Raport Badawczy

Research Report

RB/15/2003

Selection of variables for systems analysis – application of a fuzzy statistical test for independence

O. Hryniewicz

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 8373578

fax: (+48) (22) 8372772

Kierownik Pracowni zgłaszający pracę:
Prof. dr hab. inż. Olgierd Hryniewicz

Warszawa 2003

Selection of variables for systems analysis - application of a fuzzy statistical test for independence

Olgierd Hryniewicz
Systems Research Institute, Newelska 6, 01-447 Warsaw, Poland
hryniewi@ibspan.waw.pl

Abstract

The problem of the selection of variables for systems analysis is considered. Variables are selected according to the statistical analysis of experts opinions. We propose to describe possible ambiguous experts opinions by possibility distributions, and thus by fuzzy sets. For such fuzzy data we propose a fuzzy version of the Pearson's chi-square test of independence.

Keywords: Fuzzy data, Statistical test of independence, Possibility.

1 Introduction

Systems analysis deals with complex problems and processes described by many possible variables that may be uncertain and imprecise. These problems and processes we treat as certain systems described by mathematical models. Building of such mathematical models requires a large amount of information, especially when input information is of a random nature, and the models are presented in a form of regression equations. When this information is available we could use well known methods of mathematical statistics to select the best set of variables that describe the considered system. Using such methods we could select explanatory variables which, on the one hand, are highly correlated with the main characteristics of the system (so called outcome variables) but, on the other hand, are mutually independent. However, in many cases this information is not available, and the cost of its acquisition is very high. For example, in building mathematical models for complex socio-economic phenomena we need results of costly polls. In such a case, we need to select the appropriate variables in advance in order to reduce the amount of necessary information. In order to do this we propose to use experts opinions.

The simplest way to select variables using expert opinions is to ask them about possible dependencies between different variables. This approach may be not efficient, especially in the case of different or even conflicting opinions. Therefore, there is a need to propose a more objective method of acquiring expert opinions. In the second section of this paper we propose a statistical procedure for establishing possible dependencies between variables of interest which utilises the statistical test of independence for categorical data. In this procedure we divide the range of possible values of each variable of interest into a finite number of categories. This means that for sets of possible values of the considered variables we assign some labels. These categories may be defined precisely (e.g. by defining numerical intervals such as $10 < X \leq 20$) or imprecisely (e.g. by using imprecise linguistic notions as "high income", "low risk", etc.). We acquire necessary information by asking questions such as "If the value of the explanatory variable X belongs to the category x_i what is a corresponding value of the outcome variable Y?". In the case of unambiguous answers (i.e. when variables are strongly depen-

dent, and experts are able to indicate only one possible category for the outcome variable Y) we propose to use the well known Pearson's chi-square test of independence in order to find information about possible dependencies. However, we cannot expect such unambiguous answers - especially in the case of independent or weakly correlated variables. Thus, we may face imprecise answers that may be described in terms of possibility distributions. We consider this case in the third section of this paper where we introduce the fuzzy version of the chi-square test of independence. Finally, in the fourth section of the paper, we discuss the obtained results and indicate the problems for future investigation.

2 Test of independence using unambiguous expert opinions

Suppose that we have to investigate a possible dependence between an explanatory variable X and the outcome variable Y. Let $\{x_1, x_2, ..., x_k\}$ be a set of labels (categories) that describe the possible values of the explanatory variable X, and $\{y_1, y_2, ..., y_r\}$ be a set of labels (categories) that describe the possible values of the outcome variable Y. For each of n experts we *randomly* choose one value of X. Then we ask each expert "If $X = x_i$ which is the most plausible value of Y?". We expect that the expert indicates only *one* value of Y. This answer could be described in the following form:

X/Y	y_1	...	y_j	...	y_r
...	...	...	...	...	...
x_i	0	0	1	0	0
...	...	...	...	...	...

In such a case the results of questioning can be summarised in a form of a two-way kxr contingency table.

X/Y	y_1	...	y_j	...	y_r	$\sum_j$
x_1	n_{11}	...	n_{1j}	...	n_{1r}	$n_{1\cdot}$
...	...	...	...	...	...	...
x_i	n_{i1}	...	n_{ij}	...	n_{ir}	$n_{i\cdot}$
...	...	...	...	...	...	...
x_k	n_{k1}	...	n_{kj}	...	n_{kr}	$n_{k\cdot}$
$\sum_i$	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot r}$	n

where n_{ij} describes the number of indications (or *observations*) in the ij-th cell, and

$$n_{i\cdot} = \sum_{j=1}^{r} n_{ij} \tag{1}$$

$$n_{\cdot j} = \sum_{i=1}^{k} n_{ij}. \tag{2}$$

Pearson introduced the notion of "the expected number of observations". This is the expected number of observations in each cell of the contingency table, calculated under the assumption that both variables X and Y are mutually independent. He proposed to calculate these values from the formula

$$\hat{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}. \tag{3}$$

Then, he proposed to measure the "distance" between the observed contingency table and the ideal one using the famous chi-square statistics

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}. \tag{4}$$

If n is sufficiently large, and $n_{ij} > 5$, $i = 1, ..., k$, $j = 1, ..., r$, then the χ^2 statistics has asymptotically the chi-square distribution with $(k-1)(r-1)$ degrees of freedom. We should reject the hypothesis of independence at the significance level δ when the value of the chi-square statistics is too large, namely larger than the quantile $\chi^2_{(k-1)(r-1),1-\delta}$ of the corresponding chi-square distribution. The statistical test described above is well known as Pearson's chi-square test of independence.

Further information about the chi-square test of independence the reader can find in any statistical textbook (e.g. by Bickel and Doksum [2]). Information about more advanced methods of testing independence for categorical data can be found in Agresti [1].

If we need to build a model with a limited number of explanatory variables we should choose those that are mutually independent, and highly correlated with the outcome variable. To select the variables which has the

highest correlation with the outcome variable we can use the well known Tchouproff's index

$$T^2_{XY} = \frac{\chi^2}{n\sqrt{(k-1)(r-1)}}. \qquad (5)$$

Explanatory variables with the highest value of this index have probably the strongest correlation with the outcome variable.

3 Test of independence using fuzzy expert opinions

The assumption that the expert will indicate only one value of Y without hesitation in many cases is obviously unrealistic. Consider for example a situation when both variables X and Y are independent. Then it is quite probable that the expert is not able to indicate the most plausible value. In such a case he either chooses the answer randomly or indicates two (or more) values as equally plausible. In the case of kxr contingency table his answer could be like the one presented in the table:

X/Y	y_1	...	y_j	y_{j+1}	...	y_r
...	...	...	...	...	...	...
x_i	0	0	1	1	0	0
...	...	...	...	...	...	...

This result looks like a well known in statistics case of multiple answers. However, its interpretation is quite different. In the problem of multiple answers two (or more) values of Y may occur *simultaneously*. An example of such a problem was described by Loughin and Scherer [10] who investigated the association between the educational background of farmers and the source of some veterinary information. In such a case it is quite probable that people would indicate different sources from which they obtained this information. In our case the situation is quite different. The reason for a multiple answer is either a lack of knowledge or even an inherent inability, as in the case of independent variables.

Let us consider the interpretation of the "multiple" answer that was presented in the example given above. What does it mean in practice? One possible answer is that the expert sees both outcomes y_j and y_{j+1} as equally *probable*. Another possibility is that the expert sees both outcomes as equally *possible*. Let us discuss the difference between these two interpretations. In the first case, the expert evaluates the expected frequency of the occurrence of those outcomes, and the reported numbers may be easily transformed to probabilities (0.5 and 0.5 in the considered case). In the second case, the expert says that both outcomes y_j and y_{j+1} are equally possible but not necessarily equally probable, and their probabilities belong to the interval $(0, 1]$. Thus, in a real experiment we could expect both results

X/Y	y_1	...	y_j	y_{j+1}	...	y_r
...	...	...	...	...	...	...
x_i	0	0	0	1	0	0
...	...	...	...	...	...	...

and

X/Y	y_1	...	y_j	y_{j+1}	...	y_r
...	...	...	...	...	...	...
x_i	0	0	1	0	0	0
...	...	...	...	...	...	...

Now, let us consider a more general situation when the answer is given in the following form

X/Y	y_1	...	y_j	...	y_r
...	...	...	...	...	...
x_i	μ_{i1}	...	μ_{ij}	...	μ_{ir}
...	...	...	...	...	...

where $\mu_{ij} \in [0, 1]$, $i = 1, ..., k$; $j = 1, ..., r$ and $\max_{i,j} \mu_{ij} = 1$. We may interpret the values of μ_{ij} as the degrees of possibility that for the $X = x_i$ we will observe $Y = y_j$. Hence, for a given possibility μ_0 we may observe in a real experiment all values whose possibilities are not smaller than μ_0.

Let's denote $n_{i\cdot}$ by n_i, then we have $\sum_{i=1}^{k} n_i = n$. Now, a single expert's opinion can be described in a general form by the following vector:

$M_{i;q} = (\mu_{i1;q}, \mu_{i2;q}, ..., \mu_{ir;q})$, $q = 1, ..., n_i$, $i = 1, .., k$, $0 \le \mu_{ij;q} \le 1$, $j = 1, ..., r$, $\max_j \mu_{ij;q} = 1$.

We assume that this vector represents a *possibility distribution* given in a form of a *fuzzy set* that describes the expert's answer, and numbers $\mu_{ij;q}$ are the values of the membership function assigned by the expert to possible observations (x_i, y_j).

For a given α-cut $(0 < \alpha \leq 1)$ the expert's opinion is described by the following vector

$$M^{\alpha}_{i;q} = \left(M^{\alpha}_{i1;q}, M^{\alpha}_{i2;q}, ..., M^{\alpha}_{ir;q}\right) ,$$
$$q = 1, ..., n_i \ , \ i = 1, .., k \ , \ 0 < \alpha \leq 1$$

where

$$M^{\alpha}_{ij;q} = \begin{cases} 1 & \text{if } \mu_{ij;q} \geq \alpha \\ 0 & \text{otherwise} \end{cases}$$

If the expert does not hesitate, and its answer is unambiguous, we can describe formally this situation by the following vector:

$$S_{i;q} = (S_{i1;q}, S_{i2;q}, ..., S_{ir;q}) \ ,$$
$$q = 1, ..., n_i \ , \ i = 1, .., k$$

where

$$S_{ij;q} = \begin{cases} 1 & \text{if } \mu_{ij;q} = 1 \text{ and } \sum_{j=1}^{r} \mu_{ij;q} = 1 \\ 0 & \text{otherwise} \end{cases}$$

Hence, we can find the number of unambiguous answers for each pair (x_i, y_j) from the following expression

$$n^0_{ij} = \sum_{q=1}^{n_i} S_{ij;q} \ , \ i = 1, ..., k \ , \ j = 1, ..., r. \quad (6)$$

Now, let us find the number of observations in the ij-th cell of the contingency table for the given α-cut. It can be calculated from the formula

$$n^{\alpha}_{ij} = \sum_{q=1}^{n_i} M^{\alpha}_{ij;q}, \quad i = 1, ..., k \ , \ j = 1, ..., r \ , \quad 0 < \alpha \leq 1. \quad (7)$$

Now, the number of ambiguous observations in the ij-th cell of the contingency table for the given α-cut is given by

$$n^{\alpha,u}_{ij} = n^{\alpha}_{ij} - n^0_{ij} \ , \quad i = 1, ..., k \ , \ j = 1, ..., r \ , \quad 0 < \alpha \leq 1. \quad (8)$$

The total number of ambiguous cases for the i-th level of the variable X is now

$$n^u_i = n_(i) - n^0_i \ , \ i = 1, ..., k \quad (9)$$

where

$$n^0_i = \sum_{j=1}^{r} n^0_{ij} \ , \ i = 1, ..., r \ . \quad (10)$$

Having defined all these quantities we define the chi-square statistics for the case of ambiguous opinions. First, let us introduce the set $\mathcal{M}^{\alpha}$ of auxiliary variables $m^{\alpha}_{ij} \in \{0, 1, ...\}$, $i = 1, ..., k$, $j = 1, ..., r$, $0 < \alpha \leq 1$ such that $0 \leq m^{\alpha}_{ij} \leq n^{\alpha,u}_{ij}$ and $\sum_{j=1}^{r} m^{\alpha}_{ij} = n^u_i$. Let

$$\hat{n}^{\alpha}_{ij} = \frac{n_i}{n} \sum_{w=1}^{k} \left(n^0_{wj} + m^{\alpha}_{wj}\right) \quad (11)$$

and

$$\chi^2_{\alpha} = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{\left(n^0_{ij} + m^{\alpha}_{ij}\right)^2}{\hat{n}^{\alpha}_{ij}} - n \quad (12)$$

Now, for a given α-cut we can find the smallest possible value of the chi-square statistic (12) from the formula

$$\chi^2_{\alpha,\min} = \min_{\mathcal{M}^{\alpha}} \chi^2_{\alpha}, \quad (13)$$

and the largest possible value of the chi-square statistic (12) from the formula

$$\chi^2_{\alpha,\max} = \max_{\mathcal{M}^{\alpha}} \chi^2_{\alpha}. \quad (14)$$

We can consider these two values as the limits of the α-cut interval $\chi^2_{\alpha} = \left[\chi^2_{\alpha,\min}, \chi^2_{\alpha,\max}\right]$ of a fuzzy chi-square statistic $\tilde{\chi}^2_{\alpha}$ whose membership function can be found from the following expression:

$$\mu(\chi^2) = \sup\{\alpha I_{\chi^2_{\alpha}}(\chi^2) : \alpha \in [0, 1]\}, \quad (15)$$

where $I_{\chi^2_{\alpha}}(\chi^2)$ denotes the characteristic function of the set χ^2_{α}.

When the obtained test statistics is fuzzy we can use several methods for the interpretation of test results. The introduction of vagueness to the problem of statistical testing leads to a

new class of statistical tests which have been proposed by many authors such as Casals et al. [3], Kruse and Meyer [9], Watanabe and Imaizumi [12], Römer and Kandel [11], and Grzegorzewski [5]. For deeper discussion and critical review of the problems considered there we refer the reader to the paper by Grzegorzewski and Hryniewicz [6].

Possibilistic approach to testing statistical hypotheses with fuzzy data has been proposed by Hryniewicz [7], [8]. We apply this approach in the considered case of the fuzzy chi-square test of independence. To reject the hypothesis of independence on the significance level δ we have to evaluate the relation $\tilde{\chi}^2_\alpha > \chi^2_{(k-1)(r-1),1-\delta}$. In order to do this we propose to use the concept of possibility indices (see: Dubois and Prade [4]): *necessity of strict dominance* (NSD), and *possibility of strict dominance* PSD).

Possibility of strict dominance index PSD for two fuzzy sets A and B described by their membership functions $\mu_A(x)$ and $\mu_B(y)$, respectively, is defined by the following formula:

$$PSD = Poss(A \succ B) = \sup_x \inf_{y:y\geq x} \min\{\mu_A(x), 1-\mu_B(y)\}. \quad (16)$$

PSD is the measure for a possibility that the set A strictly dominates the set B.

Necessity of strict dominance index is defined as

$$NSD = Ness(A \succ B) = 1 - \sup_{x,y:x\leq y} \min\{\mu_A(x), \mu_B(y)\}. \quad (17)$$

NSD represents a necessity that the set A strictly dominates the set B.

In the considered case of the fuzzy chi-square test of independence we should evaluate the dominance of the fuzzy test statistics $\tilde{\chi}^2_\alpha$ over the crisp value $\chi^2_{(k-1)(r-1),1-\delta}$. In such a case the values of possibility indices can be found straightforwardly. First, let us introduce two sets: $\chi^2_{\alpha,L} = \left[\chi^2_{\alpha,\min}, \infty\right]$ and $\chi^2_{\alpha,R} = \left[0, \chi^2_{\alpha,\max}\right]$. We use these sets to define two membership functions:

$$\mu_L(\chi^2) = \sup\{\alpha I_{\chi^2_{\alpha,L}}(\chi^2) : \alpha \in [0,1]\}, \quad (18)$$

where $I_{\chi^2_{\alpha,L}}(\chi^2)$ denotes the characteristic function of the set $\chi^2_{\alpha,L}$, and

$$\mu_R(\chi^2) = \sup\{\alpha I_{\chi^2_{\alpha,R}}(\chi^2) : \alpha \in [0,1]\}, \quad (19)$$

where $I_{\chi^2_{\alpha,R}}(\chi^2)$ denotes the characteristic function of the set $\chi^2_{\alpha,R}$.

The PSD index is given as

$$PSD = \mu_R(\chi^2_{(k-1)(r-1),1-\delta}). \quad (20)$$

and the NSD index is given as

$$NSD = 1 - \mu_L(\chi^2_{(k-1)(r-1),1-\delta}). \quad (21)$$

There exists a positive necessity of the rejection of the hypothesis of independence when the critical value $\chi^2_{(k-1)(r-1),1-\delta}$ is located to the left of the core of the fuzzy set $\tilde{\chi}^2_\alpha$. If this critical value is situated to the left of the support of the fuzzy set $\tilde{\chi}^2_\alpha$, then the necessity of the rejection of the hypothesis of independence is equal to one. We have a positive possibility of the rejection of the hypothesis of independence when the critical value $\chi^2_{(k-1)(r-1),1-\delta}$ is to the left of or belongs to the core of the fuzzy set $\tilde{\chi}^2_\alpha$.

When the number of explanatory variables has to be limited we should choose those with the strongest correlation with the outcome variable (variable of interest). In order to do so we can use Zadeh's extension principle, and build a fuzzy equivalent of the Tchouproff's index given by (5). Unfortunately, there does not exist a single method for the ordering of fuzzy numbers. Therefore, we need to choose one method of defuzzification, and then to order defuzzified values of the Tchouproff's index, and to choose those explanatory variables with the highest values of the defuzzified Tchouproff's index. The problem of the selection of an appropriate defuzzification method does not seem, however, to be a crucial one. Our methodology should be used only for a *preliminary selection* of possible variables. The final selection has to be done using real statistical data, and it is rather improbable that during the preliminary selection we would omit statistically significant explanatory variables.

4 Conclusions

We have proposed a methodology for testing hypotheses of mutual independence using fuzzy experts opinions about possible values of two variables. By applying this methodology we can eliminate those variables that seem not to influence the variable of interest (outcome variable). Moreover, the proposed test could be used for choosing only those explanatory variables that are mutually independent. It has to be noted, however, that the proposed methodology can be used only for the initial selection of *possible* variables. The final selection of the variables should be done using real statistical data and the appropriate statistical procedures.

The proposed methodology has a serious drawback. Optimisation problems (13) and (14) are not convex, and thus very difficult. If the number of ambiguous answers is relatively small, the optimal solutions could be found by a full search procedure. Otherwise, time-efficient optimisation procedures (e.g. genetic algorithms) may find only local extrema. If we remember, however, that our procedure has to be used only for the preliminary selection of variables, this drawback may not be considered as very serious.

References

[1] Agresti A., Categorical Data Analysis, J.Wiley, New York, 1990.

[2] Bickel P., Doksum K., Mathematical statistics. Basic ideas and selected topics, Holden-Day, Inc., San Francisco, 1977.

[3] Casals R., Gil M.A., Gil P., The fuzzy decision problem: an approach to the problem of testing statistical hypotheses with fuzzy information, *Europ. Journ. of Oper. Res.*, vol.27 (1986), 371–382.

[4] Dubois D., Prade H., Ranking fuzzy numbers in the setting of possibility theory, *Information Sciences*, vol.30 (1983), 184–244.

[5] Grzegorzewski P., Testing statistical hypotheses with vague data, *Fuzzy Sets and Systems*, vol.112 (2000), 501–510.

[6] Grzegorzewski P., Hryniewicz O. Testing hypotheses in fuzzy environment, *Mathware and Soft Computing*, vol.4 (1997), 203–217.

[7] Hryniewicz O., Possibilistic interpretation of fuzzy statistical tests. In: C.Bertoluzza, M.A.Gil, D.A.Ralescu (Eds.):*Statistical Modeling, Analysis and Management of Fuzzy Data*. Physica Verlag, Heidelberg and New York, 2002, 226–238.

[8] Hryniewicz O., Possibilistic decisions and fuzzy statistical tests, *Fuzzy Sets and Systems* (submitted).

[9] Kruse R., Meyer K.D. Statistics with Vague Data, Riedel, Dodrecht, 1987.

[10] Loughin T.M., Scherer P.N. (1998) Testing for Association in Contingency Tables with Multiple Column Responses, *Biometrics*, vol.54, 630 - 637.

[11] Römer Ch., Kandel A., Statistical tests for fuzzy data, *Fuzzy Sets and Systems*, vol.72 (1995), 1–26.

[12] Watanabe N., Imaizumi T. A fuzzy statistical test of fuzzy hypotheses,*Fuzzy Sets and Systems*, vol.53 (1993), 167–178.