Aspects in Development of Statistic Data Analysis in Romanian Sanitary System

DANA SIMIAN[1], CORINA SIMIAN[1], OANA DANCIU[2], LAVINIA DANCIU[3]

[1] Faculty of Sciences
University "Lucian Blaga" Sibiu
Str. Ion Ratiu 5-7, 550012, Sibiu
ROMANIA
dana.simian@ulbsibiu.ro
corinafirst@yahoo.com

[2] Faculty of Medicine
Carol Davila Bucharest
B-dul Eroilor Sanitari 8
ROMANIA
microwawe_oana@yahoo.com

[3] Hospital of Pneumology Sibiu
Aleea Filozofilor 3-5 Sibiu
ROMANIA
medlavi@yahoo.com

Abstract: - The aim of this paper is to present a model for the statistical analysis of data in sanitary system from Romania. We establish the factors involved in the evolution of diseases, for appreciating the tendency of evolution of these affections in the following years, in order to create an efficient system of monitoring and prevention. We evaluate software products for their use in analysis of data provided from the Romanian Informational Program in Sanitary System. A practical study for the case of tuberculosis is made.

Key-Words: - data analysis, statistics, software

1 Introduction and motivation

Romania is a country which finds itself at the beginning of using informational programs in the Sanitary System. These programs have been used at a national level only for the last five years, and mostly in Public Health Programs. This situation has the following impact: on one hand, it enables us to complete an information data base, but on the other hand it does not allow us to use or value it efficiently, for studying different illnesses and their impact on the population. This is mainly due to the fact that informational programs require a certain amount of time for processing data and to the fact that our Sanitary System confronts itself with a lack of specialized personnel.

The present National Informational Program contains a data base that includes the identification elements of the patient (name, surname, personal identity number, gender, age, residence-urban or rural), his/her level of studies, occupation, as well as data concerning the type of illness, its infectiousness, the investigations required, treatment, evaluation of treatment (results of treatment) and healing process. It is highly important to use the data base as efficiently as possible, in order to extract information regarding the factors that influence the treatment, the evolution and the healing perspective of these patients.

An interesting and useful possibility is to improve and use the existing data base for analyzing the correlations between different factors involved in the evolution of diseases, for appreciating the tendency of evolution of these affections in the following years, in order to create an efficient system of monitoring and prevention.

For this purpose it is necessary first to establish the model for the statistical analysis of data, that is the useful statistical data to be computed and statistical tests to be performed. On the other hand, one software product might be choice for implementing our model.

The aim of this paper is to present a model for the statistical analysis of data in sanitary system from Romania and to make a comparison of the most important software products which can be used for implemented our model, taking into account the performances of these software products, their user interface, accessibility for no specialist persons, graphical facilities for results rendering, possibility of import and export different types of data. For the purpose of this paper, have been used SPSS and MATLAB. For modeling purposes AnyLogic and Synapse are proposed..

Our study also allows improving the performance of the sanitary informational system determining more parameters to be included in the data base and using statistical analysis of data for obtaining useful information in the shortest time possible.

Since now, in our medical system, primary statistical analysis of data has been used only for comparison the distribution of patients with different diseases, using simple criteria, like sex or geographical distribution. None correlation study was made.

The paper is organized as followed: in section 2 is presented the statistical model used for our analysis. Section 3 presents and compares software products for data analysis and for modeling and prediction. In section

4 we make a practical study, for case of tuberculosis, using the National Tuberculosis Control Program's data base.

2 Data analysis

The aim of this section is to present the main types of statistical analyses which might be done using the data from the sanitary national informational data base. We also give the possibilities of interpretation of the results.

Frequencies analysis

The data from the data base can be group in categories or classes, using different criteria: The criteria we take into account and the classes are:

1. Sex - two classes (*m* = male, *f*=female).
2. Age - four classes, denoted by integer numbers, 1…4 (*1*= age less than 20, *2*= age between 20 and 40, *3*= age between 40 and 65, *4*= age greater than 65)
3. Place of residence - two classes (*town, village*)
4. Level of education – four classes (*low, elementary, medium, height*)
5. Treatment – five classes (*cured, treatment complete, abandon*, *decades*, *lost* from evidence)
6. Reaction to the treatment - four classes (*none, mould, moderate, height*)
7. Intolerance to the treatment - four classes (*none, mould, moderate ,severe*)

Ones of these data groups (1-4) can be obtained directly from the information given by the patient, others suppose a period of treatment and observation of the patient (5-7).

First type of analysis consists in reports for absolute or relative frequencies of the classes presented before (frequency counts, percentages, mean, minimum and maximum values) This situation is given analytic and graphic, as bar or pie charts, at the end of a temporal interval (month, quarter, year) and suppose filtering the data and computing the frequency values of the classes, or importing the data in a data analysis software and makethis analysis using this software.

Descriptive statistics

The descriptive statistics for data is also useful and will be provided in tabular form.

.

Summarize statistics

In a larger period of time we need to calculate subgroup statistics for variables within categories of one or more grouping variables. Summarize statistics includes sum, number of cases, mean, median, grouped mean, grouped median, minimum, maximum.

Frequencies analysis, descriptive statistics and summarize statistics are usually required to be reported for medical analysis of a health sector or sub sector (pulmonary diseases, cardiac diseases, etc.). Our aim is to improve the utility of national health databases, providing results regarding possible dependencies between different variables. First we must determine if a relationship exists between these quantities.

Covariance and correlation coefficients

Correlations measure how variables or rank orders are related. Correlation quantifies the strength of a linear relationship between two variables. We are interested in the degree of relationship between variables; therefore we might calculate correlation coefficients. These coefficients are derived from covariances. The covariance matrix is

$$\begin{bmatrix} \sigma^2_{11} & \sigma^2_{12} & \sigma^2_{13} \\ \sigma^2_{21} & \sigma^2_{22} & \sigma^2_{23} \\ \sigma^2_{31} & \sigma^2_{32} & \sigma^2_{33} \end{bmatrix}$$

$$\sigma^2 ij = \sigma^2 _{ji}$$

The correlation coefficient $r_{X,\,Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is their covariance normalized by their standard deviations. The correlation coefficients range from -1 to 1. Values close to or equal to 0 suggest there is no linear relationship between the data columns. We can use Pearson's correlation coefficient or partial correlation coefficients that describe the linear relationship between two variables while controlling for the effects of one or more additional variables.

Two variables can be perfectly related, but if the relationship is not linear, a correlation coefficient is not an appropriate statistic for measuring their association.

Regression analysis

If two values are linear dependent we can use linear regression which estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.

Ordinal regression allows us to model the dependence of a polytomous ordinal response on a set of predictors, which can be factors or covariates.

The estimated coefficients reflect how changes in the predictors affect the response. The response is assumed to be numerical, in the sense that changes in the level of the response are equivalent throughout the range of the response. A classical

example for using ordinal regression is the study of patient reaction to drug dosage.
There is also the possibility that the dependence is described by another curve. In this case other regression models might be used: logarithmic, inverse, quadratic, cubic, power, growth, and exponential, etc.
Many details about frequencies analysis**,** covariance, correlation and regression analysis can be found in [1], [2].

3 Tools for data analysis

The national health databases have not implemented procedures for performing data analysis. Frequencies analysis, descriptive statistics and summarized statistics could be implemented directly, as procedures and functions working with database. However, a complex statistical analysis, together with dependencies analysis and designing of various models using the results of these analyses, require dedicated statistical analysis and modeling software for this goal.
The aim of this section is to analyze possible software to be used for accomplishing this task.
We have two possibilities: to use existing dedicated software using as input data the information from health databases or designing our own dedicated software.
In this section we analyze only the first possibility.
There are many free and licensed software products for data analysis ([4], [5]). Taking into account the importance of the national health informational program, the possibility of connection to the European programs, we are looking for licensed software for data analysis.
We considerate the following programs: SPSS ([6], [7]) and MATLAB ([9]). We analyze them taking into account usability aspects, data analysis aspects and distributed analysis aspects.

Criteria
Usability aspects
Ability to import and export data
Ability to organize and manage data
Existence of an editor for defining, entering, editing and and displaying data*
Clarity and simplicity of interface*
Visualization
Command language
Accessibility for no specialized persons*
Local installation
Help system
Data analysis aspects
Features for data transformations*
Feature for frequencies statistics
Descriptive statistics
Regression analysis
Graphical representation
Factorial analysis
Distributed analysis mode
Tools support features

The criteria marked with * are better for SPSS than for MATLAB in the environment of our goal. Therefore we recommend SPSS for performing data analysis in our case, especially for a better accessibility for no specialized persons.

For modeling we recommend to use model development software, such as AnyLogic and which allows developing models and for prediction we recommend software for development of adaptive systems such as Synapse ([8]). Both AnyLogic and Synapse offer complete tutorials, user guides and demos. Adaptive systems can be applied for prediction and medical diagnosis.

4 Data analysis for tuberculosis database

Even though tuberculosis is known since Ancient times, it still remains a matter of national interest, not only for European countries, but for all those which deal with a great number of infected or ill individuals. In Romania, the incidence of this ailment is much higher than the average in Europe, being of 120/100000, this having a correspondence of 23000 ill people in population. These values make us the leading country in Europe regarding tuberculosis, pointing the fact that it represents a matter of Public Health. For this reason and not only, our government has implemented a surveillance program for population monitoring in an Informational System.
In the context of free international circulation, the most efficient management of tuberculosis should not represent only a matter of national interest, but one of international concern.
The aim of this section is to make a data analysis using the data from national tuberculosis database and to render the results. The analyze we made underlines the statistic which is recommended to be done using the data from the health informational system for allowing a clear evidence of the factors which influence the disease.
Taking into account the dynamics of the disease, it is useful to realize a monthly analyze. A summarized temporal analysis must be realized in each year.
The access to the database is allowed to district level, that is, the personnel working in a geographic district area has access only to the data from this district. The access to the whole data base is allowed only for few specialized person from the central forum. It is necessary and very useful to be done a data analysis for each district.

In this section we use data from tuberculosis data base, district Sibiu, month January 2009.

Singular data frequencies analysis

We can make frequencies analysis for only one group of data in the sense defined in section 2. The results can be display in tabular (fig. 1) or graphic way (fig. 2).

Statistics

SEX

N	Valid	51
	Missing	0

SEX

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MASCULIN	41	80.4	80.4	80.4
	FEMININ	10	19.6	19.6	100.0
	Total	51	100.0	100.0	

```
RECODE
  varsta
  (Lowest thru 20=1)  (21 thru 40=2)  (41 thru 65=3)  (65 thru Highest=4)
  INTO  v1 .
EXECUTE .
GRAPH
  /HISTOGRAM(NORMAL)=varsta .
```

Fig. 1 – Tabular frequencies analysis on sex groups

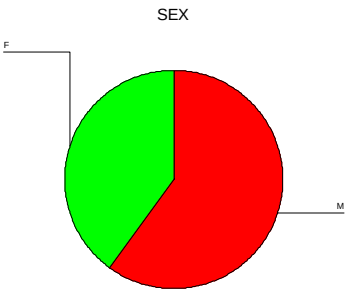

Fig. 2 – Graphical frequencies analysis on sex group (pie chart)

Singular data frequencies analysis is relevant only for counting the patients from groups created using a given criteria. We can see how many ill persons are male or females or which is their distribution in classes of age.
This kind of analysis is recommended to be presented in a pie chart.

Summarized frequencies data analysis

The connection between groups obtained using different criteria can be illustrated using summarized frequencies data analysis. We can analyze how the results of analysis are distributed taking account different criteria sex (fig. 3, fig. 4, fig. 5), age (fig. 6), etc. The results can also be displayed in tabular or graphical way (fig. 3, fig. 4)
This kind of analysis is more relevant and the rendering of the results is more complex.

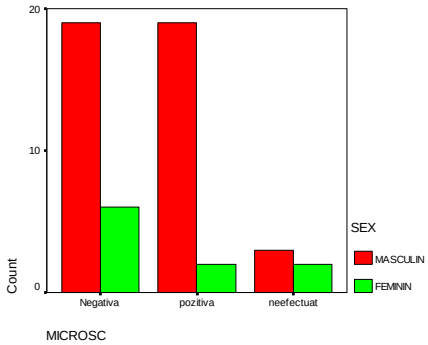

Fig. 3 – Distribution of the results using sex criteria

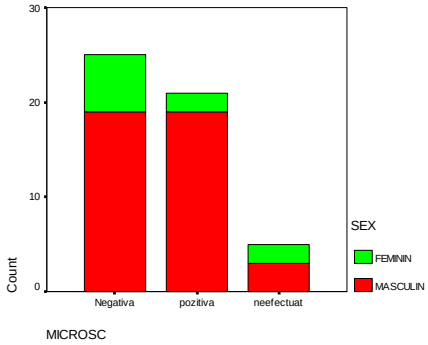

Fig. 4 – Distribution of the results using sex criteria

Frequency Table

SEX

MICROSC			Frequency	Percent	Valid Percent	Cumulative Percent
Negativa	Valid	MASCULIN	19	76.0	76.0	76.0
		FEMININ	6	24.0	24.0	100.0
		Total	25	100.0	100.0	
pozitiva	Valid	MASCULIN	19	90.5	90.5	90.5
		FEMININ	2	9.5	9.5	100.0
		Total	21	100.0	100.0	
neefectuat	Valid	MASCULIN	3	60.0	60.0	60.0
		FEMININ	2	40.0	40.0	100.0
		Total	5	100.0	100.0	

Fig.5 – Distribution of the results using sex criteria

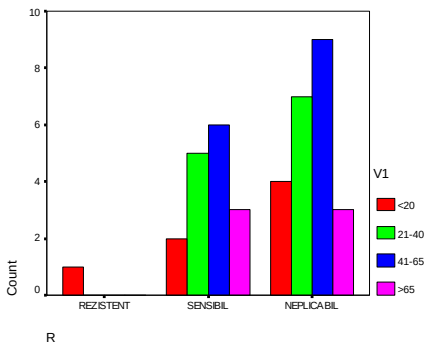

Fig.6 –Resistance to virus,, using age criteria

Using many analysis of this type, suggestions about possible dependences can be obtained. By example an analysis of treatment situation using age criteria (fig. 7),

together with the analysis of the results obtained to the resistance to virus using the same criteria suggests a connection between these elements.

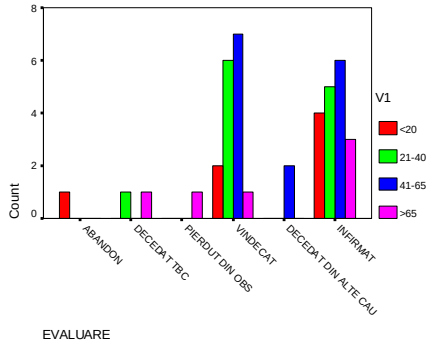

Fig.7 Treatment results, using age criteria

Empirical verification can be done using the analysis of treatment results using criteria of resistance to virus (fig. 8).

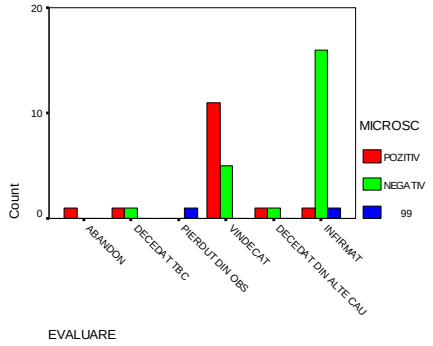

Fig.8 – Treatment results using resistance to virus criteria

Descriptive statistics

Descriptive statistics in tabular form (fig. 10) or graphical form is also useful (fig. 11). A part of descriptive statistics for analyzing the data grouped in multilevel is presented in fig. 9.

Statistics

MICROSC			SEX	LOC	OCUPATIE
Negativa	N	Valid	25	25	25
		Missing	0	0	0
	Mean		1.24	1.48	2.20
	Median		1.00	1.00	2.00
	Mode		1	1	2
	Std. Deviation		.44	.51	.50
	Skewness		1.297	.085	.435
	Std. Error of Skewness		.464	.464	.464
	Kurtosis		-.354	-2.174	.490
	Std. Error of Kurtosis		.902	.902	.902
	Minimum		1	1	1
	Maximum		2	2	3

Fig. 9 – Descriptive statistics

The statistic distribution can also be illustrated in *a* graphical way. Figure presents the low of patients' distribution (taking into account the age criteria) and the values a descriptive statistics.

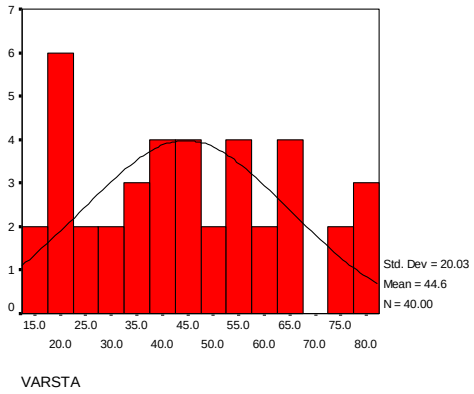

Fig. 10 – Distribution low representation

Determination of the type of distribution is useful, further for calculate covariance, correlation coefficients and find a regression rule.

This task will be accomplished in another article, because it requires access to a larger volume of data (access to the national data base not only the data for a district).

5. Conclusions and further direction of study

In this article we presented the main types of data analyses which are relevant to be made with data from the national health databases. An analysis of the software products recommended for this analyzes is made. The analysis indicates SPSS to be use together with a software for prediction and modeling.

We intend to develop a GUI based on these software products for making the analysis for unspecialized personnel.

Moreover, using the data analysis, we could find out what other factors should be introduced in the data base in order to be able to predict the evolution of this affection in the following years. If, for example, we deduce that the most significant impact on tuberculosis is brought by a certain age group, we can create a system of monitoring and prevention, designed especially for that certain age group. If we identify a certain group susceptible to tuberculosis due to its socio-economical status, or due to its level of studies, we can approach this group of patients with specifically designed programs.

A program designed as a "site" or " interactive webpage" would also be useful, because it would allow people to bring about their ideas, suggestions, their own

experience regarding the disease and its impact on the individual's life. This would represent a new way of approaching the relation between patient and the Sanitary System, which is actually designed for this purpose.

Taking into consideration the fact that young people are extremely interested in using the internet as a source of information, we could develop a program of sanitary education which would be much more efficient and with significant lower costs, compared to the present programs, which use banners, commercials, posters and other types of displays. This program can include also a GUI for presented the evolution of different disease taking into account many factors.

Moreover, the national health program could be connected to the European programs, and the ones involved in the management of different affections could cooperate more easily, by exchanging data and information regarding patients, costs, efficiency.

References:

[1] Dodge Y., *The concise encyclopedia of statistics*, SpingerVerlag, 2008, 616p.

[2] Luca N., XIV, 299 p *The statistical analysis of MRI data*, Series Statistics for Biology and Health, Springerverlag 2008, 299p

[3] Lance A. Waller, Carol A I., *Applied Spatial Statistics for Public Health Data,* 112p, online book

[4] Statistical Science Web: Free Statistical Programs http://www.statsci.org/free.html

[5] Interactive Statistical Calculation Pages http://www.r-project.org/

[6] Resources to help you learn and use SPSS http://statpages.org/

[7] SPSS 10 Guía para el análisis de datos, www2.uca.es/serv/ai/formacion/spss/Pantalla/verguia.pdf

[8] http://www.peltarion.com/

[9] http:// www.matworks.com