Ann. Ist. Super. Sanità
Vol. 27, N. 3 (1991), pp. 377-384

STATISTICAL TOOLS IN THE CLINICAL LABORATORY

A. CHIECCHIO and A. BO

Servizio di Fisica Sanitaria, Ospedale Mauriziano, Torino

Summary. - *Method evaluation, control of data and transformation of laboratory results into diagnoses all involve a decision step. A survey of the statistical tools available to organize the information and check the congruity of decision making is provided is focused on: (a) the use of classical statistical tools (including computer based simulation and replication techniques) which enable theoretical distributions to be obtained and their optimal limits to be defined for classification purposes; (b) the analysis of multivariate distributions, which evidences the relationships among the variables involved, whatever they might be: e.g. results obtained on the same specimens with different methods (in test evaluation), different laboratory data related to the same pathophysiological situations (in making diagnoses), etc. As for the latter, the most common techniques of statistical analysis of data (discriminant and cluster analysis, principal components analysis) are also illustrated by general examples.*

KEY WORDS: univariate and multivariate distributions, statistical sampling, statistical analysis.

Riassunto (La strumentazione statistica nel laboratorio). - *La valutazione dei metodi, il controllo dei dati e la loro trasformazione in diagnosi comportano senza eccezioni un passaggio decisionale. L'organizzazione dell'informazione e la verifica della congruità della decisione non sono possibili senza far ricorso ad un'opportuna strumentazione statistica. Scopo dell'articolo è presentare una rassegna di questa strumentazione, con particolare riferimento a: l'uso dei classici strumenti statistici (comprese le tecniche di simulazione e replicazione al calcolatore) che consentono di ottenere le distribuzioni teoriche dei dati, definendone i limiti ottimali ai fini della classificazione; l'analisi delle distribuzioni a più variabili, che evidenzia le relazioni tra tutte le variabili in gioco, intese o come i risultati ottenuti per gli stessi campioni con metodi differenti (nella valutazione metodologica), o come i diversi dati di laboratorio riferiti alle stesse situazioni fisiopatologiche (nel processo diagnostico), o altro ancora. Per quanto riguarda l'ultimo punto, viene esemplificato l'uso delle più comuni tecniche statistiche di analisi, quali l'analisi discriminante, l'analisi dei "cluster" e l'analisi delle componenti principali.*

PAROLE CHIAVE: distribuzione univariata e multivariata, campionamento statistico, analisi statistica.

Introduction

A survey of statistical tools for diagnosis is presented. The term "diagnosis" is intended here in a wide sense, namely, the end-point of a decisional path associated with the evaluation of a method, the test of the data provided by the evaluated method and the allocation of tested data into a diagnostic protocol. Moreover, diagnosis is the "judgment" ending an analysis, e.g. a statement about patient's state of health, quality of an industrial or scientific product, and related data. The guidelines followed here are: the exhaustive utilization of the numerical information possessed by the operator, the use of statistics in terms of computer science [1] (so that the operator can find answers not only in probability tables but also at a computer terminal) and the reduction to a common cultural and technical denominator of topics similar with regard to language, approach and method.

This is applied to two problems: firstly, statistics of univariate distributions and, secondly, statistics of multivariate distributions.

Statistics of univariate distributions

Definition

In decisional problems, where proper classification (*i.e.* a diagnosis with a low level of uncertainty) is needed, the knowledge of the characteristics of the data distribution is critical.

Two descriptions of distributions can be provided, namely: (a) a mathematical description, when the knowledge of the distribution is related to the knowledge of the probability density function and (b) a statistical description, when the distribution is known if a "sufficiently" large number of outcomes of the random variable under study is possesed [2]. Fig. 1 shows two distributions with known mathematical shape. These distributions might be associated with two different pathologies or with the replicates of measurements of a single sample as obtained by two different methods.

Materials

The working material consists of the numerical data related to a single variable (supposed as continuous for sake of simplicity).

Only one characteristic of the data, *i.e.* their number, will be taken into account, while their properties will be the basis of knowledge concerning the situation under study [3]. "Knowledge" is intended as the process of acquiring information with a degree of uncertainty, which depends on each particular situation. Fig. 1 shows the histograms of data of two samples extracted from the theoretical distributions in the same illustration.

Theoretical and empirical models

The simplest answer to the demand for knowledge about distributions is the hypothesis of a model. Very often the gaussian model is assumed, since it actually fits most statistical variables [4] and good tests exist to check the hypothesis underlying this model [2]. Nevertheless, when this hypothesis is rejected, other more appropriate models are not always taken into account, even when they do not require any theoretical assumptions (*i.e.* empirical models [2, 3]).

Empirical models are based on transformations reducing the values of the original variable to percentiles of a standard gaussian variate, *i.e.* gaussian with zero mean and unitary standard deviation (SD). (Note: the αth percentile is the value that as a proportion α percent of sample distribution below it). The histograms of Fig. 1, fitted according to two models, are shown in Fig. 2.

The gaussian sum

Let every sample value be the most probable among the possible results of the measurement, and let these possible results be distributed around this value in a gaussian way, with the same SD as for the whole sample. The distribution of sample [5] can be considered as the sum of the distributions around the single data, mathematically smoothed to reduce roughness without the likelihood of lack of fitting. The theoretical model and the best fitting gaussian sum superimposed to sample data are shown in Fig. 3.

Descriptive statistics

Every analysis of data is aimed to characterize the distributions by means of some experimental statistics (descriptives) [2]. Mean (moment of first order) and median (50th percentile) provide information about the values around which the numerical data tend to cluster. Variance (second-order moment of data centered around the mean) and SD describe the data spread. Skewness (third-order moment of data centered around the mean, free from data variability) quantifies the asymmetry of a distribution,

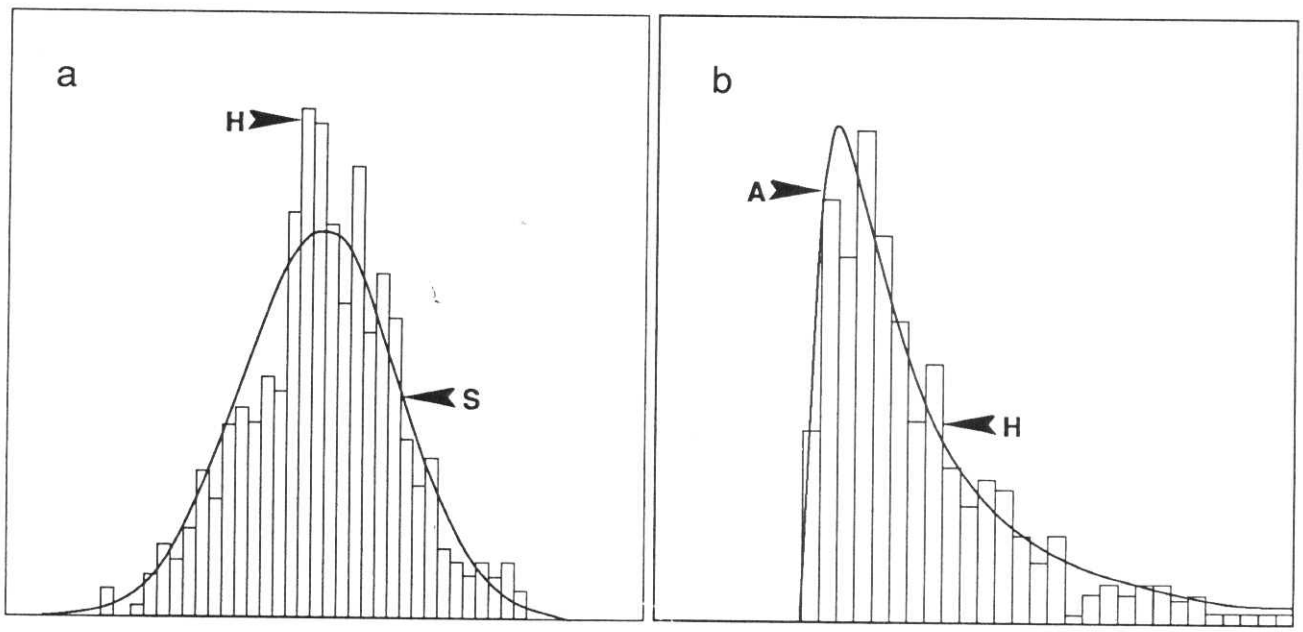

Fig. 1. - Theoretical distributions and histograms (H) of 400 data points, derived from the same distributions. (a) Symmetrical distribution (S), (b) asymmetrical, skewed to the right (A).

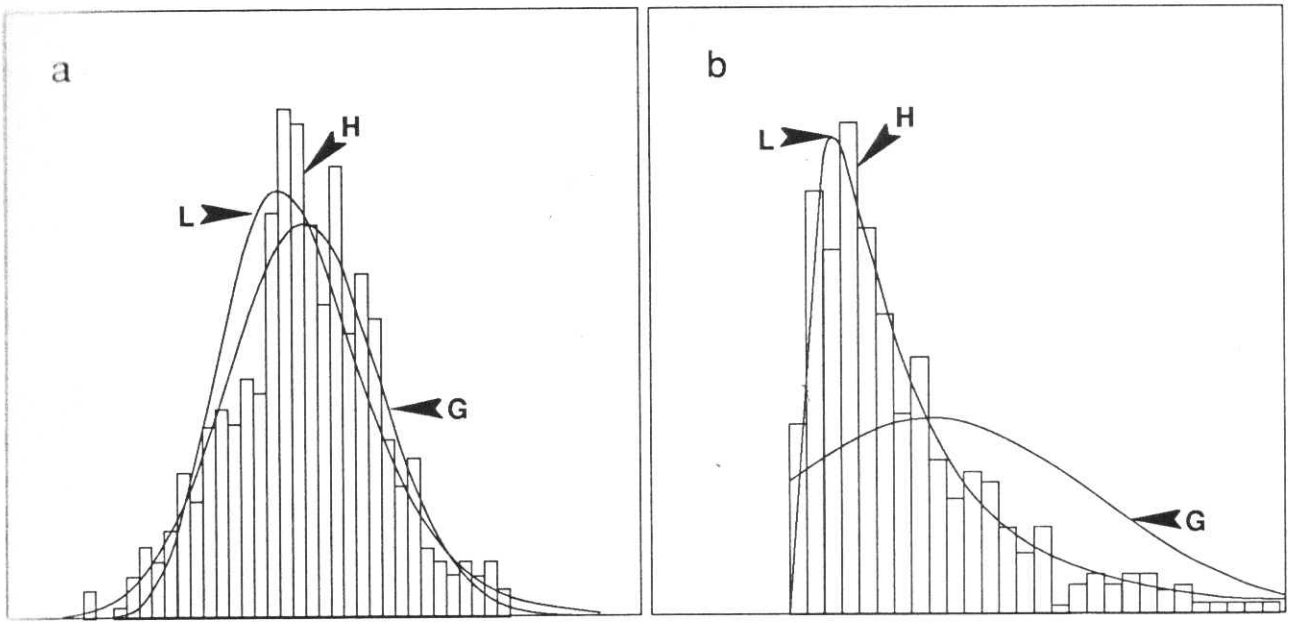

Fig. 2. - Density probability functions computed from (a) symmetrical and (b) asymmetrical histogram (H) of Fig. 1, under the hypothesis of a symmetrical empirical model (G) and a asymmetrical one (L).

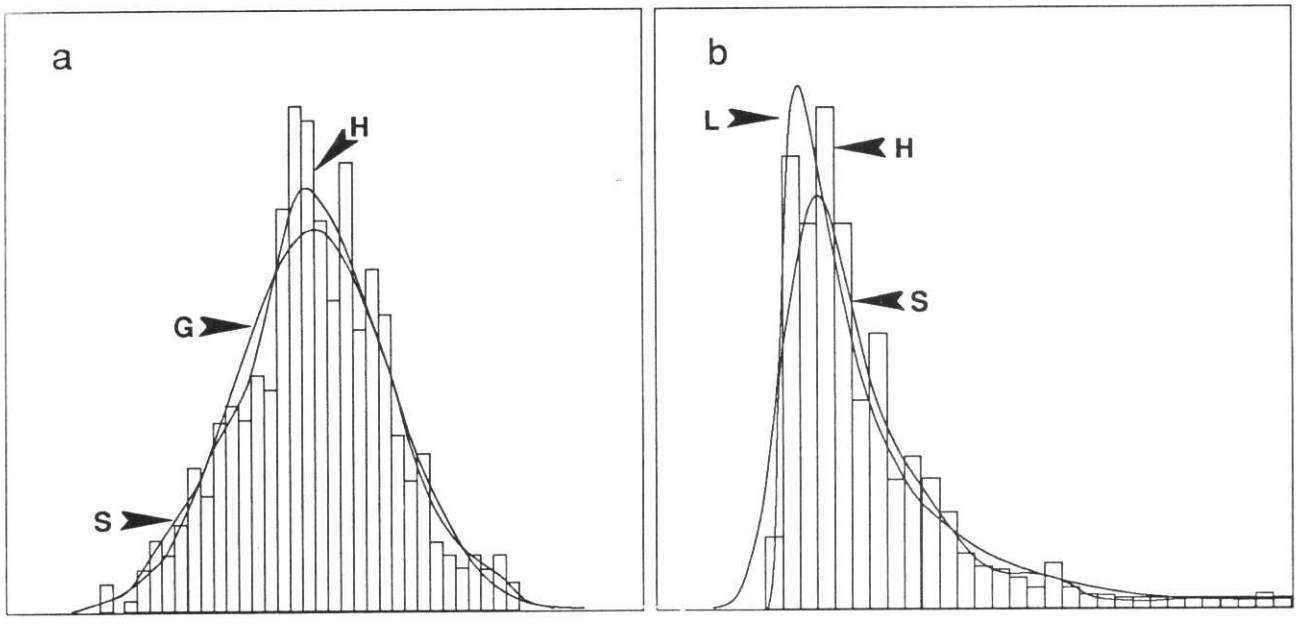

Fig. 3. - Fit of density probability function by means of the method of gaussian sum computed from the histograms (H) of Fig. 1, as compared to the theoretical distributions, for (a) symmetrical (G) and (b) asymmetrical (L) distribution.

while kurtosis (fourth-moment of data centered around the mean, free from data variability) relates to the "peakedness" of a distribution. Table 1 reports some descriptives of the data shown in Fig. 2, together with their associated errors (computed from the original data by the resampling techniques hereafter described [6]) and the corresponding theoretical values.

Descriptives report the macroscopic properties of distributions, so that only these quantities can be referred to for the characterization of the distributions themselves. This approximation, while simplifying the statistical treatment of data, neglects the microscopic aspects. Good results are provided for the center of distribution, but not for its tails. As a matter of fact, tails are the regions that clinical chemists have to analyze in evaluating methods or that physicians take into account for making diagnoses [3].

When more precise descriptions of data are sought, either a greater number of statistical moments can be taken into account or the local characteristics of data must be evaluated *via* percentiles.

Table 1. - *Expected values (m) and errors (s) of some descriptive statistics, in a symmetrical (A) and asymmetrical (B) sample (N = 400), together with their theoretical values for gaussian (g) and lognormal model (l)*

Sample	Parameter	m	s	g model	l model
A	Mean	9.995	0.061	9.995	9.995
	Median	9.969	0.055	9.995	2.885
	SD	0.960	0.161	0.960	0.960
	Skewness	- 0.014	1.948	0	-0.014
	Kurtosis	2.975	15.96	3	3.004
B	Mean	14.07	0.10	14.07	14.07
	Median	13.47	0.06	14.07	13.53
	SD	1.81	0.23	1.81	1.81
	Skewness	3.00	1.77	0	3.00
	Kurtosis	13.88	4.71	3	22.40

Variability

Unfortunately, the last statement is essentially the ideal and fully reliable information can be obtained only in the absence of errors. In fact, all data are affected by errors and, when several series of experimental data referring to the same population are obtained, the single value of statistics results in a cluster distributed around a central value.

An estimate of the cluster dispersion could be useful, but generally only single data sets are available, because of time and cost involved in repeating series of measurements.

Simulation and resampling techniques

Simulation with Montecarlo techniques [2] provides information on the statistical distribution of data by randomly producing a great number (e.g. 10^6) of possible values belonging to that distribution. The generation of these data requires knowledge of the mathematical form of the distribution. For example, this technique was used to simulate the samples of Fig. 1.

Resampling techniques [1, 6] (often referred to as *bootstrap*) are strictly related to the Montecarlo simulation techniques, but differ with regard to the basis of extraction of random values. In fact, in this case, the basis is not infinite but is given by the distinct combinations of sample data for the statistics of interest.

The bootstrap extends the characteristics of a sample to the population, taking into account that the latter contains *in nuce* all that is known and can be stated about the population itself. From this point of view, the population is nothing else but a resampling of the original data from which a higher number of other samples can be extracted. However, it is worth noting that resampling techniques extract only the information content of the sample without increasing it. In Fig. 4, examples of bootstrap resampling are shown.

By these techniques, but only by means of a computer, the variability of distribution statistics can be analysed and the consequences of changing either sample dimension, or decisional thresholds, or control rules [7], or other productive strategies can be investigated.

Reference limits and tests

The description of the local characteristics of a single distribution is not sufficient, nor necessary, in some cases. More interestingly, the relationship between different distributions can be analyzed, since different situations can be actually confused and the same unknown value cannot be associated unambiguously with a single situation [3]. In the process of classification of the single result, which transforms the result itself from "datum" into "diagnosis", the knowledge of reference limits, and the related uncertainty, allows the result to be accepted or rejected according to the degree of homogeneity with the data population.

The task assigned to statistical tests is a comparison of statistics or a check of hypotheses. Some tests verify if the means of two samples can be considered as the outcomes of the same mean (e.g. Student's test). With other tests the variability and the random errors (imprecisions) of two samples are compared: e.g. the chi-square test defines the range of an error and the Fisher's test evaluates the ratio betweeen two variances.

Statistics of multivariate distributions

The variables

A single situation gives very little information with which to characterize a pathophysiological situation, and the individual variability may predominate. Clinical chemists and physicians, therefore, try to collect information from several situations in order to describe the situation on "an average".

In the same way, when the single value is not sufficient to characterize the individual case, clinical chemists and physicians look for other data (the variables of case) to obtain more information. Fig. 5 shows a two-dimensional plot of two variables, which may be the results of the measurements either of two analytes in the same analytical sample, or of the same analyte with two different methods, or else of a variable and its variance, e.g. the response/error relationship [8].

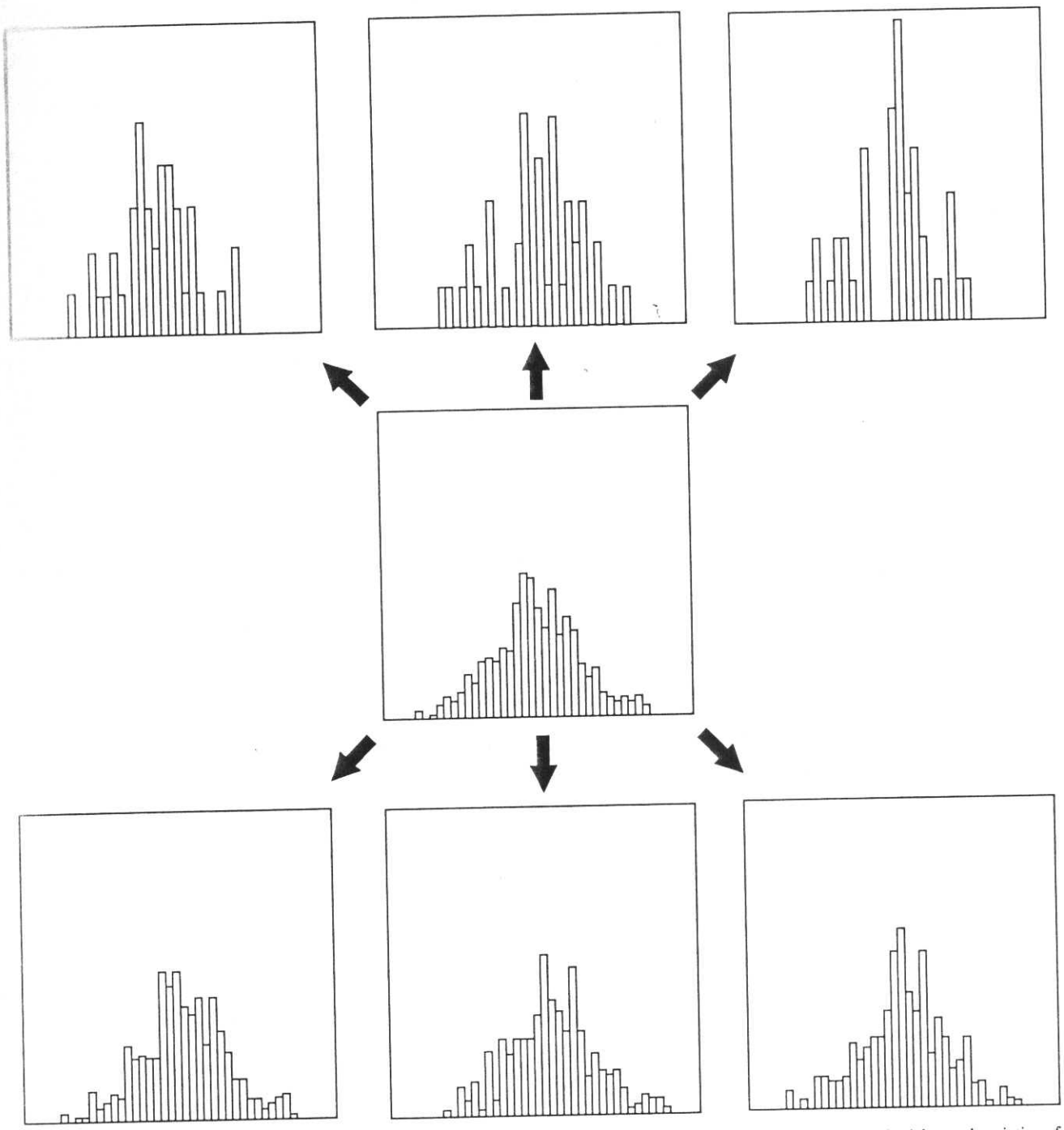

Fig. 4. - Bootstrap resamples with N = 40 (top) and N = 400 (bottom) of the sample shown in the middle (same as in Fig. 1a). A better description of the distribution is apparent for greater sample size.

In any case, the relationship among data must be investigated. In the first case, the independence of the variables indicates a high information content of both; in the second, the correlation of results demonstrates a between-method commutability and, in the third, quantifies a quality function to control the assay.

Regression

Multiple regression evaluates the dependence of the variables according to a given fitting (usually linear) model, using the information contained in the correlation matrix of data and a decisional criterion (e.g. the analysis of residual variance), through least-squares or maximum-likelihood methods [9]. The regression curve for the two variables of Fig. 5 is shown in Fig. 6.

Classification

Once a set of useful variables has been selected, a classification of the cases could be required, exactly as previously discussed for univariate analysis.

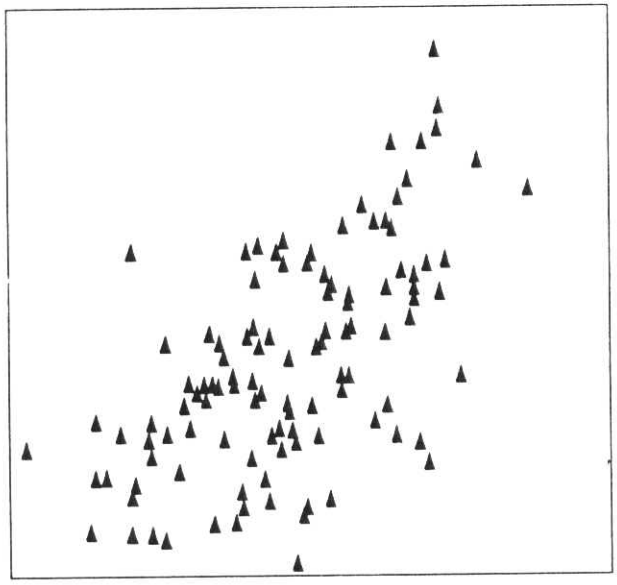

Fig. 5. - Bidimensional plot of two variables (N = 113). Every triangle represents a single case, while the plane is the space of variables.

The aim of classification is the determination of the domain of influence of different states or classes (errors, pathologies, etc.), the correct classification of the known cases (learning sample) and the allocations, according to some criterion, of the unknown cases.

Classification can be performed according to two main categories of method of different complexity, exemplified by: (a) discriminant analysis, where the objects are classified into known prototype classes [10] and (b) cluster analysis, where the unknown classes are also identified by the method through an iterative process of refinement [11].

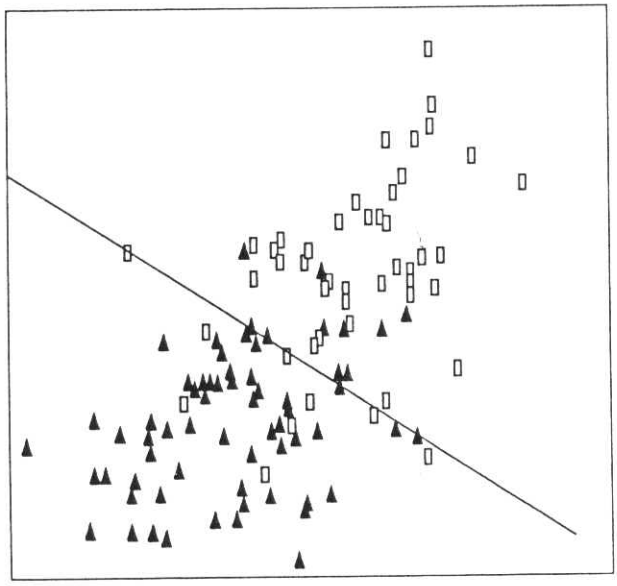

Fig. 6. - Regression curve between the variable of Fig. 5. The error range around the mean curve refers to a confidence level of 95%.

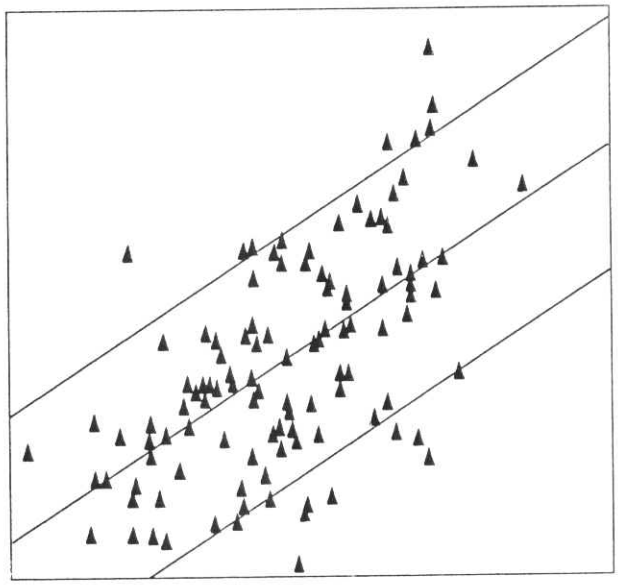

Fig. 7. - Results of discriminant analysis applied to the data of Fig. 5, considered as being related to two distinct situations. In this two-dimensional case, the "hyperplane" of separation is the shown straigth line.

Discriminant analysis allows the determination of a discriminant function, depending on the original variables, that assigns to each point (case) in the space of variables a value as to its degree of membership to a given class. All the cases are then allocated in the given classes by means of optimal cutoffs (from the decisional point of view), that graphically draw separating surfaces (hyperplanes). In Fig. 7, discriminant analysis is applied to the data plotted in Fig. 5, considered as two learning samples extracted from two different states (classes). In this two-dimensional case, the hyperplane of separation is a straight line.

The essential feature of clustering is that objects are sorted into subsets containing data points as alike as possible. Allocation into the subsets is made according to some decisional function, following logical paths optimized as far as costs (iteration number) is concerned. Fig. 8 exemplifies a cluster analysis (same data as Fig. 7), as obtained through stopping the process at two different steps ("thresholds") of clustering.

Principal components analysis

When a case is defined by several variables, it is useful to reduce the redundancy of information by selecting those uncorrelated variables holding the richest information content.

The analysis of principal components consists in changing the reference system from the space of the original variables into a new space, where the compression of information is optimal [12]. In this new space, the operator selects the number of variables (linear combinations of the original variables) which can be rejected with a given loss

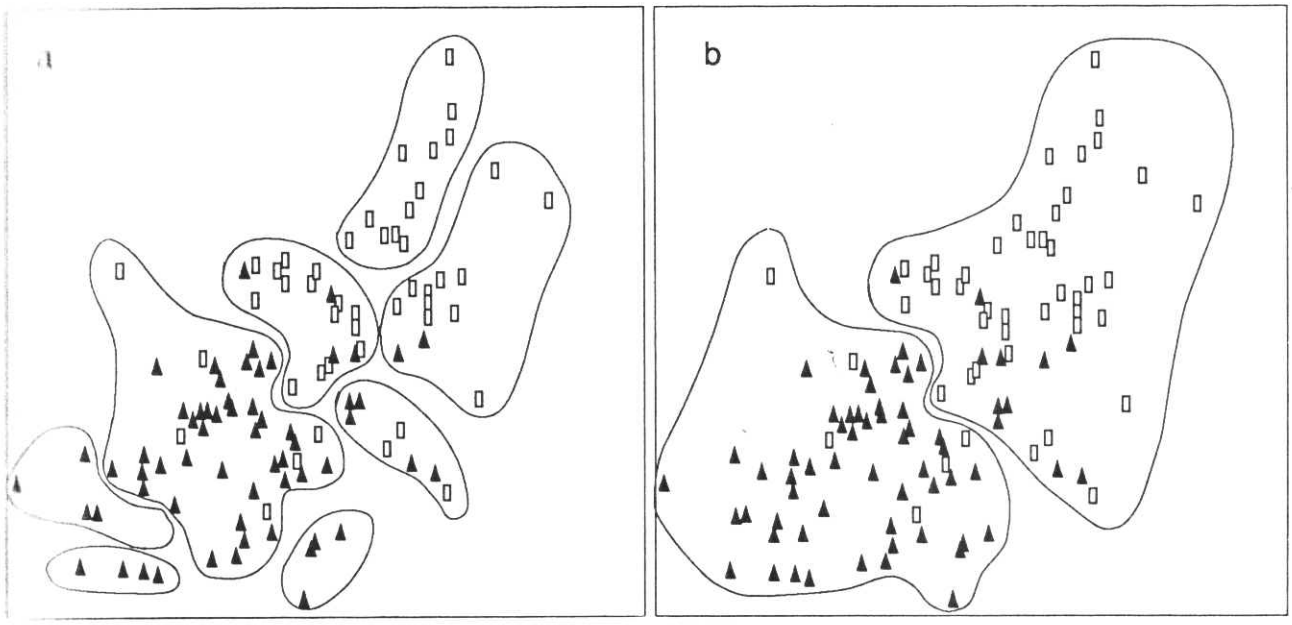

Fig. 8. - Results of cluster analysis, when the automatic classification of data of Fig. 5, is stopped at (a) minimal threshold or (b) optimal threshold. The clusters have been superimposed on the data of Fig. 7, to point out the efficacy and residual misclassifications.

of information. When many variables exist, principal components analysis associated with correlation or multiple regression allows the compression of information, as described above, to be transferred to some of the original variables.

As a simple (two dimensional) example, easy to represent, the principal components analysis is applied in Fig. 9 to the data of Fig. 7. The example demonstrates the great information content (81%) of one of the new variables. This single variable could be sufficient to characterize the samples and to allow a first classification.

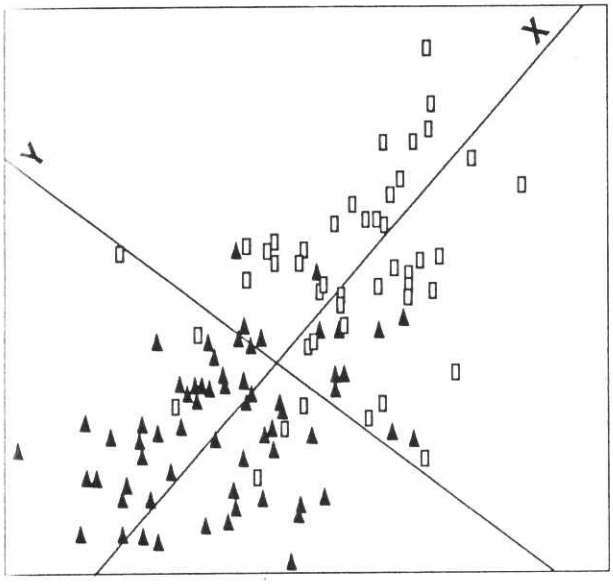

Fig. 9. - Results of the application of principal component analysis to the data of Fig. 7. The two inclined X and Y axes represent the new reference system. The data are sufficiently aligned with the X axis, whose associated variable has high information content (81%).

Conclusion

Making decisions is unavoidable. Decision must be made even when elements for decision are unavailable ("undecidability") but alternatives are not allowed by external constraints. In fact, decision is a complex problem based on a hierarchic organization of possible choices and checks for congruity of any decision step. But neither the organization nor the check are possible without statistical tools related to "statistics of decision", which save decisions from acritical automatisms.

Statistics of decision [13, 14] will be not treated here, as several topics involved are authoritatively dealt with in this issue [15-17]. It is suggested, as a conclusion, that there is a possible analogy between epidemiology of disease and epidemiology of error, and a desirable unification of tools and approaches in discriminating between normal and pathological values or correct and uncorrect results.

Review submitted on invitation by the Editorial Board of the *Annali*.
Accepted for publication: 5 February, 1991.

REFERENCES

1. EFRON, B. 1982. *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics, Philadelphia.

2. HAHN, G.J. & SHAPIRO, S.S. 1967. *Statistical models in engineering*. Wiley, New York.

3. CHIECCHIO, A., BO, A. & MIGLIARDI, M. 1987. Analysis of distributions and determination of references limits in short series of data. Application to RIA. *J. Nucl. Med. All. Sci.* **31**: 195-200.

4. HEALY, M.J.R. 1979. Outliers in clinical chemistry quality-control schemes. *Clin. Chem.* **25**(5): 675-677.

5. HERMANS, J. & HABBEMA, J.D.F. 1976. *Manual for the ALLOC discriminant analysis program*. Deptartment of Medical Statistics, University of Leiden.

6. EFRON, B. & TIBSHIRANI, R. 1988. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Sci.* **1**(1): 54-77.

7. GROTH, T., FALK, H. & WESTGARD, J.O. 1981. An interactive computer simulation program for the design of statistical control procedures in clinical chemistry. *Computer Progr. Biomed.* **13**: 73-86.

8. EKINS, R.P. 1981. The "precision profile": its use in RIA assessment and design. *Ligand Q.* **4**: 33-44.

9. DRAPER, N.R. & SMITH, H. 1966. *Applied regression analysis*. Wiley, New York.

10. ANDERSON, T.W. 1958. *An introduction to multivariate statistical analysis*. Wiley, New York.

11. HARTIGAN, J.A. 1975. *Clustering algorithms*. Wiley, New York.

12. YAGLOM, A.M. 1962. *Stationary random functions*. Prentice-Hall International, London.

13. GALEN, R.S. & GAMBINO, S.R. 1975. *Beyond normality: the predictive value and efficiency of medical diagnoses*. Wiley, New York.

14. GREEN, D.M. & SWETS, J.A. 1974. *Signal detection theory and psychophysics*. Krieger, Huntington.

15. CAREY, R.N. 1991. Implementation of multirule quality control procedures. *Ann. Ist. Super. Sanità* **27**(3): 419-426.

16. SCANDELLARI, C. 1991. Bayesian approach in evaluating and planning diagnostic data. *Ann. Ist. Super. Sanità* **27**(3): 385-394.

17. GAMBINO, S.R. 1991. The misuse of predictive value - or why you must consider the odds. *Ann. Ist. Super. Sanità* **27**(3): 395-400.