Raport Badawczy
Research Report

RB/10/2004

# Goodman-Kruskal $\gamma$ measure of dependence for fuzzy ordered categorical data

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 2004

# Goodman-Kruskal $\gamma$ measure of dependence for fuzzy ordered categorical data

Olgierd Hryniewicz

Systems Research Institute, Newelska 6, 01-447 Warsaw, POLAND

**Summary.** In the paper we the generalisation of Goodman-Kruskal $\gamma$ for the measurement of the strenght of dependence (association) between two categorical variables with ordered categories. We consider the case when some data is not precise, and observation are described by possibility distributions over a set of categories of one variable. For such a data we define the fuzzy $\gamma$ statistic, and present the methodology for the statistical inference related to this fuzzy measure of dependence.

## 1 Introduction

Looking for dependencies is one of the most frequently used applications of statistics. Methods for the statistical analysis of dependencies have been intensively developed for more than one hundred years, especially in economical,agricultural, medical, and social sciences. Recently, these methods are also used in information technology and computer sciences, especially in data mining and knowledge discovery. Traditionally, statistical methods used for the analysis of dependencies are divided into two groups: methods for continuous variables, and methods for categorical data. In both groups there exist many procedures and test for dealing with precise data. However, when data are not precise the situation is not so clear. When we deal with imprecise statistical data measured on the real line the number of available procedures is quite large. For example, there are many papers devoted to the problem of a fuzzy regression, and - to less extend - to the problem of the analysis of correlation. However, when we deal with categorical data the number of papers devoted to this problem is relatively small. This situation is somewhat astonishing, as in real problems we often have to analyse statistical data in the presence of imprecisely defined categories. Consider, for example, the analysis of dependence between health and smoking. Patients filling questionnaires may identify themselves as "nonsmoker", "smoker", and "heavy smoker". Note, that the border between the last two categories is definitely vague. The attempt to clarify this situation by introducing a sharp border (in terms of the number of

cigarettes per day) does not solve the problems, as for many individuals this number may vary in a way which is difficult to describe formally. Therefore, many people may assign for both categories some "weights" making the problem similar to the statistical analysis of multiple answers. We claim, however, that this similarity is highly misleading. In classical problems of the statistical analysis of multiple answers a statistician faces situations when two (or more) categories may occur *simultaneously*. For example, when we ask people for the source of a particular information the may indicate multiple categories as, e.g., television, newspapers, etc. This situation is, in our opinion, entirely different than the problem described above. The lack of precision in the case of imprecisely defined categories is not of a probabilistic nature, and requires the application of a "soft" methodology like the possibility theory or the theory of fuzzy sets.

In a recent paper Hryniewicz [8] considers a fuzzy version of a well known Pearson's chi-square test of independence. The chi-square statistics works very well for that purpose, but is not suitable for the measurement of the strength of dependence, even in its standardised version (indices proposed by Tchouproff and Cramer). The reason for this is its lack of operational meaning. A set of better measure of dependence was proposed in a seminal paper by Goodman and Kruskal [4]. One of the most popular measures of dependence proposed by Goodman and Kruskal is based on a $\gamma$ statistic that was proposed for the analysis of ordered categories. In the second section of this paper we recall its definition and basic properties. Then, in the third section we adopt the similar approach as in Hryniewicz [8], in order to generalise the Goodman-Kruskal statistic $\gamma$ for fuzzy categorical data. Finally, we propose a possibilistic interpretation for statistical tests which are based on the fuzzy $\gamma$ statistic. The paper is the extended verion of the paper of Hryniewicz [9] presented during the SMPS'2004 Conference in Oviedo.

## 2 Measurement of association for ordered categorical data. Goodman-Kruskal's $\gamma$

Consider the situation when we have to measure the strength of dependence (or association) between two categorical variables $X$ and $Y$. Let's assume that $X$ has $k$ categories ordered in the following way $x_1 < x_2 < ... < x_k$, and $Y$ has $r$ categories ordered as follows: $y_1 < y_2 < ... < y_k$ . For a fully multinomial statistical experiment (i.e. when the total number of observation in each category of $X$ and $Y$ is a random variable) the results of the experiment 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 observations in the $ij$-th cell, and

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

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

Let's consider two individual observations of the random vector $(X, Y)$ described as $(x_i, y_j)$ and $(x_m, y_n)$, where $(i \neq m) \vee (j \neq n)$. We call this pair *concordant* if $(i < m) \wedge (j < n)$, and *disconcordant* if either $(i < m) \wedge (j > n)$ or $(i > m) \wedge (j < n)$. Goodman and Kruskal [4] proposed a measure of association called $\gamma$ that is based on two probabilities: of observing a concordant pair of observation $\Pi_c$, and of observing a disconcordant pair of observation $\Pi_d$. To calculate these probabilities let's assume that the contingency table is generated by a multinomial distribution with probabilities given as

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

Goodman and Kruskal have shown that probability $\Pi_c$ is given by

$$\Pi_c = 2 \sum_{i,j} p_{ij} \Pi_{I;ij} \tag{3}$$

where

$$\Pi_{I;ij} = \sum_{i'>i} \sum_{j'>j} p_{i'j'}, \tag{4}$$

and probability $\Pi_d$ is given by

$$\Pi_d = 2 \sum_{i,j} p_{ij} \Pi_{IV;ij}, \tag{5}$$

where

$$\Pi_{IV;ij} = \sum_{i'>i} \sum_{j'<j} p_{i'j'}. \tag{6}$$

The measure of association $\gamma$ was then defined by Goodman and Kruskal [4] as

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}. \tag{7}$$

The value of $\gamma$ belongs to the interval $[-1, 1]$, and when $X$ and $Y$ are independent, we have $\gamma = 0$. When $\gamma < 0$ the considered variables are associated negatively, and when $\gamma > 0$ they are associated positively (Note that the sign of association depends entirely upon the ordering of categories).

Let $G$ be the maximum likelihood estimator of $\gamma$. the formula for $G$ is obtained straightforwardly by inserting in (3) - (7) $n_{ij}/n$ instead of $p_{ij}, i = 1, ..., k$ , $j = 1, ..., r$. It is easy to show that for the calculation of $G$ it suffices to change $p_{ij}$ to $n_{ij}$ in all those formulae. The exact probability distribution of $G$ is very difficult to calculate even when the considered variables are independent. However, for a large number of observations Goodman and Kruskal [5], [6] found an asymptotic distribution of $G$. They showed that $\sqrt{n}(G - \gamma)$ is asymptotically normally distributed with the expected value equal to zero, and the variance given by

$$\sigma^2 = \frac{16}{(\Pi_c + \Pi_d)^4} \sum_{i,j} p_{ij} \left[\Pi_c (\Pi_{III;ij} + \Pi_{II;ij}) - \Pi_d (\Pi_{I;ij} + \Pi_{III;ij})\right]^2 \tag{8}$$

where $\Pi_c$, $\Pi_d$, $\Pi_{I;ij}$, $\Pi_{IV;ij}$ are defined as previously, and

$$\Pi_{II;ij} = \sum_{m<i} \sum_{n>j} p_{mn}, \tag{9}$$

$$\Pi_{III;ij} = \sum_{m<i} \sum_{n<j} p_{mn}. \tag{10}$$

The maximum likelihood estimator $\widehat{\sigma}$ of $\sigma = \sqrt{\sigma^2}$ is obtained by inserting $n_{ij}/n$ (or simply $n_{ij}$) instead of $p_{ij}, i = 1, ..., k\ ,\ j = 1, ..., r$ into respective formulae. Thus, the two-sided asymptotic confidence interval for $\gamma$ (for a given confidence level $\beta$) is given as

$$\left(G - y_{(1+\beta)/2}\widehat{\sigma}/\sqrt{n}\ ,\ G + y_{(1+\beta)/2}\widehat{\sigma}/\sqrt{n}\right) \tag{11}$$

where $y_{(1+\beta)/2}$ is the quantile of the $(1+\beta)/2$ order from the standard normal distribution. This confidence interval can be used for testing the hypothesis that both variables $X$ and $Y$ are mutually independent.

## 3 Goodman-Kruskal's $\gamma$ for fuzzy statistical data

In the classical statistics it is assumed that each statistical datum is given as a pair $(X_q, Y_q), q = 1, .., n$ that defines the observed cell $(i_q, j_q)$ of the contingency table. The value $Z_{ij} \in \{0, 1\}, i = 1, ..., k\ ,\ j = 1, ..., r$ assigned to each cell of the contingency table is equal to 1 when $(i = i_q, j = j_q)$ and 0 otherwise. It means that for $q$-th observation we have only *one* value of $X$ and *one* value of $Y$. This type of statistical data could be generalised to the case of so called multiple responses, when for a given value of $X$ we observe *simultaneously* several values of $Y$ (or vice versa). A good example of multiple responses is when we analyse questionnaires in which people indicate one or *more* sources of particular information from among a set of *different* sources. This generalisation, to the best of our knowledge, has not been considered yet for Goodman-Kruskal's $\gamma$. However, there exists another generalisation of categorical statistical data introduced by Hryniewicz [8] who assumed that for a given value of $X$ we observe a fuzzy value of $Y$ described by a certain *possibility* distribution. To explain this situation let us assume the following experiment. Our aim is to find the measure of association between education and smoking habits. Assume now that smoking habits (variable $Y$) are described by three categories: "non-smoker", "smoker", and "heavy smoker". it seems to be quite obvious that for many smokers it becomes difficult to indicate only *one* value of $Y$ ("smoker" or "heavy smoker"). Note, that this situation is entirely different from that of multiple response data, as it arises rather from a vagueness of these two categories than from their simultaneous "existence". Further explanation of the difference between the considered situation and the case of multiple responses may be the following. In the case of multiple responses we can always create a set of new additional clearly defined categories (being combinations of the existing ones) in order to remove all multiple responses from our data set. This is not possible, however, in the considered case where the vagueness of responses has an intrinsic nature, and cannot be removed without loosing some information.

To simplify the analysis of imprecise (fuzzy) categorical data let us assume that imprecise observations are related only to the values of $Y$, and the observations of $X$ are always crisp. Thus, for a *single* observation the contingency

table may look like (Hryniewicz [8])

| $X/Y$ | $y_1$ | ... | $y_j$ | ... | $y_r$ |
|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... |
| $x_i$ | $\mu_{i1}$ | ... | $\mu_{ij}$ | ... | $\mu_{ir}$ |
| ... | ... | ... | ... | ... | ... |

where $\mu_{ij} \in [0,1]$, $i = 1,...,k$; $j = 1,...,r$ and $\sup_{i,j} \mu_{ij} = 1$. We may interpret the values of $\mu_{ij}$ as the degrees of possibility that for the $X = x_i$ the variable $Y$ adopts the value $y_j$. Now, each observation is described by a pair $\left(X_q, \widetilde{Y}_q\right), q = 1,..,n$, where $X_q$ represents the observed category of the variable $X$, and $\widetilde{Y}_q = y_1|\mu_{i_q 1} + y_2|\mu_{i_q 2} + ... + y_r|\mu_{i_q r}$ is a *fuzzy set* that describes imprecise observation of the variable $Y$. Fuzzy set $\widetilde{Y}_q$ may be interpreted as a *possibility distribution* over the categories of $Y$ for the $q$-th observation.

Let $Y_q^\alpha$ be the $\alpha$-cut $(0 < \alpha \leq 1)$ of $\widetilde{Y}_q$. Thus, $Y_q^\alpha = \left\{M_{i_q 1}^\alpha, M_{i_q 2}^\alpha, ..., M_{i_q r}^\alpha\right\}$, where

$$M_{i_q j}^\alpha = \begin{cases} 1 \text{ if } \mu_{i_q j} \geqslant \alpha \\ 0 \text{ otherwise} \end{cases}, \; i_q \in \{1,...,k\}, \; j = 1,...,r, \; q = 1,...,n,$$

is an ordinary set. Let $Y^\alpha = \{Y_1^\alpha, Y_2^\alpha, ..., Y_n^\alpha\}$ be the set of $\alpha$-cuts for all observations, and $S^\alpha \subseteq Y^\alpha$ be its subset consisting of those elements of $Y^\alpha$ for which $\sum_{j=1}^r M_{i_q j}^\alpha = 1$. It means that for each element of $S^\alpha$ we have only one value 1, and $r-1$ zeros. Now, define $n_{ij,\min}^\alpha$ to be the number of observations in the $(i,j)$-th cell, calculated over all observations that belong to the set $S^\alpha$, and $n_{ij,\max}^\alpha$ to be the number of observations in the $(i,j)$-th cell, calculated over all observations that belong to the set $Y^\alpha$. Having these quantities defined we proceed to the definition of the fuzzy equivalent of Goodman-Kruskal's $\gamma$ measure of dependence (association).

Let's define the following set

$$N^\alpha = \left\{ n_{ij} \; , \; i = 1,...,k \; , \; j = 1,...,r \; : n_{ij,\min}^\alpha \leq n_{ij} \leq n_{ij,\max}^\alpha \; , \sum_{j=1}^r n_{ij} = n_{i\cdot} \right\}$$

To give an interpretation of $N^\alpha$ let's introduce the notion of an observation that for a given $\alpha$-cut is *compatible* with a given fuzzy observation. A *crisp* observation $Z(i_0, j_0)$, such that $\mu_{ij} = 1$ for $(i = i_0) \cap (j = j_0)$ and $\mu_{ij} = 0$ otherwise, is for a given $\alpha$-cut *compatible* with a given fuzzy observation $\widetilde{Y}_q$ $(q = 1,...,n)$ if for a given $i_q = i_0$ we have $M_{i_0 j_0}^\alpha = 1$ . (Note, that the each $\alpha$-cut of $\widetilde{Y}_q$ can be represented as a superposition of all crisp observations that for the given value of $\alpha$ are compatible with $\widetilde{Y}_q$.). Thus, $N^\alpha$ is the set of all possible samples consisted of *crisp* observations that are compatible with the observed fuzzy sample.

Now, using the approach proposed in Hryniewicz [8], we define a fuzzy index of dependence (association) $\widetilde{\gamma}$ by the set of $\alpha$-cuts $(\gamma^{\alpha}_{\min}, \gamma^{\alpha}_{\max})$, where

$$\gamma^{\alpha}_{\min} = \inf_{N^{\alpha}} \gamma \tag{12}$$

and

$$\gamma^{\alpha}_{\max} = \sup_{N^{\alpha}} \gamma. \tag{13}$$

Calculation of (12) and (13) may be a hard computational task. In order ease these computations we will prove the following Lemma.

**Lemma 1.** *Let* $\{n_{ij}, i = 1, ..., k\ ,\ j = 1, ..., r\}$ *describes a set of crisp observations for which the value of Goodman-Kruskal* $\gamma$ *statistic equals* $\gamma_0$. *Moving of one observation from the cell* $(i_0, j_0)$ *to the cell* $(i_0, j_0 - 1)$ *increases (decreases) the value of Goodman-Kruskal* $\gamma$ *statistic if*

$$(DW_{i_0,j_0} - CU_{i_0,j_0}) + (CW_{i_0,j_0-1} - DU_{i_0,j_0-1}) > (<) 0, \tag{14}$$

*where*

$$C = \sum_{i,j} n_{ij} \sum_{i'>i} \sum_{j'>j} n_{i'j'}, \tag{15}$$

$$D = \sum_{i,j} n_{ij} \sum_{i'>i} \sum_{j'<j} n_{i'j'}, \tag{16}$$

$$U_{i_0,j} = \sum_{i=1}^{i_0-1} n_{ij},\ j = 1, ...r, \tag{17}$$

*and*

$$W_{i_0,j} = \sum_{i=i_0+1}^{k} n_{ij},\ j = 1, ...r. \tag{18}$$

*Proof.* Note, that the maximum likelihood estimator of $\gamma$ for a given set of crisp observations is given by

$$\widehat{\gamma} = \frac{C - D}{C + D}. \tag{19}$$

From the structure of $C$ given by (15) we see that by moving one observation from $(i_0, j_0)$ to $(i_0, j_0 - 1)$ we change $C$ by

$$\Delta_C^- = -\sum_{i=1}^{i_0-1} n_{ij_0-1} + \sum_{i'>i_0}\sum_{j'>j_0-1} n_{i'j'} - \sum_{i'>i_0}\sum_{j'>j_0} n_{i'j'} =$$

$$-U_{i_0,j_0-1} + \sum_{i=i_0+1}^{k} n_{ij_0} = -U_{i_0,j_0-1} + W_{i_0j_0}.$$

Similarily, from the structure of $D$ given by (16) we see that by moving one observation from $(i_0, j_0)$ to $(i_0, j_0 - 1)$ we change $D$ by

$$\Delta_D^- = \sum_{i=1}^{i_0-1} n_{ij_0} + \sum_{i'>i_0}\sum_{j'<j_0-1} n_{i'j'} - \sum_{i'>i_0}\sum_{j'<j_0} n_{i'j'} =$$

$$U_{i_0,j_0} - \sum_{i=i_0+1}^{k} n_{ij_0-1} = U_{i_0,j_0} - W_{i_0j_0-1}.$$

Hence, the total change of $\gamma$ is given by

$$\Delta_\gamma^- = \frac{(C-D) + (\Delta_C^- - \Delta_D^-)}{(C+D) + (\Delta_C^- + \Delta_D^-)} - \frac{C-D}{C+D} = \frac{\Delta N}{\Delta D},$$

where

$$\Delta N = 2\left[(DW_{i_0,j_0} - CU_{i_0,j_0}) + (CW_{i_0,j_0-1} - DU_{i_0,j_0-1})\right].$$

It is easy to show that $\Delta D > 0$. Thus, the sign of the change of $\gamma$ is governed by the sign of $\Delta N$, and the appropiate conditions for that are given by (14)□.

**Lemma 2.** *Let* $\{n_{ij}, i = 1, ..., k\ ,\ j = 1, ..., r\}$ *describes a set of crisp observations for which the value of Goodman-Kruskal* $\gamma$ *statistic equals* $\gamma_0$. *Moving of one observation from the cell* $(i_0, j_0)$ *to the cell* $(i_0, j_0 + 1)$ *increases (decreases) the value of Goodman-Kruskal* $\gamma$ *statistic if*

$$(DU_{i_0,j_0} - CW_{i_0,j_0}) + (CU_{i_0,j_0+1} - DW_{i_0,j_0+1}) > (<)\,0, \qquad (20)$$

*where description of* (20) *is given by* (15) – (18).

*Proof.* The proof is similar to that of Lemma 1 □.

**Corollary 1.** *When* $i_0 = 1$, *then by moving observations to the left we increase the value of* $\gamma$.

**Corollary 2.** *When* $i_0 = k$, *then by moving observations to the right we decrease the value of* $\gamma$.

Thus, for the calculation of the maximum (minimum) value of $\gamma$ it is necessary to represent all fuzzy observations from the first (last) row by their leftmost (rightmost) compatible crisp observations. Hence, if $k = 2$ , then the calculation of (12) and (13) is straightforward.

Next four Lemmas state additionally some sufficient conditions that may be useful for the efficient calculations of (12) and (13).

**Lemma 3.** *Let* $\{n_{ij}, i = 1,...,k\ ,\ j = 1,...,r\}$ *describe a set of crisp observations for which the value of Goodman-Kruskal statistic* $\gamma$ *is* $\gamma_0 > 0$. *If the following two conditions are met*

$$(W_{i_0j_0} + W_{i_0j_0-1}) - (U_{i_0j_0} + U_{i_0j_0-1}) > 0, \tag{21}$$

*and*

$$U_{i_0j_0} < W_{i_0j_0-1} \tag{22}$$

*then moving of one observation from the cell* $(i_0, j_0)$ *to the cell* $(i_0, j_0 - 1)$ *increases the value of Goodman-Kruskal* $\gamma$.

*Proof.* Note, that if (22) holds and $\gamma_0 > 0$ (i.e. $C - D > 0$) then

$$\frac{(C-D)+(\Delta_C^- - \Delta_D^-)}{(C+D)+(\Delta_C^- + \Delta_D^-)} = \frac{(C-D)+(\Delta_C^- - \Delta_D^-)}{(C+D)+(\Delta_C^- - \Delta_D^-)+2\Delta_D^-} > \frac{(C-D)+(\Delta_C^- - \Delta_D^-)}{(C+D)+(\Delta_C^- - \Delta_D^-)}$$

Hence, if (21) holds, then

$$\frac{(C-D)+(\Delta_C^- - \Delta_D^-)}{(C+D)+(\Delta_C^- - \Delta_D^-)} > \frac{C-D}{C+D} = \gamma_0 > 0,$$

and this ends the proof.□

**Lemma 4.** *Let* $\{n_{ij}, i = 1,...,k\ ,\ j = 1,...,r\}$ *describe a set of crisp observations for which the value of Goodman-Kruskal statistic* $\gamma$ *is* $\gamma_0 > 0$. *If the following two conditions are met*

$$-(W_{i_0j_0} + W_{i_0j_0+1}) + (U_{i_0j_0} + U_{i_0j_0+1}) > 0, \tag{23}$$

*and*

$$U_{i_0j_0+1} > W_{i_0j_0} \tag{24}$$

*then moving of one observation from the cell* $(i_0, j_0)$ *to the cell* $(i_0, j_0 + 1)$ *increases the value of Goodman-Kruskal* $\gamma$.

*Proof.* The proof is similar to that of Lemma 3 □.

If the conditions of Lemma 3 and Lemma 4 are not fulfilled, the maximisation of $\gamma$ is either impossible or requires recalculation of $C$ and $D$ after each step of the procedure.

**Lemma 5.** *Let* $\{n_{ij}, i = 1, ..., k\ ,\ j = 1, ..., r\}$ *describe a set of crisp observations for which the value of Goodman-Kruskal statistic* $\gamma$ *is* $\gamma_0 < 0$. *If the following two conditions are met*

$$(W_{i_0j_0} + W_{i_0j_0-1}) - (U_{i_0j_0} + U_{i_0j_0-1}) < 0, \tag{25}$$

*and*

$$U_{i_0j_0} < W_{i_0j_0-1} \tag{26}$$

*then moving of one observation from the cell* $(i_0, j_0)$ *to the cell* $(i_0, j_0 - 1)$ *decreases the value of Goodman-Kruskal* $\gamma$.

*Proof.* Note, that if (22) holds and $\gamma_0 < 0$ (i.e. $C - D > 0$) then

$$\frac{(C-D) + (\Delta_C^- - \Delta_D^-)}{(C+D) + (\Delta_C^- + \Delta_D^-)} = \frac{(C-D) + (\Delta_C^- - \Delta_D^-)}{(C+D) + (\Delta_C^- - \Delta_D^-) + 2\Delta_D^-} < \frac{(C-D) + (\Delta_C^- - \Delta_D^-)}{(C+D) + (\Delta_C^- - \Delta_D^-)}$$

Hence, if (25) holds, then

$$\frac{(C-D) + (\Delta_C^- - \Delta_D^-)}{(C+D) + (\Delta_C^- - \Delta_D^-)} < \frac{C-D}{C+D} = \gamma_0 > 0,$$

and this ends the proof.□

**Lemma 6.** *Let* $\{n_{ij}, i = 1, ..., k\ ,\ j = 1, ..., r\}$ *describe a set of crisp observations for which the value of Goodman-Kruskal statistic* $\gamma$ *is* $\gamma_0 < 0$. *If the following two conditions are met*

$$(W_{i_0j_0} + W_{i_0j_0+1}) - (U_{i_0j_0} + U_{i_0j_0+1}) > 0, \tag{27}$$

*and*

$$U_{i_0j_0+1} < W_{i_0j_0} \tag{28}$$

*then moving of one observation from the cell* $(i_0, j_0)$ *to the cell* $(i_0, j_0 + 1)$ *decreases the value of Goodman-Kruskal* $\gamma$.

*Proof.* The proof is similar to that of Lemma 5 □.

If the conditions of Lemma 5 and Lemma 6 are not fulfilled, the minimisation of $\gamma$ is either impossible or requires recalculation of $C$ and $D$ after each step of the procedure.

Practical importance of Lemmas 3 to 6 is obvious. They state conditions whose fulfilment allows to decrease dramatically the number of computations required for the calculations of (12) and (13).

## 4 Statistical inference for fuzzy Goodman-Kruskal $\gamma$.

In the considered case the test statistic is fuzzy, so 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. [1], Kruse and Meyer [10], Watanabe and Imaizumi [12], Römer and Kandel [11], and Grzegorzewski [3]. In this paper we adopt the approach proposed by Grzegorzewski [3], whose methodology follows that of Kruse and Meyer [10], and is based on fuzzy confidence intervals of the considered fuzzy statistic. In the case of the fuzzy index $\widetilde{\gamma}$ we will limit ourselves to the asymptotic case mentioned in the second section of the paper.

For the statistical inference about the fuzzy dependence (association) measure $\widetilde{\gamma}$ we will use its fuzzy confidence interval. In the second section we have recalled Goodman-Kruskal's asymptotic result expressed by (11). A fuzzy version of this interval can be given in terms of its $\alpha$-cuts as follows:

$$(G_l^\alpha, G_u^\alpha), \ 0 < \alpha \leq 1, \tag{29}$$

where

$$G_l^\alpha = \gamma_{\min}^\alpha - y_{(1+\beta)/2}\widetilde{\sigma}_{\min}^\alpha/\sqrt{n}, \ 0 < \alpha \leq 1, \tag{30}$$

and

$$G_u^\alpha = \gamma_{\max}^\alpha + y_{(1+\beta)/2}\widetilde{\sigma}_{\max}^\alpha/\sqrt{n}, \ 0 < \alpha \leq 1. \tag{31}$$

One might assume that $\widetilde{\sigma}_{\min}^\alpha$ and $\widetilde{\sigma}_{\max}^\alpha$ should be calculated independently from $\gamma_{\min}^\alpha$ and $\gamma_{\max}^\alpha$, respectively. We claim, however, that $\widetilde{\gamma}$ and $\widetilde{\sigma}$ are *interconnected*. Therefore, the value of $\widetilde{\sigma}_{\min}^\alpha$ should be calculated from the sample equivalent of (8) using the same set of crisp observations that have been already used for the calculation of $\gamma_{\min}^\alpha$. By analogy, for the calculation of $\widetilde{\sigma}_{\max}^\alpha$ we should use the same set of crisp observations that have been already used for the calculation of $\gamma_{\max}^\alpha$. This claim is in accordance with the intuition that the lower (upper) limit of the confidence interval of $\widetilde{\gamma}$ should be calculated using the same data as the lower (upper) fuzzy assesment of this parameter. Having the fuzzy confidence interval for $\widetilde{\gamma}$ we can use the results of Grzegorzewski [3] for the construction of statistical tests.

Another approach was proposed by Hryniewicz [7] who proposed to look at statistical tests as a procedure for the possibilistic comparison of the fuzzy test statistics and its crisp critical value. We propose to apply this approach in the case of testing hypotheses about $\gamma$ using the fuzzy statistic $\widetilde{\gamma}$. For example, we could test the hypothesis of independence $\gamma = 0$. To reject the hypothesis of independence on the significance level $\delta$ we have to evaluate the relation $\widetilde{\gamma}_s = \sqrt{n}|\widetilde{\gamma}/\widetilde{\sigma}| > y_{1-\delta}$. The membership function of $\widetilde{\gamma}_s$ should be calculated using the same sets of crisp observations which have been used for the calculation of $\gamma^{\alpha}_{\min}$ and $\gamma^{\alpha}_{\max}$, respectively.

In order to make this comparison we propose to use the concept of possibility indices (see: Dubois and Prade [2]): *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:

$$PSD = Poss\,(A > B) = \sup_{x} \inf_{y:y \geq x} \min\{\mu_A(x), 1 - \mu_B(y)\}. \tag{32}$$

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

Necessity of strict dominance index is defined as

$$NSD = Ness(A > B) = 1 - \sup_{x,y:x \leq y} \min\{\mu_A(x), \mu_B(y)\}. \tag{33}$$

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

In the considered case of the fuzzy test of independence based on $\widetilde{\gamma}$ we should evaluate the dominance of the fuzzy test statistics $\widetilde{\gamma}_s = |\widetilde{\gamma}/\widetilde{\sigma}|$ over the crisp value $y_{1-\delta}$. In such a case the values of possibility indices can be found straightforwardly. First, let us introduce two sets: $\gamma^{\alpha}_{s,L} = [\gamma^{\alpha}_{s,\min}, \infty]$ and $\gamma^{\alpha}_{s,R} = [0, \gamma^{\alpha}_{s,\max}]$. We use these sets to define two membership functions:

$$\mu_L(\gamma_s) = \sup\{\alpha I_{\gamma^{\alpha}_{s,L}}(\gamma_s) : \alpha \in [0,1]\}, \tag{34}$$

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

$$\mu_R(\gamma_s) = \sup\{\alpha I_{\gamma^{\alpha}_{s,R}}(\gamma_s) : \alpha \in [0,1]\}, \tag{35}$$

where $I_{\gamma^{\alpha}_{s,R}}$ denotes the characteristic function of the set $\gamma^{\alpha}_{s,R}$.

The $PSD$ index is given as

$$PSD = \mu_R(y_{1-\delta}). \tag{36}$$

and the $NSD$ index is given as

$$NSD = 1 - \mu_L(y_{1-\delta}). \tag{37}$$

There exists a positive necessity of the rejection of the hypothesis of independence when the critical value $y_{1-\delta}$ is located to the left of the core of the fuzzy set $\widetilde{\gamma}_s$. If this critical value is situated to the left of the support of the fuzzy set $\widetilde{\gamma}_s$, 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 $y_{1-\delta}$ is to the left of or belongs to the core of the fuzzy set $\widetilde{\gamma}_s$.

## 5 Numerical example and discussion

To illustrate the theoretical results let's consider a set of (fictive) data which have been collected in order to investigate a possible association between education and smoking habits. The results of the poll are the following:

- 40 persons with the education described as "High School or less" indicated the category "Non-smoker";
- 15 persons with the education described as "High School or less" indicated the category "Smoker";
- 10 persons with the education described as "High School or less" indicated the category "Heavy smoker";
- 30 people with the education described as "University" indicated the category "Non-smoker";
- 8 persons with the education described as "University" indicated the category "Smoker";
- 8 persons with the education described as "University" indicated the category "Heavy smoker".

  Moreover, the some persons presented the following fuzzy responses:
- 5 persons with the education described as "High School or less" presented their indication as 1|"Smoker"+0,5|"Heavy smoker";
- 5 persons with the education described as "High School or less" presented their indication as 0,5|"Smoker"+1|"Heavy smoker";
- 2 persons with the education described as "University" presented their indication as 1|"Smoker"+0,5|"Heavy smoker";
- 2 persons with the education described as "University" presented their indication as 0,5|"Smoker"+1|"Heavy smoker".

  Hence, for the $\alpha$-cut level $\alpha = 1$ we do not observe any fuzziness, and the corresponding contingency table is the following:

| | Non-smoker | Smoker | Heavy smoker | |
|---|---|---|---|---|
| High school or less | 40 | 20 | 15 | 75 |
| University | 30 | 10 | 10 | 50 |
| | 70 | 30 | 25 | 125 |

For the $\alpha$-cut level $\alpha = 0,5$ fuzzy responses are presented by vectors of the form $(0,1,1)$, and the crisp compatible observations are given either as $(0,1,0)$ or as $(0,0,1)$. Following Corollary 1 and Corollary 2 we see that the maximum value of $\gamma$ is attained for such an allocation of these observations that the corresponding contingency table looks like:

| | Non-smoker | Smoker | Heavy smoker | |
|---|---|---|---|---|
| High school or less | 40 | 25 | 10 | 75 |
| University | 30 | 8 | 12 | 50 |
| | 70 | 33 | 22 | 125 |

On the other hand, the minimum value of $\gamma$ is attained for such an allocation of these observations that the corresponding contingency table looks like:

| | Non-smoker | Smoker | Heavy smoker | |
|---|---|---|---|---|
| High school or less | 40 | 15 | 20 | 75 |
| University | 30 | 12 | 8 | 50 |
| | 70 | 27 | 28 | 125 |

For these input data we can calculate the $\alpha$-cut representation of $\tilde{\gamma}$. For $\alpha = 1$ the result is crisp: $\gamma_{\min}^{1,0} = \gamma_{\max}^{1,0} = -0,09091$, and for $\alpha = 0,5$ we have: $\gamma_{\min}^{0,5} = -0,1674$, and $\gamma_{\max}^{0,5} = -0,01345$. The confidence interval (on the confidence level $\beta = 0,9$) for $\alpha = 1$ is $\left(G_l^{1,0} = -0,35677, G_u^{1,0} = 0,174954\right)$, and $\left(G_l^{0,5} = -0,42441, G_u^{0,5} = 0,2560\right)$ for $\alpha = 0,5$. Hence, the data reveals small negative dependence between education and smoking habits (better educated persons smoke less than the other), but this result is not statisticaly very significant as the fuzzy confidence interval contains zero. It is also easy to show that both necessity and possibility to reject the hypothesis of independence (lack of association) on normally used significance levels are equal to zero.

Let us notice, that an opponent of fuzzy statistics might argue that the presented data can be treated using classical methods. A possible argumentation may be the following: persons whose answer was 1|"Smoker"+0,5|"Heavy smoker" may be assigned to a new class "Smoker+", and persons whose answer was 0,5|"Smoker"+0,1|"Heavy smoker" may be assigned to a class, say "Smoker++". In such a case the contingency table might look like

| | Non-smoker | Smoker | Smoker+ | Smoker++ | Heavy smoker | |
|---|---|---|---|---|---|---|
| H. Sch. or less | 40 | 15 | 5 | 5 | 10 | 75 |
| University | 30 | 8 | 2 | 2 | 8 | 50 |
| | 70 | 23 | 7 | 7 | 18 | 125 |

For such crisp data the value of Goodman-Kruskal's $\gamma$ is equal to $-0,0815$, and this value is similar to that obtained using fuzzy methodology. However, the calculation of the asymptotic confidence intervals seems to be not possible, as certain cells contain too small number of observations. Moreover, let

us notice that in this crisp case we tacitly assumed that smoker who were not able to present crisp answers smoke more than ordinary "Smokers". It may not be true if the hesitation results from the impecise description of categories. Therefore, while building a new extended contingency table we introduce some new information which is not necesarily present in the data. Using the fuzzy approach proposed in this paper do not loose imprecision contained in statistical data.

## 6 Conclusions

In the paper we present a new methodology for the assesment of the strength of dependence (association) of two variables described by ordered categorical data when the observation are not crisp. We propose to use a certain possibility distribution over a set of categories of one variable in order to describe imprecise data. For this purpose we use a fuzzy version of the Goodman-Kruskal $\gamma$ statistic. We present also a methodology for the calculation of fuzzy confidence intervals of $\gamma$, and a methodology for testing statistical hypotheses about the values of this measure. The results presented in this paper may be generalised to the case when the data are fuzzy with respect to both variables. However, in such a case a required computational effort (optimisation over a possibly very large set of alternatives) could be rather prohibitive unless we find some hidden structure of the optimisation algorithms.

## References

1. Casals R., Gil M.A., Gil P. (1986), The fuzzy decision problem: an approach to the problem of testing statistical hypotheses with fuzzy information, *Europ. Journ. of Oper. Res.*, vol.27, 371–382.
2. Dubois D., Prade H. (1983), Ranking fuzzy numbers in the setting of possibility theory, *Information Sciences*, vol.30, 184–244.
3. Grzegorzewski P. (2000), Testing statistical hypotheses with vague data, *Fuzzy Sets and Systems*, vol.112, 501–510.
4. Goodman L.A., Kruskal W.H. (1954), Measures of Association for Cross Classifications, *Journ. of the Amer. Stat. Assoc.*, vol.49, 732-764.
5. Goodman L.A., Kruskal W.H. (1963), Measures of Association for Cross Classifications, III: Approximate Sampling Theory, *Journ. of the Amer. Stat. Assoc.*, vol.58, 310-364.
6. Goodman L.A., Kruskal W.H. (1972), Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances. *Journ. of the Amer. Stat. Assoc.*, vol.67, 414-421.
7. Hryniewicz O. (2003), Possibilistic decisions and fuzzy statistical tests, *Fuzzy Sets and Systems* (submitted).
8. Hryniewicz O. (2004), Selection of variables for systems analysis - application of a fuzzy statistical test for independence, Proceedings of IPMU'2004, Perugia, vol.3, 2197–2204.

9. Hryniewicz O. (2004), Measures of dependence for fuzzy ordered categorical data, In: Soft Methodology and Random Information Systems, M.Lopez-Diaz, M.A.Gil, P.Grzegorzewski, O.Hryniewicz, J.Lawry (Eds.), Springer, Berlin, 503–510.
10. Kruse R., Meyer K.D. (1987), Statistics with Vague Data, Riedel, Dodrecht, 1987.
11. Römer Ch., Kandel A. (1995), Statistical tests for fuzzy data, *Fuzzy Sets and Systems*, vol.72, 1–26.
12. Watanabe N., Imaizumi T. (1993), A fuzzy statistical test of fuzzy hypotheses, *Fuzzy Sets and Systems*, vol.53, 167–178.