2)$ for $\boldsymbol{\xi} \in \mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2$; and we define the **empirical joint histogram** Z_1 and Z_2 as the graphical set $[\boldsymbol{\xi}, \frac{1}{m} O_{Z_1 Z_2}(\xi)]$,

When Z_1 and Z_2 are quantitative symbolic variables, we can also define symbolic sample covariance and correlation functions. The **symbolic sample covariance function** between the symbolic quantitative multi-valued variables Z_1 and Z_2 is given by

$$S_{Z_1,Z_2} \equiv S_{12} = \left[\frac{1}{m} \sum_{\xi_1, \xi_2} (\xi_1 \times \xi_2) O_{Z_1,Z_2}(\xi_1, \xi_2) \right] - (\bar{Z}_1)(\bar{Z}_2) \tag{45}$$

where $\bar{Z}_i$, $i = 1, 2$, are the symbolic sample means of the univariate variables Z_i, $i = 1, 2$, respectively, defined in (17).

The **symbolic sample correlation function** between the multi-valued symbolic variables Z_1 and Z_2 is given by

$$r(Z_1, Z_2) = S_{Z_1,Z_2} / \sqrt{(S^2_{Z_1})(S^2_{Z_2})} \tag{46}$$

where $S^2_{Z_1}$, $i = 1, 2$, are the respective univariate symbolic sample variances, as defined earlier in (18).

An Example

We illustrate these statistics with the cancer data given previously in Table 3 above. We have that $O_{Z_1,Z_2}(\xi_1 = 0, \xi_2 = 0) \equiv O(0, 0)$, say, is

$$O(0,0) = \frac{0}{1} + \frac{1}{3} + \frac{0}{1} + \frac{0}{2} + \frac{1}{1} + \frac{0}{2} + \frac{0}{2} + \frac{0}{2} = \frac{4}{3}$$

where, e.g., $\pi(0, 0; 1) = 0/1$ since the individual vector (0,0) occurs no times out of a total of $|vir(d_1)| = 1$ vectors in $vir(d_1)$; likewise, $\pi(0, 0; 2) = 1/3$ since (0,0) is one of three individual vectors in $vir(d_2)$ and so on and where the summation does not include the $u = 5$ term since $|vir(d_5)| = 0$. Similarly, we can show that

$$\begin{aligned} O(0,1) &= 0; \quad O(0,2) = 0; \quad O(0,3) = 0; \\ O(1,0) &= 1/3; \quad O(1,1) = 4/3; \quad O(1,2) = 5/2; \quad O(1,3) = 5/2. \end{aligned}$$

Hence, the symbolic sample covariance is, from equation (45),

$$S_{12} = \frac{1}{8}(\frac{83}{6}) - (\frac{5}{6})(\frac{83}{48}) = 0.2882.$$

Therefore, the symbolic correlation function is, from equation (46),

$$r(Z_1, Z_2) = \frac{(.2282)}{\sqrt{(0.1239)(1.1975)}} = 0.7482.$$

6.2 Interval-valued Variable

Some Definitions

For interval-valued variables, let us suppose the specific variables of interest Z_1 and Z_2 have observations on the rectangle $Z(u) = Z_1(u) \times Z_2(u) = ([a_{1u}, b_{1u}], [a_{2u}, b_{2u}])$ for each $u \in E$. As before, we make the intuitive choice of assuming individual vectors $x \in vir(d_u)$ are each uniformly distributed over the respective intervals $Z_1(u)$ and $Z_2(u)$. Therefore, the joint distribution of (Z_1, Z_2) is a copula $C(z_1, z_2)$; see Nelson, (1999); and Schweizer and Sklar (1983). An expanded discussion of the role of copulas in symbolic data will be presented elsewhere.

We define the **empirical joint density function** for (Z_1, Z_2) as

$$f(\xi_1, \xi_2) = \frac{1}{m} \sum_{u \in E} \frac{I_u(\xi_1, \xi_2)}{||Z'(u)||} \tag{47}$$

where $I_u(\xi_1, \xi_2)$ is the indicator function that (ξ_1, ξ_2) is or is not in the rectangle $Z'(u)$ and where $||Z'(u)||$ is the area of this rectangle. Analogously with (26), we can find the **joint histogram** for Z_1 and Z_2 by graphically plotting $\{R_{g_1 g_2}, p_{g_1 g_2}\}$ over the rectangles $R_{g_1 g_2} = \{[\xi_{1,g_1-1}, \xi_{1 g_1}) \times [\xi_{2,g_2-1}, \xi_{2 g_2})\}$, $g_1 = 1, \ldots, r_1$, $g_2 = 1, \ldots, r_2$, where

$$p_{g_1, g_2} = \frac{1}{m} \sum_{u \in E} \frac{||Z'(u) \cap R_{g_1 g_2}||}{||Z'(u)||}, \tag{48}$$

i.e., $p_{g_1 g_2}$ is the probability an arbitrary individual description vector lies in the rectangle $R_{g_1 g_2}$. Then, if the "volume" on the rectangle $R_{g_1 g_2}$ on the histogram represents the probability, its height would be

$$f_{g_1 g_2} = [(\xi_{1 g_1} - \xi_{1,g_1-1})(\xi_{2 g_2} - \xi_{2,g_2-1})]^{-1} p_{g_1 g_2}. \tag{49}$$

The **symbolic sample covariance** function is obtained analogously to the derivation of the mean and variance for a univariate interval-valued variable, viz.,

$$\begin{aligned} Cov(Z_1, Z_2) = S_{Z_1 Z_1} \\ &= \int_{-\infty}^{\infty} (\xi_1 - \bar{Z}_1)(\xi_2 - \bar{Z}_2) f(\xi_1, \xi_2) d\xi_1 d\xi_2 \\ &= \frac{1}{m} \sum_{u \in E} \frac{1}{(b_{1u} - a_{1u})(b_{2u} - a_{2u})} \int \int_{(\xi_1, \xi_2) \in Z(u)} \xi_1 \xi_2 d\xi_1 d\xi_2 - \bar{Z}_1 \bar{Z}_2 \\ &= \frac{1}{4m} \sum_{u \in E} (b_{1u} + a_{1u})(b_{2u} + a_{2u}) - \frac{1}{4m^2} [\sum_{u \in E} (b_{1u} + a_{1u})][\sum_{u \in E} (b_{2u} + a_{2u})]. \end{aligned} \tag{50}$$

Hence, the **symbolic sample correlation function** is defined by

$$r(Z_1, Z_2) = S_{Z_1 Z_2} / \sqrt{S^2_{Z_1} S^2_{Z_2}} \tag{51}$$

where $S^2_{Z_1}$, $i = 1, 2$, were given in (29). As before, the summation in (50) and (51) is only over those u for which $vir(d_u)$ is not empty.

An Example

We illustrate these concepts with the data of Table 4. In particular, let us consider the systolic and diastolic blood pressure levels, Y_2 and Y_3, respectively. Suppose further there is a logical rule that specifies that the diastolic blood pressure must be less than the systolic blood pressure, that is, $Y_2(u) \leq Y_1(u)$. The observations $u = 1, \ldots, 10$, all satisfy this rule. However, had there been another observation represented as given in the table by the $u = 11$ data point, then the rule v is violated since $Y_2(11) > Y_1(11)$. Therefore, $|vir(d_{11})| = 0$, and so our subsequent calculations would not include this observation.

Suppose we want to find the joint histogram for these two variables. Let us suppose we want to construct the histogram on the rectangles $R_{g_2 g_3} = \{[\xi_{2,g_2-1}, \xi_{2,g_2}) \times [\xi_{3,g_3-1}, \xi_{3,g_3})\}$, $g_2 = 1, \ldots, 10$, $g_3 = 1, \ldots, 6$, where the sides of these rectangles are the intervals (given in Table 10) used in constructing the univariate histogram function for each of these variables.

Then, from equation (48), we can calculate the probability $p_{g_2 g_3}$ that an arbitrary description vector lies in the rectangle $R_{g_2 g_3}$. For example, for $g_2 = 6$, $g_3 = 4$,

$$\begin{aligned} p_{64} &= P\{x \in [140, 150) \times [80, 90)\} \\ &= \frac{1}{10}\{0 + 0 + 0 + \frac{2}{32} \cdot \frac{10}{28} + 0 + \frac{2}{8} \cdot \frac{10}{30} + \frac{10}{30} \cdot \frac{10}{14} + \frac{10}{80} \cdot \frac{10}{40} + 0 + \frac{10}{40} \cdot \frac{10}{22}\} \\ &= .04886. \end{aligned}$$

Table 10 provides all the probabilities $p_{g_2 g_3}$. The table also gives the marginal totals which correspond to the corresponding probabilities p_{g_j}, $j = 2, 3$, of the univariate case obtained from (26). A plot of the joint histogram is given in Figure 5 where the heights $f_{g_2 g_3}$ are calculated from equation (49).

Further, using equations (50) and (51), we can show that the covariance between the systolic and diastolic blood pressure is $Cov(Y_2, Y_3) = 257.92$ and that the correlation coefficient between these two variables is $Corr(Y_2, Y_3) = r_{Y_2 Y_3} = 0.858$. Similarly, the covariance between pulse rate and systolic blood pressure is $Cov(Y_1, Y_2) = 194.17$ with correlation coefficient $r(Y_1, Y_2) = 0.685$; and the covariance between pulse rate and diastolic blood pressure is $Cov(Y_1, Y_3) = 141.04$ with correlation coefficient $r(Y_1, Y_3) = 0.820$.

Therefore, we can show that if the systolic and diastolic blood pressures are the predictor variables and pulse rate is the dependent variable, then for the data of Table 4, the linear regression model becomes $Y_1 = 14.2 - 0.04 Y_2 + 0.83 Y_3$.

6.3 Modal-valued Variables

Some definitions

Let us develop corresponding results for quantitative modal-valued variables where the data are recorded as histograms. Suppose interest centers on the two specific variables Z_1 and Z_2. Suppose for each object u, each variable $Z_j(u)$ takes values on the subintervals $\xi_{juk} = [a_{juk}, b_{juk})$ with relative frequency p_{juk}, $k = 1, \ldots, s_{ju}$, $j = 1, 2$, and $u = 1, \ldots, m$, with $\sum_{k=1}^{s_{ju}} p_{juk} = 1$. When $s_{ju} = 1$ and $p_{juk} = 1$ for all j, u, k values, the data are interval-valued realizations.

Then, by extending the derivations of subsection *6.2*, we can show that the **empirical joint density function** for (Z_1, Z_2) at the value (ξ_1, ξ_2) is

$$f(\xi_1, \xi_2) = \frac{1}{m} \sum_{u \in E} \{ \sum_{k_1=1}^{s_{1u}} \sum_{k_2=1}^{s_{2u}} \frac{p_{1uk_1} p_{2uk_2} I_{uk_1,k_2}(\xi_1, \xi_2)}{||Z_{k_1k_2}(u)||} \} \tag{52}$$

where $||Z_{k_1k_2}(u)||$ is the area of the rectangle $Z_{k_1k_2}(u) = [a_{1uk_1}, b_{1uk_1}) \times [a_{2uk_2}, b_{2uk_2})$, and $I_{uk_1k_2}(\xi_1, \xi_2)$ is the indicator variable that the point (ξ_1, ξ_2) is (is not) in the rectangle $Z_{k_1k_2}(u)$. Analogously with (36) and (48), the joint histogram for Z_1 and Z_2 is found by plotting $\{R_{g_1g_2}, p_{g_1g_2}\}$ over the rectangles $R_{g_1g_2} = \{[\xi_{1,g_1-1},\ \xi_{1g_1}) \times [\xi_{2,g_2-1},\ \xi_{2g_2})\}$, $g_1 = 1, \ldots, r_1$, $g_2 = 1, \ldots, r_2$ with

$$p_{g_1g_2} = \frac{1}{m} \sum_{u \in E} \sum_{k_1 \in Z(g_1)} \sum_{k_2 \in Z(g_2)} \frac{||Z(k_1, k_2; u) \cap R_{g_1g_2}||}{||Z(k_1, k_2; u)||} p_{1uk_1} p_{2uk_2} \tag{53}$$

where $Z(g_j)$ represents all the intervals $Z(k_1; u) \equiv [a_{juk_j}, b_{juk_j})$, $j = 1, 2$, which overlaps with the rectangle $R_{g_1g_2}$ for each given u value.

We can then derive the symbolic covariance function as

$$Cov(Y_1, Y_2) = \frac{1}{4m} \sum_{u \in E} \{ \sum_{k_1=1}^{s_{1u}} \sum_{k_2=1}^{s_{2u}} p_{1uk_1} p_{2uk_2} (b_{1uk_1} + a_{1uk_1})(b_{2uk_2} + a_{2uk_2}) \}$$

$$- \frac{1}{4m^2} [\sum_{u \in E} \{ \sum_{k_1=1}^{s_{1u}} p_{1uk_1} (b_{1uk_1} + a_{1uk_1}) \}] [\sum_{u \in E} \{ \sum_{k_1=2}^{s_{2u}} p_{2uk_2} (b_{2uk_2} + a_{2uk_2}) \}]. \tag{54}$$

The symbolic mean and symbolic variance for each of Z_1 and Z_2 can be obtained from (40) and (41), respectively. *An Example*

The data of Table 10 record histogram hematocrit (Y_1) values and hemoglobin (Y_2) values for each of $m = 5$ objects. Then, from (54), we can calculate the covariance between Y_1 and Y_2 as $Cov(Y_1, Y_2) = 6.256$. Billard and Diday (2002a) have recently extended these results to the problem of fitting a regression equation to histogram data. Applying their methodology to these data, we can show that $Y_1 = -2.134 + 3.172 Y_2$.

7 Principal Components Analysis

A principal component analysis is designed to reduce p-dimensional observations into s-dimensional (where usually $s << p$) components. More specifically a principal component is a linear combination of the original variables, and the goal is to find those s principal components which together explain most of the underlying variance-covariance structure of the p variables.

Cazes et al. (1997), Chouakria (1998), and Chouakria et al. (1998) develop a method of conducting principal component analysis on symbolic data for which each symbolic variable Y_j, $j = 1, \ldots, p$, takes interval values $\xi_u = [a_{uj}, b_{uj}]$, say, for each object $u = 1, \ldots, m$, and where each object represents n_u individuals; for simplicity, we take $n_u = 1$.

Each symbolic data point is represented by a hyperrectangle with 2^p vertices. Thus, each hyperrectangle can be represented by a $2^p \times p$ matrix $\boldsymbol{M}_u$ with each row containing the coordinate values of a vertex R_k, $k = 1, \ldots, 2^p$, of the hyperrectangle. Then, a $(m \cdot 2^p \times p)$ matrix $\boldsymbol{M}$ is constructed of the $\{\boldsymbol{M}_u, u = 1, \ldots, m\}$, viz.,

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_1 \\ \vdots \\ \boldsymbol{M}_m \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} a_{11} \ldots a_{1p} \\ \ddots \\ b_{11} \ldots b_{1p} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} a_{m1} \ldots a_{mp} \\ \vdots \\ b_{m1} \ldots b_{mp} \end{bmatrix} \end{pmatrix}.$$

For example, if $p = 2$, the data $\xi_u = ([a_{u1}, b_{u1}], [a_{u2}, b_{u2}])$ is transformed to the $2^p \times p = 2^2 \times 2$ matrix

$$\boldsymbol{M}_u = \begin{bmatrix} a_{u1} & a_{u2} \\ a_{u1} & b_{u2} \\ b_{u1} & a_{u2} \\ b_{u1} & b_{u2} \end{bmatrix},$$

and likewise for $\boldsymbol{M}$.

The matrix $\boldsymbol{M}$ is now treated as though it represents classical p-variate data for $n = m \cdot 2^p$ individuals. Chouakria (1998) has shown that the basic theory for a classical analysis carries through; hence, a classical principal component analysis can be applied. If observations have weights $p_i \geq 0$, then in this transformation of the symbolic to the classical matrix $\boldsymbol{M}$, each vertex of the hyperrectangle now has the same weight $(p_i 2^{-p})$. Let $Y_1^*, \ldots, Y_s^*$, $s \leq p$, denote the first "numerical" principal components with associated eigenvalues $\lambda_1 \geq \ldots \geq \lambda_s \geq 0$ which result from this analysis. We then construct the interval principal components $Y_1^I, \ldots, Y_s^I$ as follows.

Let L_u be the set of row indices in $\boldsymbol{M}$ identifying the vertices of the hyperrectangle R_u, i.e., L_u represents the rows of $\boldsymbol{M}_u$ describing the symbolic data ξ_u.. For each $k = 1, \ldots, 2^p$, in L_u, let y_{kv} be the value of the numerical principal component Y_u^*, $v = 1, \ldots, s$, for that row k. Then, the interval principal component Y_v^I for object u, is given by

$$y_{uv} = [y_{uv}^a, y_{uv}^b]$$

where $y_{uv}^a = \min_{k \in L_u}(y_{kv})$ and $y_{uv}^b = \max_{k \in L_u}(y_{kv})$.

Then, plotting gives us the data represented as hyperrectangles of principal components in s-dimensional space. Thus, taking $v = 1, 2$, we have each object $u = 1, \ldots, m$, represented by a rectangle with axis corresponding, respectively, to the first two principal components.

As an alternative to using the vertices of the hyperrectangles R_u as above, the centers of the hyperrectangles could be used. In this case, each object data point

$$\xi_u = ([a_{u1}, b_{u1}], \ldots, [a_{up}, b_{up}])$$

is transformed to

$$x_u^c = (x_{u1}^c, \ldots, x_{up}^c), \quad u = 1, \ldots, m,$$

where

$$x_{uj}^c = (a_{uj} + b_{uj})/2, \quad j = 1, \ldots, p.$$

Thus, the symbolic data-matrix $\boldsymbol{X}$ has been transformed to a classical $m \times p$ matrix $\boldsymbol{X}^c$ with classical variables $Y_1^c, \ldots, Y_p^c$, say.

Then, the classical principal component analysis is applied to the classical data $\boldsymbol{X}^c$. The v^{th} center principal component for object u is, for $v = 1, \ldots, s$,

$$y_{uv}^c = \sum_{j=1}^{p} (x_{uj}^c - \bar{x}_j^c) w_{uv}$$

where the mean of the values for Y_j is

$$\bar{x}_j^c = \frac{1}{m} (\sum_{u=1}^{m} x_{uj}^c),$$

and where $w_v = (w_{1v}, \ldots, w_{pv})$ is the v^{th} eigenvector. Since each centered coordinate x_{uj}^c lies in the interval $[a_{uj}, b_{uj}]$ and since the principal components are linear functions of x_{uj}^c, we can obtain the interval principal components as $[y_{uv}^{ca}, y_{uv}^{cb}]$ where

$$y_{uv}^{ca} = \sum_{j=1}^{p} \min_{a_{uj} \le x_{ij}^r \le b_{uj}} (x_{uj}^c - \bar{x}_j^c) w_{uv}$$

and

$$y_{uv}^{cb} = \sum_{j=1}^{p} \max_{a_{uj} \le x_{ij}^r \le b_{uj}} (x_{uj}^c - \bar{x}_j^c) w_{uv}.$$

The methodology for the centers method is illustrated with the interval data of Table 11, extracted from U.S. Census data. There are $m = 9$ objects representing classes/regions of the United States described by $p = 6$ symbolic variables; specifically Y_1 = median household income, Y_2 = average income per household member, Y_3 = percentage of people covered by health insurance, Y_4 = percentage of population over 25 years who have completed high school, Y_5 = percentage of population over 25 years who have earned at least a bachelor's degree, and Y_6 = travel time in minutes to go to work. For simplicity, we restrict the analysis to the first three variables Y_j, $j = 1, 2, 3$, only. The $s = 3(= p)$ eigenvalues (together with the percentage of the total sum of squares contributed by the respective principal components are $\lambda_1 = 1.909$ (63.64%), $\lambda_2 = 0.914$ (30.46%) and $\lambda_3 = 0.177$ (5.91%) Thus, the first and second components together explain 94.10% of the variation. The results are plotted in Figure 6, and the distinct clusters are evident, with the SouthE and SouthW regions forming one cluster, the Mountain and SouthA regions forming a second cluster, New England stands on its own as a third grouping, and these latter two clusters are spanned by a cluster consisting of the remaining four regions (Mid WestE, MidAtl, MidWestW, and the Pacific). Executing the same analysis but with the vertices approach gives similar results as illustrated in Bock and Diday (2000).

Thus far, in this section, the object u has been assumed to be a class of size 1. However, more generally, it may be a class of size n_u, with $n_u + \ldots + n_m = m^*$. The methods carry through readily.

Principal components as a method is designed to reduce the dimension of the data space to $s < p$. Dimensionality reduction methods for interval data have also been considered by Ichino (1988) and Ichino and Yaguchi (1994) by using generalized Minkowsky metrics, and by Nagabushan et al. (1995) by using Taylor series ideas.

These principal component methods are limited at present to interval-valued symbolic data. Symbolic principal components for other types of symbolic data such as for modal variables (as would appear, e.g., in pixel satellite or tomography data) remains as an outstanding problem.

8 Symbolic Clustering

In this section, we consider clustering methods for symbolic data. The aim is to classify the objects in Ω into clusters (or classes) $C_1, \ldots, C_m$, which are internally as homogeneous as possible and externally as distinct from each other as possible. This process is distinct from the construction of classes procedure discussed in Section 3, in which pre-assigned criteria were selected for class membership. In contrast, the traditional clustering problem seeks to find "natural" groupings from the data. There is however a link between the two procedures as we shall see later. Classical clustering methods have been well described in

Breiman et al. (1984). We follow here the approach of Chavent (1998) for criterion-based divisive clustering.

Chavent (1998) distinguishes between two types of sets of variables. The first deals with quantitative variables $Y_1, \ldots, Y_k$, in which Y_j takes a single (classical) value or Y_j takes an interval value $Y_j(u) = [a, b] \subset \Re$, for $u \in \Omega$. Or, the $Y_1, \ldots, Y_p$ can be categorical-type variables, viz., they can be either (classical) ordinal values with $Y_j(u) \in \mathcal{Y}_j$, multi-valued variable with $Y_j(u) \subset Y_j$ and $\mathcal{B}_j = P(\mathcal{Y}_j)$, or modal variables with $Y_j(u) = \pi_j$ is a probability or frequency distributions on $\mathcal{Y}_j$ and $\mathcal{B}_j$ is the set of all probability distributions on Y_j. In particular, the $Y_1, \ldots, Y_p$ cannot consist of a mixture of quantitative-type and categorical-type variables.

Let $D = (d_{uv})$ be the $n \times n$ matrix of distance measures between objects u, $v \in \Omega$. First, let us take quantitative-type variables and let us assume all Y_j are interval valued. (Adjustment for the case where some Y_j are classical variables is straightforward). Suppose we have the intervals $\xi_u = [a_{uj}, b_{uj}]$ and $\xi_v = [a_{vj}, b_{vj}]$, $j = 1, \ldots, p$, u, $v \in \Omega$. We seek a distance function $\delta_j(u, v)$ between objects u and v. There are a number of possible such functions used in the clustering process.

We can define a symbolic Hausdorff distance for Y_j as

$$\delta_j(u, v) = \max\{|a_{uj} - a_{vj}|, |b_{uj} - b_{vj}|\}. \tag{55}$$

If we take a distance function $d(u, v)$

$$d(u, v) = \left(\sum_{j=1}^{p} [\delta_j(u, v)]^2 \right)^{1/2},$$

and use the specific $\delta_j(\cdot)$ of (55), we have

$$d(u, v) = \left(\sum_{j=1}^{p} [\max\{|a_{uj} - b_{vj}|, |b_{uj} - b_{vj}|\}]^2 \right)^{1/2}. \tag{56}$$

Notice that $d(u, v)$ in (56) reduces to the Euclidean distance on $\Re^p$ when all Y_j are classical variables. The distance in (55) can be normalized to

$$d'(u, v) = \left\{ \sum_{j=1}^{p} [m_j^{-1} \delta_j(u, v)]^2 \right\}^{1/2} \tag{57}$$

for suitable choices of m_j. Chavent (1998) suggests

$$m_j^2 = (2n^2)^{-1} \sum_{u=1}^{n} \sum_{v=1}^{n} [\delta_j(u, v)]^2;$$

or

$$m_j = \text{ length of the domain of } \mathcal{Y}_j.$$

When the variables are categorical, we first construct a symbolic frequency table as follows. Suppose Y_j can take p_j possible values. If Y_j is modal, then we already have that the probability of obtaining the kth possibility is π_{jk}, $k = 1, \ldots, p$, $\Sigma_k \pi_{jk} = 1$. Each possible value is "labelled" as a new variable Y_{jk}, $k = 1, \ldots, p_j$. For a multi-valued variable $Y_j(u)$, we assume that the observed categories are uniformly distributed on $\mathcal{Y}_j(u)$ and the probability for those possibilities not observed is zero. The required frequency table then represents an $n \times t$ matrix of classical-type data $\boldsymbol{X} = (f_{uj})$, where $u = 1, \ldots, n$, $j = 1, \ldots, t$, where $t = p_1 + \ldots + p_p$ is the number of "variables". However, the entries f_{uj} are real numbers and not necessarily integers (as would be the case for classical data).

Then, the distance between two objects u and v in Ω is given by

$$d(u, v) = \left[\sum_{j=1}^{t} \frac{1}{p_{.j}} \left(\frac{p_{uj}}{p_{u.}} - \frac{p_{vj}}{p_{v.}} \right)^2 \right]^{1/2} \tag{58}$$

where

$$p_{uj} = f_{uj}/np, \quad p_{u.} = \sum_{j=1}^{t} p_{uj}, \quad p_{.j} \sum_{u=1}^{n} p_{uj}.$$

For example, suppose we have the symbolic data of Table 12 which has data for three objects relating to a cancer prognosis Y_1 (poor, medium, good) and weight Y_2 (below average, average, above average). Notice that object $u = 3$ is in fact a classical data point. Here, Y_1 is modal but Y_2 is categorical. The symbolic data $\boldsymbol{X}$ of Table 12 is transformed into classical data for 6 variables, as shown in Table 13. Hence, the distance between objects $u = 1$ and $v = 2$ is, from (58), $d(1, 2) = 0.913$; likewise, $d(1, 3) = 1.118$ and $d(2, 3) = 1.812$.

Another cluster partitioning criteria is the following symbolic version of the classical within-class variance criteria, or inertia, given by

$$I(C_i) = \frac{1}{2\mu_i} \sum_{u \in C_i} \sum_{v \in C_i} w_u w_v [d(u, v)]^2 \tag{59}$$

where w_u is the weight associated with object u, and where $\mu_i = \sum_{u \in C_i} w_u$. If we take $w_u = 1/n$, then

$$I(C_i) = (mn_i)^{-1} \sum_{u,v \in C_i} \sum_{u>v} [d(u, v)]^2$$

where $n_i = |C_i|$ is the size of the class C_i.

Suppose at the mth stage of the clustering process, we have clusters $(C_1, \ldots, C_m)$. Thus, initially when $m = 1$, $C_1 \equiv \Omega$. We wish to select which cluster C_i is to be partitioned in

two clusters C_i^1 and C_i^2 at the next $(m+1)$th stage. This is achieved by selecting that i which maximizes

$$I(C_i) - I(C_i^1) - I(C_i^2).$$

The cluster C_i is partitioned into these two subclusters (C_i^1, C_i^2) as follows. We choose the best partition from among all those "induced by the set of all possible binary questions", where "best" is specified by the distance measure adopted [e.g., minimizing $I(C_i)$]. Let the binary cut be specified by the binary function q_c where we partition C_i into C_i^1 and C_i^2 according to

$$C_1 : \{u \in C | q_c(u) = \text{true}\}$$

$$C_2 : \{u \in C | q_c(u) = \text{false}\}.$$

For interval data $[a_u, b_u]$, we define q_c, for object u for $m_u = (a_u + b_u)/2$, by

$$q_c(u) = \begin{cases} \text{true} & \text{if } m_u \leq c, \\ \text{false} & \text{if } m_u > c. \end{cases}$$

For modal variables (including categorical data after transformation to modal variable format), with data value $Y_j(u) = \pi_u$, we define q_c

$$q_c(u) = \begin{cases} \text{true} & \text{if } \sum\limits_{x \leq c} \pi_u(x) \geq 0.5, \\ \text{false} & \text{if } \sum\limits_{x \leq c} \pi_u(x) < 0.5, \end{cases}$$

where the summation is over all individual x values, $x \leq c$.

To illustrate the data of Table 11, and suppose we wish to use the normalized Hausdorff distance measure of (57) with weight m_j corresponding to the length of the domain Y_j. Thus, $m_1 = 14.52$, $m_2 = 6.10$, $m_3 = 17.90$, $m_4 = 14.60$, $m_5 = 23.00$ and $m_6 = 15.60$. We have the three binary functions as

$$\begin{aligned} q_1 &= Y_1 = \text{ median household income} \leq \$41,900 \\ q_2 &= Y_3 = \text{ percentage health insurance} \leq 89.80\% \\ q_3 &= Y_4 = \text{ percent completed high school} \leq 84.375\%. \end{aligned}$$

The clusters of Figure 7 emerge; see Chavent (2000).

In terms of the queries of assertions of Section 3, we observe, for example, that the cluster $C_3 = \{\text{MidWestE, Mountain, Pacific}\}$ matches the assertion that

$$C_3 : [Y_1 > 41.90] \wedge [Y_3 \leq 89.80] \wedge [Y_4 > 84.375];$$

i.e., the output of the clustering process gives us a class (or cluster) represented by the symbolic object for which the median household income exceeds \$41900, fewer than 89.90% have health coverage and at least 84.375% completed high school.

Chavent's method is a divisive hierarchical clustering procedure which starts with all the objects in a single change cluster, and is a form of monothetic divisive clustering method built by considering one variable at a time. Monothetic divisive methods are not dissimilar from discriminant analysis methods, see, e.g., the CART algorithm of Breiman et al. (1984) or the ID3 algorithm of Quinlan (1986); see, also, Périnal and Lechevallier (2000) for specifics on symbolic discrimation analysis. Monothetic divisive clustering for conceptual objects were first introduced in Michalski et al. (1981) and Michalski and Stepp (1983).

In contrast to divisive methods, agglomerative methods start with each object in Ω being a cluster of size one, itself; with the clustering algorithm developed to merge objects into large classes. For symbolic objects, merging criteria revolve around finding those symbolic objects which are similar (as measured by an appropriate similarity index) and relate to the assertions described in Section 3. Building upon Diday's (1986) development of pyramid clusters for classical data, Brito (1994, 1995, 2000) gives an algorithm for developing pyramid clusters for symbolic data, where pyramidical clusters are defined as families of nested overlapping classes (i.e., a class can belong to two distinct clusters, in contrast to a pure hierarchical cluster in which classes are distinct or are entirely contained within another class). Brito and DeCarvalho (1999) extended this to the case where specific hierarchical rules exist, and DeCarvalho et al. (1999) extend this to dependency rules. Polaillon (2000) develops pyramidal clusters for interval data by using Galois lattice reductions. DeCarvalho et al. (1999) look at dynamical clustering of Boolean symbolic objects based on a content dependent proximity measure. Bock and Diday (2000) provides a comprehensive coverage of these approaches along with elucidating examples. With the possible exception of Polaillon's (1998) Galois lattice reduction theory, results tend to be limited to the methods themselves with theoretical justifications remaining as outstanding problems still to be addressed.

9 Three-way Data

Factorial analysis for three-way data tables was considered by Cazes et al. (1997). More recently, Loustaunau et al. (1997) and Gettler-Summa and Pardoux (2000) have studied three-way data more broadly. In this context, there are T (say) arrays with each array consisting of a symbolic data set $\boldsymbol{X}_t$, $t = 1, \ldots, T$, for a population Ω (or E) of size n_t with p_t symbolic variables as in the previous sections. Here, the t may represent times at which measurements were taken; or they may correspond to 'experts' each providing their set of symbolic measurements X_t; or there may be spatial points; or so on. Thus, Morineau et al. (1994) studied multiple time series relating to newspaper sales over a three year period ($T = 156$) for $n_t = 1577$ outlets $i \in \Omega$ and $p_t = 157$ symbolic variables; while Loustaunau et al. (1997), represented experts each providing their assessments (as measurements X_t) on $p_t = 16$ modal variables relative to the installation of optical network units. Gettler-Summa

and Pardoux (2000) present essentially three methods to approach such data. It is assumed that missing data are estimated by appropriate classical methods or are coded (e.g., as NA) in the symbolic data framework.

The first method is the vertically apended array method. Here, it is necessary that each array has data on the same $p_t = p$, $t = 1, \ldots, T$, symbolic variables, but n_t need not be the same for each array $\boldsymbol{X}_t$. Also, for some units $i \in \Omega$, data may be missing at some times. The T vertically appended $\boldsymbol{X}_t$'s now represent the new data set $\boldsymbol{X}^*$ of size $n^* \times p$, where $n^* = \sum_{t=1}^{p} n_t$ is the new number of units $i \in \Omega^*$. If all original units $i \in \Omega = \{1, \ldots, n\}$ are recorded for all time points T, then clearly $n^* = nT$. The resulting two-way table can then be analysed using whatever method (factor analysis, etc.) is appropriate or available. For example, a symbolic cluster analysis may reveal that the units in Ω^* corresponding to $t = T$ (say) are markedly different from those at $t < T$, hence suggesting a change in temporal (or spatial, or expert, etc) pattern has occured.

The second method is the horizonally appended array method, where now it is necessary that $n_t = n$, $t = 1, \ldots, T$, but p_t can vary over t. In this case, the resulting data matrix $\boldsymbol{X}^+$ is an $n \times p^+$ matrix with $p^+ = \sum_{t=1}^{T} p_t$ symbolic variables. Then, for example, a global symbolic factor analysis can be performed, or an analysis can be undertaken in turn on each $\boldsymbol{X}_t$ data array with the other $(p^+ - p_t)$ variables being supplementary variables. The units in Ω can be followed over t by performing a symbolic multiple factorial analysis as developed by Escofier and Pages (1998). However, their method requires that for any t, each of the p_t symbolic variables are of the same type (i.e., all interval-valued, all multi-valued, etc.).

Blanchard and Gettler-Summa (1994) developed a data compression method applicable for quantitative variables when t was time. They compressed the data into $k << T$ data arrays each on an interval (of time) I_t, $t = 1, \ldots, k$, and where clearly the compressed data for each unit will now assume a modal value. Blanchard and Gettler-Summa obtained homogeneous intervals by using a contrained clustering technique, Fisher's algorithm or pyramidal clustering. Gettler-Summa and Pardoux (2000) note that for symbolic ojects linked by time, dissimilarity measures used in grouping need to account for temporal changes where they exist.

In a different direction, Ferraris et al. (1996) and Ferraris and Pardoux (1998) developed a Markov chain approach for qualitative multi-valued symbolic variables, as a way to compress the data. Clearly, other compression methods can be adoapted.

10 Probabilities in Symbolic Data Analysis

Probabilistic underpining for a few of these symbolic data analysis methods has been developed by Diday (1995), Diday and Emillion (1996) and Diday et al. (1996), primarily for

modal data associated with capacities and credibilities, and by Emillion (2002) for mixture decompositions which underlie much of the histogram methods.

Let us suppose that in general the data are arrayed in a table with n rows and p columns with entries Y_{ij}. The rows are the observations of a sample of size n from a random variable $\boldsymbol{Y} = (Y_1, \ldots, Y_p) : \Omega \to \mathcal{X}_1 \times \ldots \times \mathcal{X}_p$, where $\mathcal{X}_j$ is the set of random variables defined on the same domain, such that $Y(w_i) = (Y_{i1}, \ldots, Y_{ip})$ for object w_i. If instead of having this general model we focus on some descriptive property of the Y_{ij} (such as a confidence interval, empirical distribution density, etc.), we obtain different kinds of nonnumerical data (i.e., we obtain symbolic data). For instance, if we are interested in the probability distributions, the modal data become

$$\boldsymbol{Y} = (Y_1, \ldots, Y_p) : \Omega \to P(V_1) \times \ldots \times P(V_p)$$

where $P(V_j)$ is the set of all probability measures on a domain D_j with a σ-algebra $\mathcal{A}$. Hence, if $Y(w_i) = (\mu_{i1}, \ldots, \mu_{ip})$, the stochastic process $X_j(\cdot)(\mathcal{A})$ is a "random distribution" (by taking the terminology of some authors using Bayesian analyses; see, e.g., Ferguson, 1974).

By starting from this probabilistic framework, several approaches have been considered. For instance, suppose we are interested in the probability of an event A for an easily given random variable Y_j such that $P\{[Y_{1j} = A] \cup \ldots \cup [Y_{pj} = A]\} = F(A)$. It can be easily shown that F is a capacity. See Diday (1995) where in addition to capacity, other measures (including credibility, probability, etc.) have been studied and compared. By using special maps such as T-norms and t-conorms introduced by Schweizer and Sklar (1983), Diday describes a histogram of capacities and a histogram of credibilities which are subadditive and superadditive, respectively, in contrast to standard classical histograms of frequencies which are additive. Diday also proves the existence of limit of the height of these histograms as the interval lengths supporting the histogram tends to zero. It is therefore shown that the concepts intent and extent (described in Section 3), as well as Galois lattices (not described herein) of symbolic objects can be represented as (partial, or full) capacities or credibilities and hence by the relevant respective histograms. These results can also be applied to special cases of principal components (the vertices method), clustering (to pyramid clusters) and decision trees.

Bertrand and Goupil (2000) indicate that, for interval-valued symbolic data, using the law of large numbers, the true limiting distribution of Z as $m \to \infty$ is only approximated by the exact distribution $f(\xi)$ in (15) since the derivation depends on the veracity of the uniform distribution within each interval assumption.

Recently, clustering methods based on the estimation of mixtures of probability distribution where the observations are probability distributions, were considered by Emillion

(2002). Using a theorem of Kakutani (see, e.g., Hewitt and Stromberg, 1969, p. 453), Emillion proved a convergence result for the three cases of mixture components, viz., a Dirichlet process, a normalized gamma process, and a Kraft process. Another approach for mixture distributions suggested by Diday (2001) is based on copula theory (Schweizer and Sklar, 1983; Nelsen, 1999) by utilizing different copula models such as that of Frank (1979) and Clayton (1978). Bertrand (1995) considers structural properties of pyramids.

By and large however, probabilistic and rigorous mathematical underpining along with the need to develop relevant convergence and limiting results remain as outstanding problems requiring considerable development. This includes deriving relevant standard errors and distribution functions for the respective estimators thus far available, as well as consideration of robustness issues as to the effects of the underlying assumptions on these inference procedures. The field is wide open.

11 Conclusion

In conclusion then, given contemporary data formats and data set sizes, the need to develop statistical methods to analyse them is becoming increasingly important. In addition to these types of data, several attempts have been made to extend data analysis to fuzzy, uncertain, non-precise data (e.g., Bandemer and Nather, 1992; Viertl, 1996), or compositional data (e.g., Aitchison, 1986). In both cases by taking care of the variation inside a class of units described by such data, we obtain symbolic data. The statistical world has developed a wealth of methodology throughout the twentieth century, methods which are largely limited to small (by comparison) data sets and limited to classical data formats. There is some work on imprecise of fuzzy data but not on symbolic data. The present paper has reviewed briefly the formulation and construction of symbolic objects and classes. It then presented brief overviews of those symbolic data analyses which have emerged as a way to draw statistical inferences in some formats of symbolic data. What becomes abundantly clear, by the obvious omission of methodologies in any list of such statistical methods, is the enormous need for the development of a vast catalogue of new symbolic methodologies, along with rigorous mathematical and statistical foundations for these methods.

Acknowledgement:

As an addendum, a shortened version of this work is to appear in Billard and Diday (2003).

Partial support from the National Science Foundation USA and INRIA France is gratefully acknowledged.

REFERENCES

Aitchison, J. (1986), "*The Statistical Analysis of Compositional Data*", New York: Chapman Hall.

Aristole (IVBC; 1994). Des Categories De l'Interpretation. Oragon. *Librarie Philosophique Journal Vrin.*

Arnault, A. and Nicole, P. (1662), LaLogique ou l'art depensur" Reprinted by Froman, Stuttgart (1965).

Bandemer, H. and Nather, W. (1992), "*Fuzzy Data Analysis*", Kluwer Academic Publisher.

Bertrand, P. (1995). Structural Properties of Pyramidal Clustering. In: *Partitioning Data Sets* (eds. I. Corc, P. Hansen and B. Julesz), American Mathematical Society Series in Discrete Mathematics and Theoretical Computer Science, 19, 352-353.

Bertrand, P. and Goupil, F. (2000). Descriptive Statistics for Symbolic Data. In: *Analysis of Symbolic Data* (eds. H. -H. Bock and E. Diday), Springer, 103-124.

Billard, L., and Diday, E. (2000), "Regression Analysis for Interval-Valued Data", In: *Data Analysis, Classification, and Related Methods* (eds. H.A.L. Kiers, J. -P. Rassoon, P.J.F. Groenen and M. Schader), Berlin: Springer-Verlag, 369-374.

Billard, L., and Diday, E. (2002a), "Symbolic Regression Analysis," in *Classification, Clustering, and Data Analysis* (eds. K. Jajuga, A. Sokolowski and H. -H. bock), Berlin: Springer-Verlag, 281-288.

Billard, L., and Diday, E. (2002b), "From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association*, 98, 470-487.

Blanchard, J. L. and Gettler-Summa, M. (1994). Symbolic Approaches on Ergonomic Problems for Electric Centres. *Compstat.*

Bock, H. -H and Diday, E. (2000). Symbolic Objects. In: *Analysis of Symbolic Data* (eds. H. -H. Bock and E. Diday), Springer-Verlag, 54-77.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth.

Brito, P. (1994). Use of Pyramids in Symbolic Data Analysis. In: *New Approaches in Classification and Data Analysis* (eds. E. Diday, Y. Lechevallier and M. Schader, P. Bertrand and B. Burtschy), Springer-Verlag, 378-386.

Brito, P. (1995). Symbolic Objects: Order Structure and Pyramidal Clustering. *Annals of Operations Research*, 55, 277-297.

Brito, P. (2000). Hierarchical and Pyramidal Clustering with Complete Symbolic Objects. In: *Analysis of Symbolic Data* (eds., H. -H. Bock and E. Diday), Springer-Verlag, 312-324.

Brito, P. and DeCarvalho, F. A. T. (1998). Symbolic Clustering in the Presence of Hierarchical Rules. In: *Knowledge Extraction from Statistical Data.*

Cazes, P., Chouakria, A., Diday, E. and Schektman, Y. (1997). Extensions de l'analyse en composantes principales a des donnees de type intervalle. *Revue de Statistique Appliquee* 24, 5-24.

Chavent, M. (1998). A Monothetic Clustering Algorithm. *Pattern Recognition Letters*, 19, 989-996.

Choquet, G. (1954). *Theory of Capacities*, Annals Institute Fourier, 5, 131-295.

Chouakria, A. (1998). *Extension des methodes d'analyse factorielle a des donees de type intervalle*, Ph.D. Thesis, University of Paris.

Chouakria, A., Diday, E. and Cazes, P. (1998). An Improved Factorial Representation of Symbolic Objects. In: *Knowledge Extraction from Statistical Data*.

Clayton, D. G. (1978). A Model for Association in Bivariate Life Tables and its Applicaiton in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika*, 65, 141-151.

Codd, E. F. (1972). Further Normalization of the Data Base Relational Model. In: *Data Base Systems Courant Computer Science Series* (ed. R. Rustin) Vol. 6, Prentice-Hall, 33-64.

Csernel, M. (1997). Normalization of Symbolic Objects. In: *Symbolic Data Analysis and its Applications* (eds. E. Diday and K. C. Gowda), Ceremade, Dauphine, 32-44.

DeCarvalho, F. A. T. (1994). Proximity Coefficients Between Boolean Symbolic Objects. In: *New Approaches in Classification and Data Analysis* (eds. E. Diday, Y. Lechevallier, M. Schader, P. Bertrand and B. Burtschy), Springer-Verlag, 387-394.

DeCarvalho, F. A. T. (1995). Histograms in Symbolic Data Analysis, *Annals of Operations Research* 55, 299-322.

DeCarvalho, F. A. T. (1998). Extension Based Proximities Between Contrained Boolean Symbolic Objects. In: *Data Science, Classification and Related Methods* (eds. C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock and Y. Baba), Springer-Verlag, 370-378.

DeCarvalho, F. A. T., Verde, R. and Lechevallier, Y. (1999). A Dynamical Clustering of Symbolic Objects Based on a Content Dependent Proximity Measure. *Applied Statistical Models and Data Analysis*, 15, 237-242.

Diday, E. (1986). Orders and overlapping clusters by pyramids. In: *Multidimensional Data Analysis* (eds. J. De Leeuw, W. J. Heisen, J. J. Meulman and F. Critchley), DSWO Press, Leiden, 201-234.

Diday, E. (1987). Introduction a e'Approche Symbolique en Analyse des Donnees. *Premiere Jouneles Symbolique-Numerique*, CEREMADE, Universite Paris - Dauphine, 21-56.

Diday, E. (ed.) (1989). *Data Analysis, Learning Symbolic and Numerical Knowledge*, Nova Science, Antibes.

Diday, E. (1990). Knowledge Representation and Symbolic Data Analysis. In: *Knowledge Data and Computer Assisted Decisions* (eds. M. Schader and W. Gaul), Springer-Verlag, 17-34.

Diday, E. (1995). Probabilistic, Possibilistic and Belief Objects for Knowledge Analysis. *Annals of Operations Research* 55, 227-276.

Diday, E. and Emilion, R. (1996). Capacities and Credibilities in Analysis of Probabilistic Objects. In: *Ordinal and Symbolic Data analysis* (eds. E. Diday, Y. Lechevallier and O. Opitz), Springer-Verlag, 13-30.

Diday, E., Emilion, R. and Hillali, Y. (1996). Symbolic Data Analysis of Probabilistic Objects by Capacities and Credibilities. *Atti Della XXXVIII Riunione Scientifia*, 5-22.

Diday, E. (2001). A Generalization of the Mixture Decomposition Problem in the Symbolic Data Analysis Framework. Dauphine: Ceremade.

Elder, J. and Pregibon, D. (1996). A Statistical Perspective on Knowledge Discovery in Databases. In: *Advances in Knowledge Discovery and Data Mining* (eds. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy). AAAI Press, 83-113.

Emillion, R. (2002), "Clustering and Mixtures of Stochastic Processes," Dauphine: Ceremade.

Escofier, B. and Pages, J. (1998). *Analyses Factorielles Simples et Multiples* (3rd ed), Dunod Press.

Esposito, F., Malerba, D. and Semeraro, G. (1991). Flexible Matching of Noisy Structural Descriptions. In: *Conference on Artifical Intelligence* (eds. J. Mylopoulos and R. Reiter), Morgan Kaufman Publishers, 658-664.

Ferraris, J. and Pardoux, C. (1998). Comparison de Stratégies de Peche par Analyses Factorielles sur Tableaux D'échanges. *Biometrie et Halieutique*.

Ferraris, J., Gettler-Summa, M., Pardoux, C. and Tong, H. (1996). Knowledge Extraction Using Stochastic Matrices Application to Elaborate Fishing Strategies. In: *Ordinal and Symbolic Data Analysis* (eds. E. Diday, Y. Lechevallier and O. Opitz), Springer-Verlag, 103-112.

Frank, M. J. (1979), "On the Simultaneous Associativity of $F(x, y)$ and $x + y - F(x, y)$," *Aequationes Mathematics*, 19, 194-226.

Gettler-Summa, M. (1999). Approaches MGS (Marquage et Generalisations Symboliques) pour le suivi de typologies dans le temps. *Journees de Statistique*, 31, in press.

Gettler Summa, M. (2000). Marking and Generalization by Symbolic Objects in the Symbolic Official Data Analysis Software. In: *Data Analysis, Classification, and Related Methods* (eds. H. A. L. Kiers, J. -P Rasson, P. J. F. Groenen, and M. Schader), Springer-Verlag, 417-422.

Gettler-Summa, M. and Pardoux, C. (2000). Symbolical Approaches for Three-Way Data. In: *Analysis of Symbolic Data* (eds. H. -H. Bock and E. Diday), Springer-Verlag, 342-354.

Gowda, K. C. and Diday, E. (1991). Symbolic Clustering Using a New Dissimilarity Measure. *Pattern Recognition*, 24, 567-578.

Hand, D. J., Blunt, G., Kelly, M. G. and Adams, N. M. (2000). Data Mining for Fun and Profit. *Statistical Science*, 15, 111-131.

Hewitt, E., and Stromberg, K. R. (1969), *Real and Abstract Analysis: A Modern Treatment of the Theory of Functions of a Real Variable*, New York: Springer-Verlag.

Ichino, M. (1988). General Metrics for Mixed Features - The Cartesian Space Theory for Pattern Recognition. In: *Proceedings IEEE International Conference on Systems, Man and Cybernetics*, 1, 494-497.

Ichino, M. and Yaguchi, H. (1994). Generalized Minkowski Metrics for Mixed Feature Type Data Analysis. *IEEE Transactions Systems Man and Cybernetics*, 24, 698-708.

Loustaunau, D., Pardoux, C. and Gettler-Summa, M. (1997). Multidimensional Analysis for the Acquisition and Processing of Expert Knowledge: Concept of On-Site Telecommunication. In: *Multidimensional Data Analysis* (eds. K. Fernandez-Aguirre and A. Morineau), CISIA-CERESTA, 293-300.

Michalski, R. S., Diday, E. and Stepp, R. E. (1981). A Recent Advance in Data Analysis: Clustering Objects into Classes Characterized by Conjunctive Concepts. In: *Progress in Pattern Recognition* (eds. L. Kanal and A. Rosenfeld), North Holland, 33-56.

Michalski, R. S. and Stepp, R. E. (1984). Learning from Observation: Conceptual Clustering. In: *Machine Learning* (eds. R. S. Michalski, J. G. Carbonell, and T. M. Mitchell), Springer-Verlag, 331-363.

Morineau, A., Sammartino, A. -E., Gettler-Summa, M. and Pardoux, C. (1994). Analyses des Donnels et Modelisation des Series Temporelles. *Review Statistique Appliquee*, 42, 61-81.

Nagabhushan, P., Gowda, K. C. and Diday, E. (1995). Dimensionality Reduction of Symbolic Data. *Pattern Recognition Letters*, 16, 219-233.

Nelson, R. B. (1999). *An Introduction to Copulas*, Springer-Verlag.

Périnel, E. and Lechevallier, Y. (2000). Symbolic Discrimination Rules. In: *Analysis of Symbolic Data* (eds. H. -H. Bock and E. Diday), Springer-Verlag, 244-265.

Polaillon, G. (2000). Pyramidal Classification for Internal Data Using Galois Lattice Reduction. In: *Analysis of Symbolic Data* (eds. H. -H. Bock and E. Diday), Springer-Verlag, 324-340.

Quinlan, J. R. (1986). Introduction of Decision Trees. *Machine Learning*, 1, 81-106.

Raju, S. R. K. (1997). Symbolic Data Analysis in Cardiology. In: *Symbolic Data Analysis and its Applications* (eds. E. Diday and K. C. Gowda), CEREMADE, Paris, 245-249.

Schafer, G. (1976). *A Mathematical Theory of Evidence*, Princeton.

Schweizer, B. (1984). Distributions are the Numbers of the Future. In: *Proceedings The Mathematics of Fuzzy Systems Meeting* University of Naples, 137-149.

Schweizer, B. and Sklar, A. (1983). *Probabilistic Metric Spaces*, North-Holland, New York.

Siebes, A. (1998). KESO: Data Mining and Statistics. In: *Knowledge Extraction from Statistical Data*, 1-13.

Stéphan, V. Hébrail, G. and Lechevallier, Y. (2000). Generation of Symbolic Objects from Relational Databases. In: *Analysis of Symbolic Data* (eds. H. -H. Bock and E. Diday), Springer-Verlag, 78-105.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley.

Verde, R. and DeCarvalho, F. A. T. (1998). Dependence Rules Influence on Factorial Representation of Boolean Symbolic Objects. In: *Knowledge Extraction form Statistical Data*.

Viertl, R. (1996), "*Statistical Methods for Non-Precise Data*", Boca Raton: CRC Press.

Table 1a

i	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}	Y_{16}	Y_{17}	Y_{18}	Y_{19}	Y_{20}	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}	Y_{26}	Y_{27}	Y_{28}	Y_{29}	Y_{30}
1	Boston	M	24	W	S	2	2	0	165	68	120	79	183	83	86	2.21	88	92	16	1.4	12	21	6.9	5.2	14.2	43.6	2.32	N	N	0
2	Boston	M	56	W	M	1	2	2	186	84	130	90	164	64	60	2.55	69	101	16	0.8	20	22	7.0	4.6	13.5	39.9	2.44	N	N	1
3	Chicago	M	48	B	M	1	3	2	175	73	126	82	229	109	122	2.10	114	80	17	1.4	13	24	7.7	4.9	14.1	44.2	2.73	Y	N	1
4	El Paso	F	47	W	M	0	1	1	141	78	121	86	239	69	74	3.45	44	90	15	1.1	14	20	6.7	4.6	13.9	40.7	2.17	Y	0	0
5	Byron	F	79	B	W	0	3	4	152	84	150	88	187	67	64	2.79	72	103	18	0.9	20	27	7.3	4.8	11.6	36.1	3.05	N	0	0
6	Concord	M	12	W	S	2	1	0	73	69	126	85	109	98	107	1.11	105	108	14	0.8	18	17	6.2	4.3	12.2	36.0	1.79	N	0	0
7	Atlanta	F	67	W	W	1	6	0	166	81	134	89	190	90	96	2.12	95	91	17	1.0	17	24	7.2	4.6	13.4	42.3	2.65	Y	6	0
8	Boston	F	73	O	M	0	2	4	164	77	121	81	181	81	84	2.24	86	112	19	0.9	22	29	8.0	4.0	14.9	43.6	3.32	N	0	0
9	Lindfield	M	29	B	M	2	0	2	227	62	124	81	214	94	101	2.28	99	89	18	1.0	18	27	7.8	4.7	15.0	43.4	3.13	N	0	0
10	Lindfield	M	44	B	M	1	3	3	216	71	125	79	218	98	107	2.23	103	83	18	1.0	18	27	7.8	4.5	12.4	37.1	3.12	Y	N	2
11	Boston	M	54	B	S	1	5	0	213	57	118	88	189	69	66	2.75	74	100	28	0.3	90	53	11.5	4.3	14.8	42.6	6.32	N	N	0
12	Chicago	F	12	W	S	2	2	0	75	69	115	81	153	54	45	2.83	58	119	20	1.0	19	31	8.3	4.4	14.3	40.7	3.59	N	0	0
13	Macon	F	73	W	M	0	3	1	152	58	123	82	188	87	93	2.15	93	69	16	1.2	13	21	6.9	4.6	12.9	37.1	2.35	N	0	0
14	Boston	M	48	B	M	0	2	4	206	73	113	72	264	72	62	3.69	49	91	14	1.2	11	16	6.1	5.0	12.9	40.5	1.67	N	N	0
15	Peoria	F	79	O	M	0	3	3	153	72	106	78	118	40	35	2.95	23	82	19	0.9	20	30	8.1	4.1	13.6	43.3	3.40	N	0	0
16	Concord	M	20	B	S	2	0	1	268	79	123	80	205	85	89	2.40	90	71	19	1.3	14	28	7.9	4.2	13.5	39.4	3.21	N	N	0
17	Boston	F	20	B	S	2	4	0	157	75	116	87	180	60	52	3.01	65	101	17	1.0	16	23	7.2	5.1	13.0	40.8	2.61	N	0	0
18	Chicago	M	17	B	S	2	2	0	161	69	114	78	169	49	39	3.45	54	96	17	1.0	16	23	7.2	4.2	13.1	40.7	2.61	N	N	0
19	Stowe	M	31	B	M	1	3	2	183	81	118	84	185	66	62	2.82	71	146	18	0.7	24	28	7.8	4.8	13.2	38.2	3.14	N	N	0
20	Tara	M	83	W	M	0	3	1	128	80	108	80	224	48	65	4.66	38	111	15	1.0	14	18	6.4	4.7	13.6	41.7	1.94	Y	N	3
21	Akron	M	20	W	S	1	3	0	182	68	114	76	150	51	40	2.94	55	58	13	1.2	11	14	5.8	4.0	13.7	40.7	1.43	N	N	0
22	Detroit	F	85	W	M	0	3	2	161	73	122	76	185	83	89	2.19	90	96	8	0.9	10	17	3.9	4.2	13.1	36.8	7.19	N	0	0
23	Ila	F	66	B	S	0	4	3	166	66	126	87	218	98	108	2.22	103	85	18	1.4	13	26	7.6	4.2	13.4	38.0	2.98	N	4	0
24	Marion	M	6	W	S	2	1	0	35	72	114	76	136	52	28	2.60	41	96	16	1.2	13	20	6.8	4.5	15.4	45.2	2.25	N	N	0
25	Albany	M	24	W	M	2	1	1	177	81	111	82	149	51	39	2.96	55	72	19	0.8	24	30	8.2	3.8	14.6	45.4	3.48	N	N	0
26	Salem	M	76	B	M	0	5	2	192	77	115	73	173	53	44	3.27	58	97	17	1.3	13	23	7.2	4.6	12.4	37.1	2.60	N	N	0
27	Quincy	M	57	O	S	1	3	2	159	72	114	75	234	131	157	1.78	139	88	17	0.8	22	24	7.2	5.0	13.0	37.5	2.65	N	N	0
28	Yuma	F	11	W	S	2	2	0	73	62	118	80	96	56	43	1.71	56	136	20	0.8	25	31	7.4	4.6	13.6	41.0	3.59	N	0	0
29	Fargo	F	27	W	M	2	2	1	124	70	114	72	167	67	63	2.49	72	104	13	0.7	20	15	5.9	4.7	13.7	41.4	1.53	N	0	0
30	Reno	F	43	B	M	2	4	4	148	66	135	97	172	52	43	3.31	57	82	17	0.8	21	25	8.3	4.2	11.7	34.0	2.83	Y	3	0
31	Amherst	F	53	W	S	1	0	3	165	65	135	96	236	134	161	1.76	141	102	9	1.0	9	4	7.0	4.6	11.2	33.9	2.18	N	1	0
32	Boston	M	14	W	S	2	1	0	132	66	125	87	149	51	39	2.96	54	120	20	1.1	18	32	8.4	5.3	12.1	38.7	3.67	N	N	0
33	New Haven	F	29	B	M	1	0	1	153	70	133	92	217	97	106	2.23	102	99	25	1.1	23	45	10.3	4.6	14.3	41.4	5.29	N	0	0
34	Kent	M	84	W	M	0	4	1	239	85	114	75	229	126	150	1.81	134	113	9	1.1	8	4	4.3	4.8	12.3	37.6	2.43	Y	N	5
35	Shelton	M	52	W	M	0	4	1	206	63	125	86	236	134	161	1.77	14	114	16	1.2	13	22	4.4	4.4	12.9	37.0	0.15	N	N	0
36	Atlanta	M	86	O	M	0	3	3	184	72	114	72	152	53	42	2.88	57	92	18	1.2	15	27	6.6	4.2	15.7	49.9	3.11	N	N	0
37	Medford	F	23	W	S	2	1	0	138	71	125	85	197	96	105	2.05	102	70	20	1.4	15	33	8.6	4.1	14.4	40.7	3.78	N	0	0
38	Banjor	M	51	W	M	2	2	2	172	81	119	78	172	73	71	2.38	77	105	17	1.3	13	23	7.2	4.8	13.5	39.3	2.60	N	N	0
39	Boston	M	70	W	M	1	6	3	183	75	114	74	151	52	42	2.90	56	79	14	0.7	21	17	6.3	5.2	12.3	36.5	1.86	Y	N	2
40	Barre	M	65	W	M	0	4	2	191	84	120	80	175	75	75	2.34	80	139	21	1.3	16	34	8.8	4.5	14.6	43.8	3.97	N	N	0
41	Trenton	M	82	W	M	0	3	4	201	79	123	84	188	87	93	2.15	93	111	18	0.8	21	26	7.5	4.6	14.5	44.6	2.91	N	N	0
42	Concord	M	60	W	S	0	4	0	175	74	117	76	163	63	58	2.58	68	112	12	1.2	9	10	5.2	4.7	13.4	38.7	0.91	N	N	0
43	Chicago	M	48	W	W	1	4	1	187	88	132	98	182	82	86	2.23	87	95	18	0.8	23	27	7.8	4.3	12.4	37.3	3.13	N	N	0
44	Omaha	M	29	W	M	1	1	1	166	59	122	82	178	78	79	2.29	83	77	22	1.5	15	38	9.3	4.3	13.5	38.9	4.43	N	N	0
45	Tampa	F	21	W	M	2	2	1	124	72	119	79	169	70	67	2.43	74	103	14	1.0	14	15	6.0	4.9	13.5	43.1	1.60	N	0	0
46	Lynn	F	81	W	M	0	5	3	161	79	128	89	210	109	124	1.93	115	86	15	0.6	23	19	7.8	4.5	11.3	32.9	2.06	Y	3	0
47	Quincy	R	70	B	M	0	3	2	178	72	119	78	230	110	124	2.09	115	87	12	1.2	10	10	5.3	4.4	12.5	39.1	8.88	Y	1	1
48	Wells	F	27	W	M	2	0	0	113	77	121	80	179	79	80	2.27	84	101	6	1.0	6	8	3.1	4.2	14.6	40.9	2.99	N	0	0
49	Buffalo	F	36	W	M	2	4	1	129	76	119	81	172	72	71	2.38	77	99	14	1.4	10	16	6.1	4.7	13.9	40.3	1.69	N	0	0
50	Quincy	M	64	W	S	1	2	0	194	81	128	89	210	109	124	1.93	115	88	19	1.1	17	28	7.9	5.0	15.3	49.1	3.23	N	N	0
51	Boston	F	87	W	W	0	5	2	157	88	128	88	171	71	70	2.40	76	87	19	1.0	19	28	7.9	4.6	13.9	44.4	7.14	N	0	0
⋮																														

+ See Table 1b for description of Y_i variables.

Table 1b Variable Identifications

Y_i	Description: Range
Y_1	City/Town of Residence
Y_2	Gender: Male (M), Female (F)
Y_3	Age (in years): ≥ 0
Y_4	Race: White (W), Afro-American (B), Other (O)
Y_5	Marital Status: Single (S), Married (M)
Y_6	Number of Parents Alive: 0, 1, 2
Y_7	Number of Siblings: 0, 1, ...
Y_8	Number of Children: 0, 1, ...
Y_9	Weight (in pounds): > 0
Y_{10}	Pulse Rate: > 0
Y_{11}	Systolic Blood Pressure: > 0
Y_{12}	Diastolic Blood Pressure: > 0
Y_{13}	Cholesterol Total: > 0
Y_{14}	HDL Cholesterol Level: > 0
Y_{15}	LDL Cholesterol Level: > 0
Y_{16}	Ratio = Cholesterol Total/HDL Level: > 0
Y_{17}	Triglyceride Level: > 0
Y_{18}	Glucose Level: > 0
Y_{19}	Urea Level: > 0
Y_{20}	Creatinine Level: > 0
Y_{21}	Ratio = Urea/Creatinine: > 0
Y_{22}	ALT Level: > 0
Y_{23}	White Blood Cell Measure: > 0
Y_{24}	Red Blood Cell Measure: > 0
Y_{25}	Hemoglobin Level: > 0
Y_{26}	Hemocrit Level: > 0
Y_{27}	Thyroid TSH: > 0
Y_{28}	Cancer Diagnosed: Yes (Y), No (N)
Y_{29}	Breast Cancer # Treaatments: 0, 1, ..., Not applicable (N)
Y_{30}	Lung Cancer # Treatments: 0, 1, ...

Table 2

u	Age	Blood Pressure	City	Type of Cancer	Gender
1	[20, 30)	(79, 120)	Boston	{Brain tumor}	{Male}
2	[50, 60)	(90, 130)	Boston	{Lung, Liver}	{Male}
3	[45, 55)	(80, 130)	Chicago	{Prostate}	{Male}
4	[47, 47)	(86, 121)	El Paso	{Breast p, Lung $(1-p)$}	{Female}
⋮	⋮	⋮	⋮	⋮	⋮

Table 3

u	Y_1	Y_2	$vir(d_u)$	$\lvert vir(d_u)\rvert$	$\lvert C_u\rvert$
1	{0,1}	{2}	{(1,2)}	1	128
2	{0,1}	{0,1}	{(0,0), (1,0), (1,1)}	3	75
3	{0,1}	{3}	{(1,3)}	1	249
4	{0,1}	{2,3}	{(1,2), (1,3)}	2	113
5	{0}	{1}	ϕ	0	2
6	{0}	{0,1}	{(0,0)}	1	204
7	{1}	{2,3}	{(1,2), (1,3)}	2	87
8	{1}	{1,2}	{(1,1), (1,2)}	2	23
9	{1}	{1,3}	{(1,1), (1,3)}	2	121

Table 4

u	Y_1 Pulse Rate	Y_2 Systolic Pressure	Y_3 Diastolic Pressure
1	[44-68]	[90-110]	[50-70]
2	[60-72]	[90-130]	[70-90]
3	[56-90]	[140-180]	[90-100]
4	[70-112]	[110-142]	[80-108]
5	[54-72]	[90-100]	[50-70]
6	[70-100]	[134-142]	[80-110]
7	[72-100]	[130-160]	[76-90]
8	[76-98]	[110-190]	[70-110]
9	[86-96]	[138-180]	[90-110]
10	[86-100]	[110-150]	[78-100]
11	[63-75]	[60-100]	[140-150]

Table 5

Y_1: Pulse Rate		Y_2: Systolic Pressure		Y_3: Diastolic Pressure	
I_g	p_g	I_g	p_g	I_g	p_g
$[40, 50)$	.0250	$[90, 100)$	.1750	$[50, 60)$	.1000
$[50, 60)$	.0868	$[100, 110)$	.0750	$[60, 70)$	.1000
$[60, 70)$	.2016	$[110, 120)$	.0938	$[70, 80)$	.1127
$[70, 80)$	.1611	$[120, 130)$	.0938	$[80, 90)$	.2609
$[80, 90)$	.2363	$[130, 140)$	.1818	$[90, 100)$	.2895
$[90, 100)$	.2606	$[140, 150)$	.1509	$[100, 110)$	.1369
$[100, 110)$	.0238	$[150, 160)$	.0946		
$[110, 120)$	.0048	$[160, 170)$	.0613		
		$[170, 180)$	.0613		
		$[180, 190)$	.0125		
Mean	$\bar{Y}_1 = 79.1$	$\bar{Y}_2 = 131.3$		$\bar{Y}_3 = 84.6$	
Variance	$S_1^2 = 162.29$	$S_2^2 = 495.41$		$S_3^2 = 182.44$	
Covariance	$S_{12} = 194.170$	$S_{23} = 257.920$		$S_{13} = 141.040$	
Correlation	$r(Y_1, Y_2) = .685$	$r(Y_2, Y_3) = .858$		$r(Y_1, Y_3) = .820$	

Table 6

	Y_1				Y_2	
u	Gas	Solid	Elec.	Other	No	Yes
1	{.87	.07	.05	.01}	{.09	.91}
2	{.71	.11	.10	.08}	{.12	.88}
3	{.83	.08	.09	.00}	{.23	.77}
4	{.76	.06	.11	.07}	{.19	.81}
5	{.78	.06	.09	.07}	{.12	.88}
6	{.90	.01	.08	.01}	{.22	.78}
7	{.87	.01	.10	.02}	{.22	.78}
8	{.78	.02	.13	.07}	{.13	.87}
9	{.91	.00	.09	.00}	{.24	.76}
10	{.73	.08	.11	.08}	{.14	.86}
11	{.59	.07	.17	.17}	{.10	.90}
12	{.90	.01	.08	.01}	{.19	.71}
13	{.84	.00	.14	.02}	{.09	.91}
14	{.82	.00	.11	.07}	{.17	.83}
15	{.88	.00	.09	.03}	{.12	.88}
16	{.85	.01	.10	.04}	{.09	.91}
17	{.71	.03	.17	.09}	{.16	.84}
18	{.87	.09	.04	.00}	{.13	.87}
19	{.32	.24	.24	.20}	{.25	.75}
20	{.50	.12	.28	.10}	{.14	.86}
21	{.69	.13	.18	.00}	{.12	.88}
22	{.79	.01	.20	.00}	{.21	.79}
23	{.72	.05	.19	.04}	{.07	.93}
24	{.43	.00	.43	.14}	{.28	.72}
25	{.00	.41	.14	.45}	{.09	.91}

Table 7
Apparent Relative Frequencies for $u = 25^*$

$Y_2 \backslash Y_1$	ξ_0	ξ_1	ξ_2	ξ_3	ξ_4	Total	Total $\|v(x) = 1$
No	.0225	$.0000^+$	$.0369^+$	$.0126^+$	$.0180^+$	.0900	.0319
Yes	$.2275^+$	.0000	.3731	.1274	.1820	.9100	.9681
Total	.2500	.0000	.4100	.1400	.2000		
Total $\|v(x) = 1$	.0319	.0000	.5292	.1807	.2582		

$^+$ Invalidated when rule v invoked

Table 8
Histogram for Weight by Age-Groups

Age	u	
20s	1	$\{[70-84), .02; [84-96), .06; [96-108), .24; [108-120), .30;$ $[120-132), .24; [132-144), .06; [144-160), .08\}$
30s	2	$\{[100-108), .02; [108-116), .06; [116-124), .40; [124-132), .24;$ $[132-140), .24; [140-150), .04\}$
40s	3	$\{[110-125), .04; [125-135), .14; [135-145), .20; [145-155), .42;$ $[155-165), .14; [165-175), .04; [175-185), .02\}$
50s	4	$\{[100-114), .04; [114-126), .06; [126-138), .20; [138-150), .26;$ $[150-162), .28; [162-174), .12; [174-190), .04\}$
60s	5	$\{[125-136), .04; [136-144), .14; [144-152), .38; [152-160), .22;$ $[160-168), .16; [168-180), .06\}$
70s	6	$\{[135-144), .04; [144-150), .06; [150-156), .24; [156-162), .26;$ $[162-168), .22; [168-174), .14; [174-180), .04\}$
80s	7	$\{(100-120), .02; [120-135), .12; [135-150), .16; [150-165), .24;$ $[165-180), .32; [180-195), .10; [195-210), .04\}$

Table 9
Histogram for Weights (over all ages)

g	I_g	Observed Frequency	Relative Frequency	(d)
1	[60 - 75)	0.00714	0.0010	–
2	[75 - 90)	0.04286	0.0061	–
3	[90 - 105)	0.24179	0.0345	0.0215
4	[105 - 120)	0.72488	0.1036	0.1133
5	[120 - 135)	1.27470	0.1821	0.1890
6	[135 - 150)	1.67364	0.2391	0.2411
7	[150 - 165)	1.97500	0.2821	0.2835
8	[165 - 180)	0.88500	0.1264	0.1264
9	[180 - 195)	0.13500	0.0193	0.0193
10	[195 - 210]	0.04000	0.0057	0.0057

Table 10
Bivariate Histogram for Blood Pressure Data

		$p_{g_2g_3} \times 10^2$						
	g_3	1	2	3	4	5	6	
g_2	$I_{g_2}I_{g_3}$	$[50, 60)$	$[60, 70)$	$[70, 80)$	$[80, 90)$	$[90, 100)$	$[100, 110)$	$p_{g_2} \times 10^2$
1	$[90, 100)$	7.500	7.500	1.250	1.250	0.000	0.000	17.500
2	$[100, 110)$	2.500	2.500	1.250	1.250	0.000	0.000	7.500
3	$[110, 120)$	0.000	0.000	1.790	3.815	2.565	1.205	9.375
4	$[120, 130)$	0.000	0.000	1.790	3.815	2.565	1.205	9.375
5	$[130, 140)$	0.000	0.000	1.492	7.446	5.303	3.943	18.185
6	$[140, 150)$	0.000	0.000	1.492	4.886	6.196	2.515	15.089
7	$[150, 160)$	0.000	0.000	1.265	2.693	4.003	1.503	9.464
8	$[160, 170)$	0.000	0.000	0.313	0.313	4.003	1.503	6.131
9	$[170, 180)$	0.000	0.000	0.313	0.313	4.003	1.503	6.131
10	$[180, 190)$	0.000	0.000	0.313	0.313	0.313	0.313	1.250
	$p_{g_3} \times 10^2$	10.000	10.000	11.266	26.093	28.950	13.690	

Table 11
Regional Profiles: Demographics

Region	MedHouse$	MeanPerson$	Health	EducHigh	EducBach	TravelTime
NewEng	[45.72, 47.74]	23.58, 24.84]	[88.50, 94.10]	[81.30, 90.00]	[24.10, 31.60]	[18.00, 21.90]
MidAtl	[43.69, 45.16]	[22.66, 23.37]	[84.80, 92.40]	[77.20, 87.30]	[15.30, 38.30]	[20.00, 28.60]
MidWestE	[44.03, 45.23]	[21.22, 21.86]	[86.50, 92.90]	[84.60, 90.80]	[17.10, 31.20]	[18.30, 25.10]
MidWestW	[43.85, 45.52]	[22.29, 23.47]	[88.50, 91.30]	[88.10, 90.40]	[24.60, 27.30]	[15.80, 17.20]
SouthAtl	[40.68, 41.60]	22.34, 23.07]	[82.70, 88.10]	[79.20, 84.00]	[19.00, 23.20]	[19.80, 22.70]
SouthE	[33.22, 35.14]	[18.74, 19.89]	[86.50, 89.70]	[77.50, 79.90]	[20.40, 22.00]	[20.70, 21.50]
SouthW	[35.98, 36.95]	[19.03, 19.82]	[78.50, 89.40]	[79.20, 86.60]	[18.40, 26.20]	[19.00, 22.30]
Mountain	[41.92, 43.41]	[20.21, 21.25]	[81.50, 88.70]	[85.50, 91.80]	[20.00, 34.60]	[13.00, 20.70]
Pacific	[45.22, 46.56]	[22.38, 23.21]	[76.20, 89.90]	[81.20, 91.80]	[19.30, 28.60]	[16.70, 24.60]

Table 12

u	Y_1 Cancer Prognosis	Y_2 Weight
1	{Medium (.06), Good (.4)}	{below average, average}
2	{Poor (.2), Medium (.6), Good (.2)}	{below average}
3	{Good (1)}	{Average}

Table 13

		Y_1		Below	Y_2	Above	
f_{uj}	Poor	Medium	Good	Average	Average	Average	$f_{u.}$
	0	.6	.4	.5	.5	0	2
	.2	.6	.2	1	0	0	2
	0	0	1	0	1	0	2
$f_{.j}$	.2	1.2	1.8	1.5	1.5	0	

Figure 1(a)

Histogram of Flying/Nonflying Birds

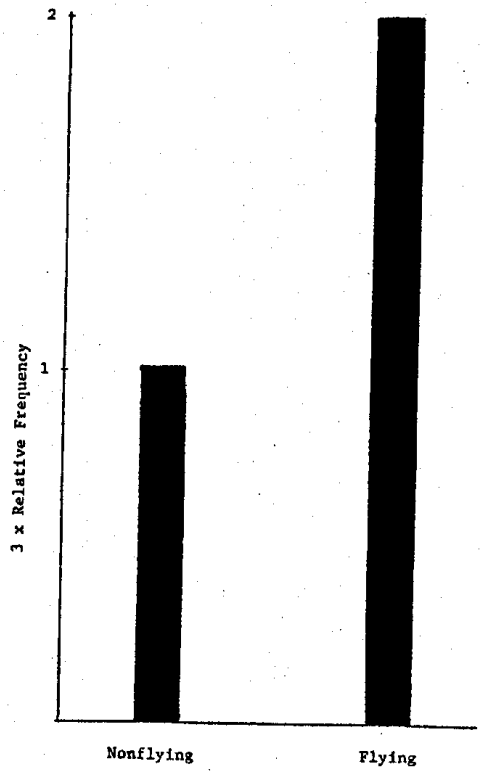


Figure 1(b)

Histogram of Flying/Nonflying Species

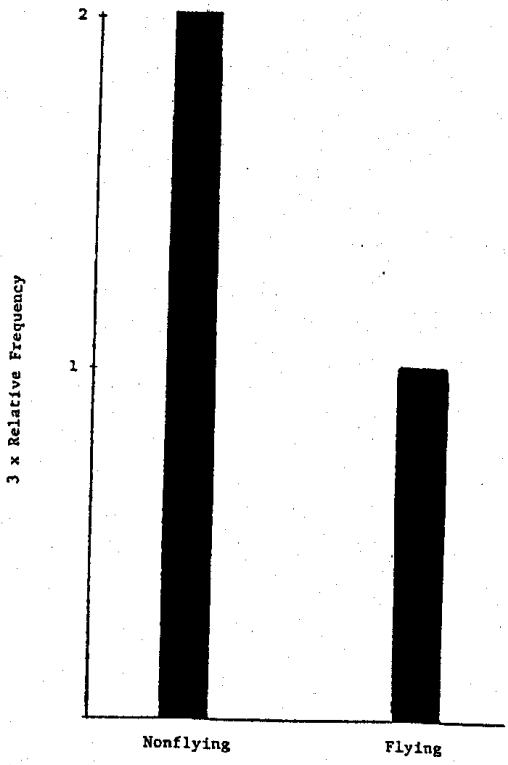


Figure 1(a)

Histogram of Flying/Nonflying Birds

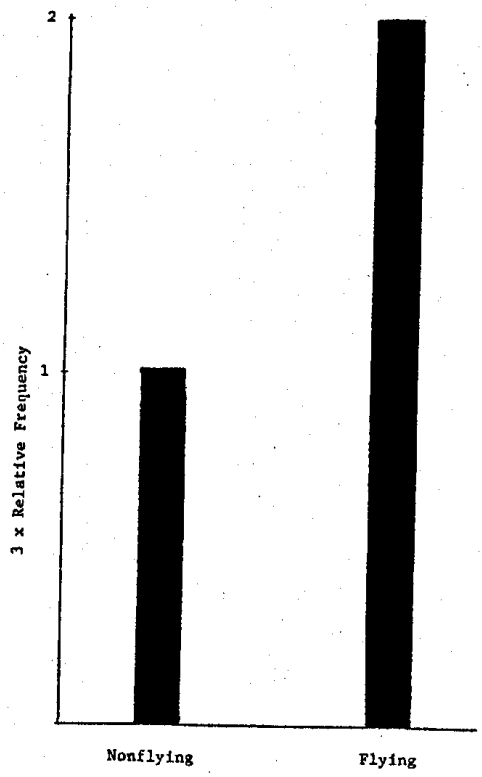


Figure 1(b)

Histogram of Flying/Nonflying Species

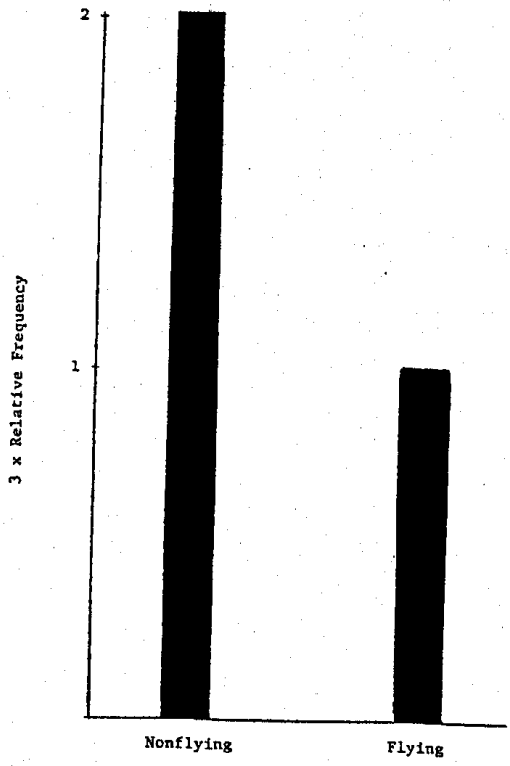
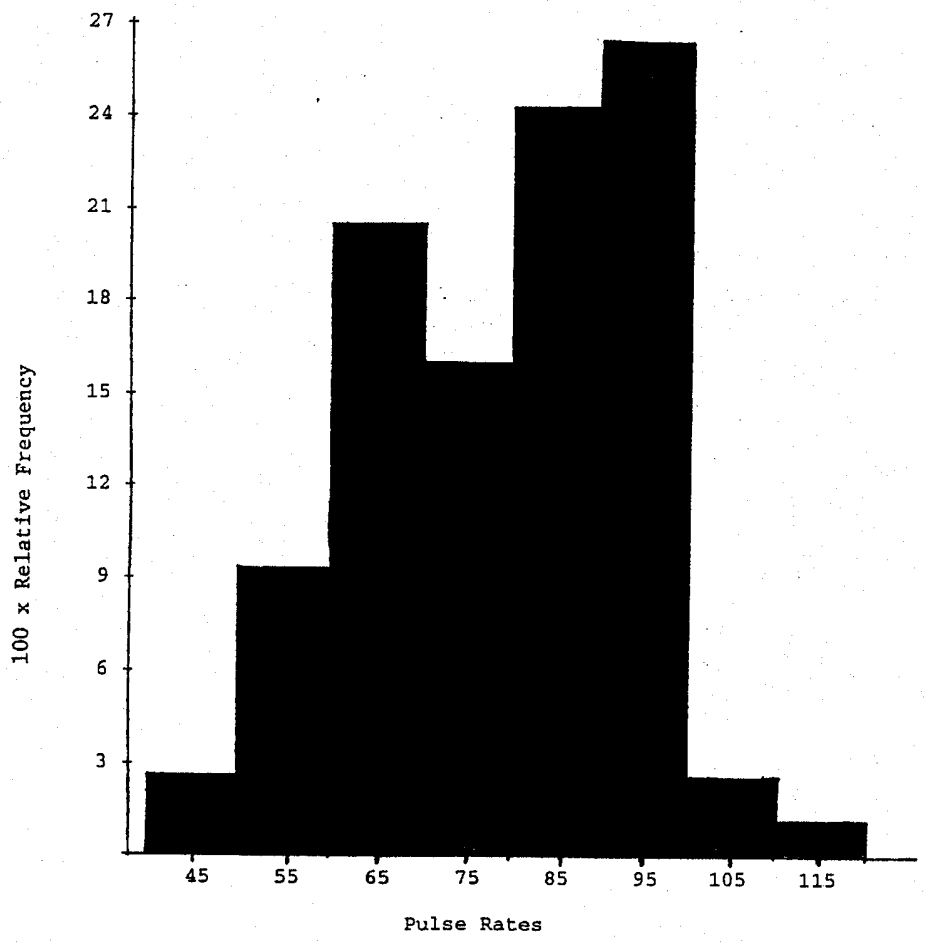


Figure 2
Histogram of Pulse Rates



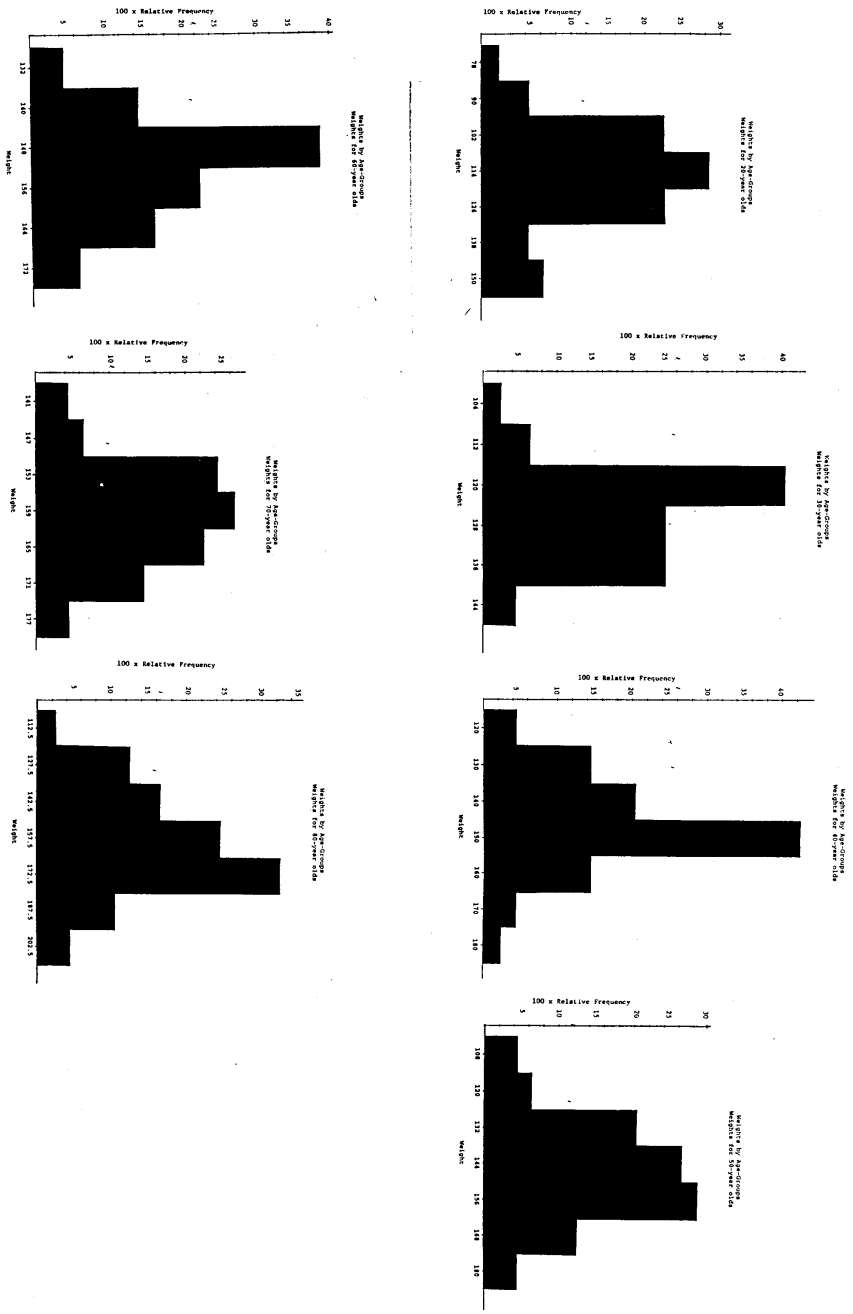

100 x Relative Frequency
Weight
Weights by age-group
Weights for 20-year olds
Weights for 30-year olds
Weights for 40-year olds
Weights for 50-year olds
Weights for 60-year olds
Weights for 70-year olds

FIGURE 3

Figure 4
Histogram of Histograms of Weights

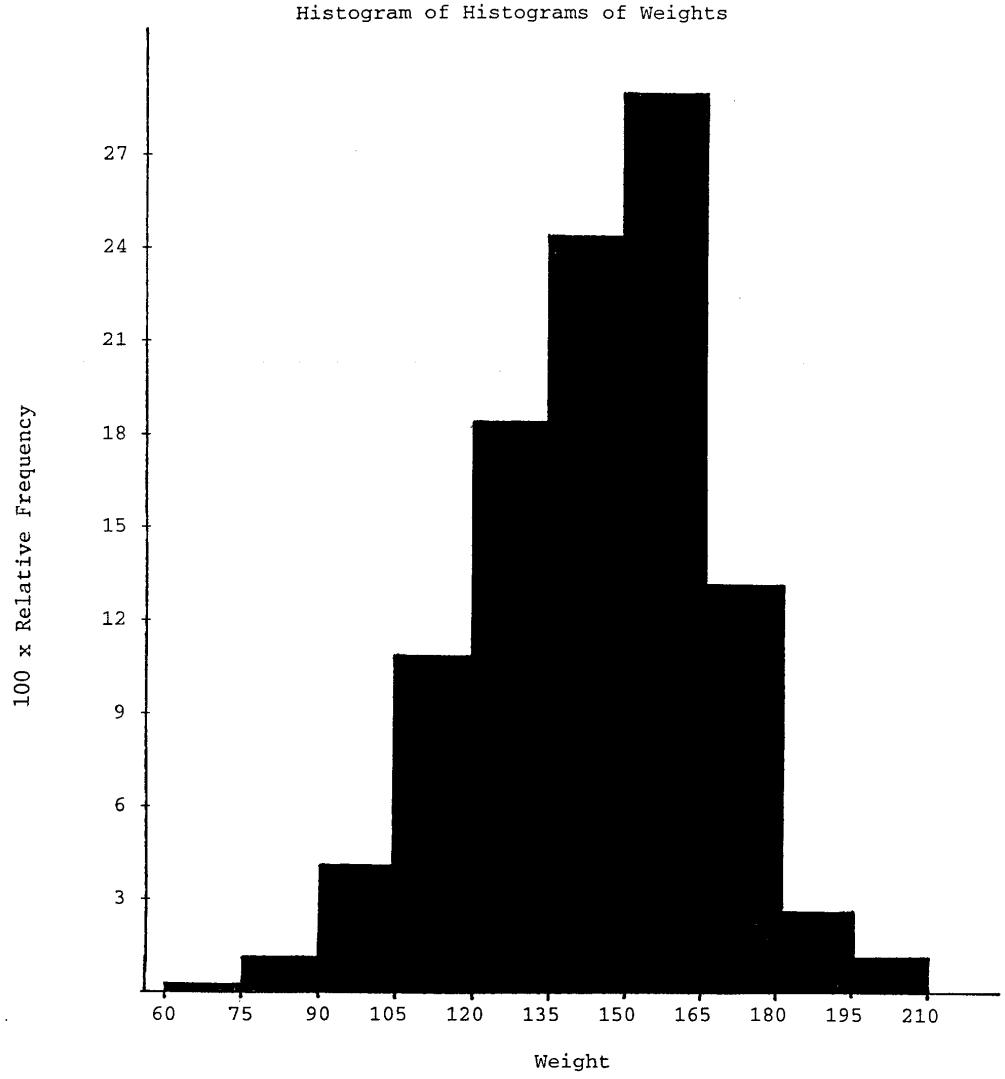


FIGURE 5
Bivariate Histogram: Blood Pressures

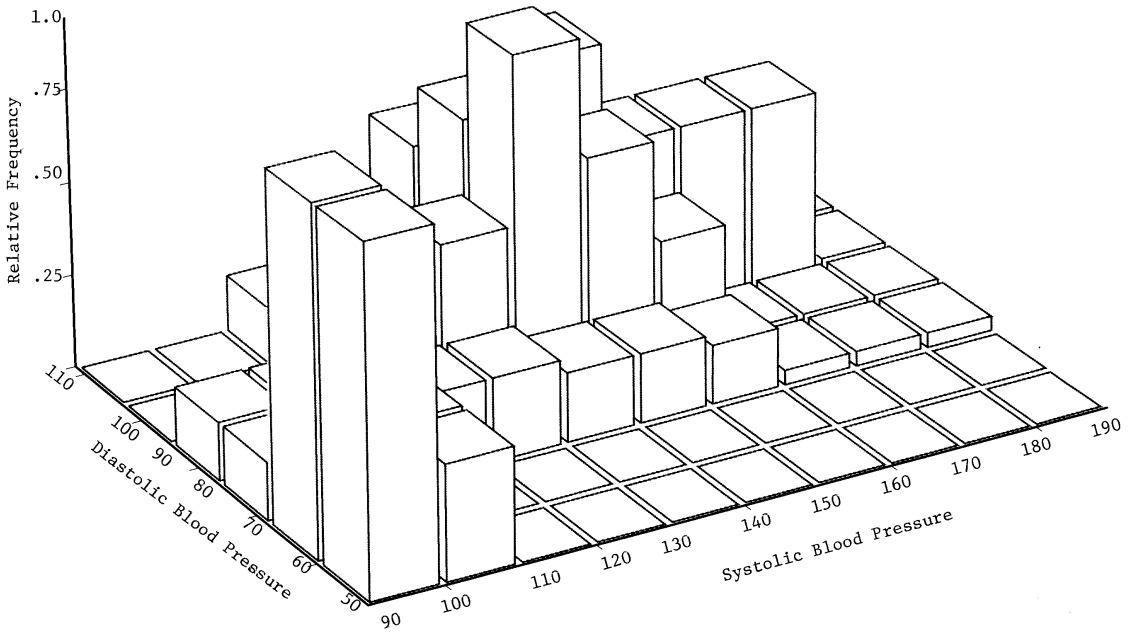


Figure 6
Principal Component Analysis: Regional Profiles

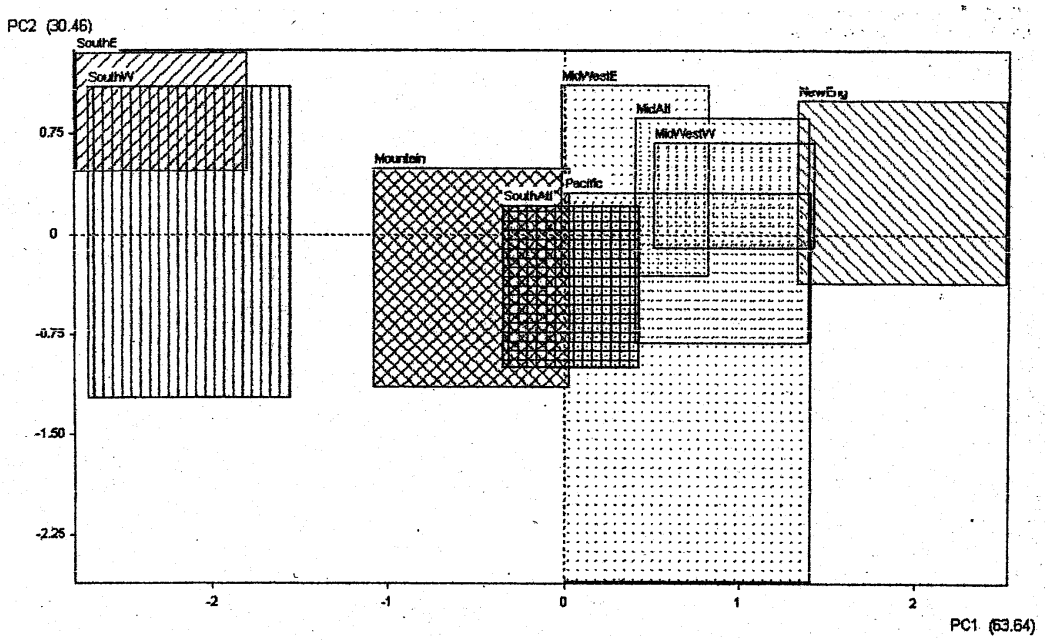


Figure 7
Divisive Clustering: Regional Profiles