KYBERNETIKA — VOLUME 47 (2011), NUMBER 1, PAGES 110–122

GOODMAN–KRUSKAL MEASURE OF ASSOCIATION FOR FUZZY–CATEGORIZED VARIABLES

S.M. Taheri and Golamreza Hesamian

The Goodman–Kruskal measure, which is a well-known measure of dependence for contingency tables, is generalized to the case when the variables of interest are categorized by linguistic terms rather than crisp sets. In addition, to test the hypothesis of independence in such contingency tables, a novel method of decision making is developed based on a concept of fuzzy p-value. The applicability of the proposed approach is explained using a numerical example.

Keywords: fuzzy frequency, fuzzy category, fuzzy Goodman–Kruskal statistic, fuzzy p-value, fuzzy significance level, NSD index

Classification: 93E12, 62A10

1. INTRODUCTION

An important class of non-parametric statistical procedures is consists in evaluating the relationship between categorized variables in a two-way contingency table. Classical procedures in these cases are commonly based on crisp (exact/nonfuzzy) categories. In real world problems, however, there are many situations in which categories based on linguistic terms are more realistic and more suitable. These situations are commonly appeared in economic, psychology, sociology, and medical studies. For example, consider a study designed to investigate the relationship between IQ and income level among a certain population. In such case, it is more realistic to categorize the possible amounts of IQ by a fuzzy partition in some linguistic terms, such as: "very low", "low", "medium", "high", and "very high". On the other hand, the more realistic categories of the amount of income should be, say, "low", "moderate", and "high". As another example, consider the relationship between different variables in health sciences. In clinical diagnosis, the performance of all screening tests depends on the cut points used to separate normal and abnormal individuals. The choice of a higher cut point leaves more cases undetected and the choice of a lower cut point classifies more healthy individuals as abnormal. For instance, there are no widely accepted or rigorously validated cut points to define positive screening tests for diabetes in non pregnant adults [5]. The American Diabetes Association (ADA) has recommended plasma glucose cut point of 140 mg and the other researchers have recommended plasma glucose cut point of 120 mg

[22]. However, assuming the same cut point in all researches does not guarantee the crispness of this point. In the other word, in the neighborhood of the cut point, a little increase or decrease in blood plasma glucose can change the individuals status from normal to abnormal or vise versa and this situation is not consistent to clinical rules where the instability of observations is routine [21]. In such case, it is better to categorize the amount of glucose by linguistic terms, such as: "low", "normal", and "high".

But, for study and analyze the above type fuzzy categorized data, it needs to develop soft statistical methods. In this regard, fuzzy set theory provides the suitable framework.

In last decades there have been a lot of attempts to combine statistical methods and fuzzy set theory, in different fields. But, as the authors know, there have been a few works on non-parametric approach in fuzzy environments. Concerning our purposes, we briefly review some of the literature on this topic. Denoeux et al. [3], using the concept of fuzzy partial ordering on closed intervals, extended the non-parametric rank-sum tests for fuzzy data. They introduced the concepts of the fuzzy p-value and the degree of rejection of the null hypothesis quantified by a degree of possibility and a degree of necessity when a given significance level is a crisp number or a fuzzy set. Grzegorzewski [9] introduced a method for inference about the median of a population using fuzzy random variables. Also, he demonstrated a generalization of some classical non-parametric tests for fuzzy random variables [10]. The last work relies on the quasi-ordering based on a metric in the space of fuzzy numbers. Also, he [11] proposed a two-sample fuzzy median test for fuzzy random variables based on the necessity index of strict dominance, suggested by Dubois and Prade [4]. In this manner, he obtained a fuzzy test showing a degree of possibility and a degree of necessity for rejecting the underlying hypothesis. In another work [12], he studied the problem of testing the equality of k-samples against the so-called "simple-tree alternative" by generalizing the two-sample fuzzy median test. Hryniewicz [14] considered a fuzzy version of the well-known Pearson's Chi-Square test of independence in a two-way contingency table. In another work [15], he used the fuzzy version of the Goodman–Kruskal statistic, described by ordered categorical data, in case that imprecise observations are related to the values of response variable while the observations of the explanatory variable are crisp, by proposing a certain possibility distribution over a set of categories of the response variable in order to describe imprecise data. Kahraman et al. [17] proposed some algorithms for fuzzy non-parametric rank-sum tests based on fuzzy random variables. For more on statistical methods with fuzzy observations, the reader is referred to the relevant literature, for example, [18, 20, 26].

In this work, we propose a procedure to extend the Goodman–Kruskal measure to the case when the categories of interest are imprecise rather than crisp. Moreover, we investigate a method of testing hypothesis of independence in contingency tables with fuzzy categories.

This paper is organized as follows: the statement of the main problem of this work is presented in Section 2. In Section 3, we recall some concepts of fuzzy numbers. In Section 4, we investigate a procedure to analyze the contingency tables

with fuzzy categories and, specially, we introduce a generalization of the Goodman–Kruskal statistic to measure the strength of dependence (or association) between two categorized variables of interest. To do this, we develop a method to construct fuzzy cell frequencies, fuzzy Goodman–Kruskal statistic, and fuzzy p-value. To accept or reject the null hypothesis of independence we use NSD index (Necessity degree of Strict Dominance [4, 16]) to compare the fuzzy p-value and fuzzy level of significance. A numerical example is provided to clarify the discussions in this paper, in Section 5. A brief conclusion is provided in Section 6.

2. STATEMENT OF THE MAIN PROBLEM

Suppose we have a random sample of observations as (x_1, y_1), (x_2, y_2) $, \ldots, (x_N, y_N)$, and there are two attributes of interest, say T and S, for each subject in the sample. Suppose there are r categories of the variable T, and c categories of the variable S, and each of N observations (x_i, y_i), $i = 1, 2, \ldots, N$ is classified into exactly one of the rc cross-categories. In a $r \times c$ contingency table, the entry in the (i, j) cell, denoted by f_{ij}, is the number of items having the cross-classification $T = t_i, S = s_j$ as shown in Table 1. The measures of association refer to a wide variety of coefficients that measure the strength of the relationship that has been described in several ways [1, 6]. A well-known measure of association in conjunction with a two-way contingency table, is the Goodman–Kruskal γ measure [1, 7, 15]. The range of γ is $[-1, +1]$, and when T and S are independent, we have $\gamma = 0$. When $\gamma < 0$ the considered variables are associated negatively, and when $\gamma > 0$ they are associated positively. The Goodman–Kruskal statistic (the estimator of γ) is given by $G = \frac{\Pi_C - \Pi_D}{\Pi_C + \Pi_D}$, where,

$$\Pi_C = 2\sum_{i=1}^{r-1}\sum_{j=1}^{c-1} f_{ij}\left(\sum_{s=i+1}^{r}\sum_{t=j+1}^{c} f_{st}\right),\ \Pi_D = 2\sum_{i=1}^{r-1}\sum_{j=2}^{c} f_{ij}\left(\sum_{s=i+1}^{r}\sum_{t=1}^{j-1} f_{st}\right),$$

in which, Π_C and Π_D denote the total number of concordant and discordant pairs of observations, respectively [1]. Consider the problem of testing hypothesis $H_0 : \gamma = 0$ of independence against the alternative hypothesis $H_1 : \gamma \neq 0$ of non-independence. The standardized test statistic is computed as $Z = G/\widehat{\sigma}_G$ which has an asymptotic standard normal distribution under the null hypothesis [2, 8]. The estimation of the variance of G, $\widehat{\sigma}^2_G$, is given by

$$\widehat{\sigma}^2_G = \frac{4}{(\Pi_C + \Pi_D)^2}\left(\sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij}(\pi^C_{ij} - \pi^D_{ij})^2 - (\Pi_C - \Pi_D)^2/N\right),$$

where, $N = \sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij}$,

$$\pi^C_{ij} = \sum_{s=1}^{i-1}\sum_{t=1}^{j-1} f_{st} + \sum_{s=i+1}^{r}\sum_{t=j+1}^{c} f_{st},\ \pi^D_{ij} = \sum_{s=1}^{i-1}\sum_{t=j+1}^{c} f_{st} + \sum_{s=i+1}^{r}\sum_{t=1}^{j-1} f_{st}.$$

Therefore, the hypothesis of independence may be rejected at the level of significance δ if $|\frac{G}{\widehat{\sigma}_G}| > \Phi^{-1}(1 - \frac{\delta}{2})$, where Φ denotes the distribution function of the

Table 1. Two-way contingency table.

	Variable S			
Variable T	s_1	s_2	...	s_c
t_1	f_{11}	f_{12}	...	f_{1c}
t_2	f_{21}	f_{22}	...	f_{2c}
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
t_r	f_{r1}	f_{r2}	...	f_{rc}

standard normal distribution. Then, the suitable test can be represented by

$$\varphi_\delta[(x_1, y_1), \ldots, (x_N, y_N)] = \begin{cases} 1 & p - \text{value} < \delta, \\ 0 & \text{otherwise}, \end{cases}$$

where, $p - \text{value} = 2[1 - \Phi(|\frac{G}{\widehat{\sigma}_G}|)]$.

Note that G and $\widehat{\sigma}_G$ are functions of $f_{11}, f_{12}, \ldots, f_{rc}$, i. e., $G \equiv G(f_{11}, f_{12}, \ldots, f_{rc})$ and $\widehat{\sigma}_G \equiv \widehat{\sigma}_G(f_{11}, f_{12}, \ldots, f_{rc})$.

Now, suppose that the variables of interest are categorized by linguistic terms, in which the boundary of categories are not precise. In specific words, suppose that instead of categories $t_1, t_2, \ldots, t_r$ of the variable T, we have $\widetilde{t}_1, \widetilde{t}_2, \ldots, \widetilde{t}_r$ as fuzzy categories of the possible values of the variable T. The main problem which is studied in this work is to extend the Goodman–Kruskal statistic and introduce a method to test the null hypothesis of independence in such kind of contingency tables. To this end, we use some concepts of fuzzy set theory, which will recall in next section.

3. FUZZY NUMBERS

A fuzzy set $\widetilde{A}$ of the universal set $\mathbf{X}$ is defined by its membership function $\mu_{\widetilde{A}} : \mathbf{X} \to [0, 1]$, with the set $\text{supp}(\widetilde{A}) = \{x \in \mathbf{X} : \mu_{\widetilde{A}}(x) > 0\}$, the support of $\widetilde{A}$. In this work, we consider $\mathbb{R}$ (the real line) as the universal set. We denote by $\widetilde{A}_\alpha$ the α-cut of the fuzzy set $\widetilde{A}$ of $\mathbb{R}$, defined for every $\alpha \in (0, 1]$, by $\widetilde{A}_\alpha = \{x \in \mathbb{R} : \mu_{\widetilde{A}}(x) \geq \alpha\}$, and $\widetilde{A}_0$ is the closure of $\text{supp}(\widetilde{A})$.

The fuzzy set $\widetilde{A}$ of $\mathbb{R}$ is called a fuzzy number if for every $\alpha \in (0, 1]$, the set $\widetilde{A}_\alpha$ is a non-empty compact interval. Such an interval will be denoted by $\widetilde{A}_\alpha = [\widetilde{A}^L_\alpha, \widetilde{A}^U_\alpha]$, where $\widetilde{A}^L_\alpha = \inf\{x : x \in \widetilde{A}_\alpha\}$ and $\widetilde{A}^U_\alpha = \sup\{x : x \in \widetilde{A}_\alpha\}$.

One of the popular forms of a fuzzy number, to be considered in this work, is the so-called trapezoidal fuzzy number $\widetilde{A} = (A^l, A^c, A^s, A^r)_T$ whose membership

function is given by

$$\mu_{\widetilde{A}}(x)=\left\{\begin{array}{ll} 0 & x<A^l,\\ \frac{x-A^l}{A^c-A^l} & A^l\le x<A^c,\\ 1 & A^c\le x<A^s,\\ \frac{A^r-x}{A^r-A^s} & A^s\le x\le A^r,\\ 0 & x>A^r. \end{array}\right\}\forall x\in\mathbb{R}.$$

If $A^c=A^s$, it is called a triangular fuzzy number and is denoted by $\widetilde{A}=(A^l,A^c,A^r)_T$. For more on fuzzy numbers, see, for example, [19].

4. CONTINGENCY TABLE WITH FUZZY CATEGORIES

In this section, we provide an approach for analyzing a two-way contingency table with fuzzy categories $T=\{\widetilde{t}_1,\widetilde{t}_2,\ldots,\widetilde{t}_r\}$ and $S=\{\widetilde{s}_1,\widetilde{s}_2,\ldots,\widetilde{s}_c\}$, based on a sample of crisp observations $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$ (briefly: a two-way contingency table with fuzzy categories). In specific words, we investigate a method of testing independence, and introduce a generalized Goodman–Kruskal measure of association, for such contingency tables.

4.1. Fuzzy frequency

Let us consider an ordinary two-way contingency table with crisp categories $T=\{t_1,t_2,\ldots,t_r\}$ and $S=\{s_1,s_2,\ldots,s_c\}$. Set

$$I_k^i=\left\{\begin{array}{ll}1 & t_i=x_k,\\ 0 & t_i\ne x_k,\end{array}\right.,\quad I_k^j=\left\{\begin{array}{ll}1 & s_j=y_k,\\ 0 & s_j\ne y_k.\end{array}\right.$$

In other words, an observation (x_k,y_k), $k=1,2,\ldots,N$, belongs to the cell ij if $\min\{I_k^i,I_k^j\}=1$. Now, if we want to consider a two-way contingency table with fuzzy categories $T=\{\widetilde{t}_1,\widetilde{t}_2,\ldots,\widetilde{t}_r\}$ and $S=\{\widetilde{s}_1,\widetilde{s}_2,\ldots,\widetilde{s}_c\}$, then it is natural to allocate the observation (x_k,y_k) to the cell ij, at level α, when $\min\{\mu_{\widetilde{t}_i}(x_k),\mu_{\widetilde{s}_j}(y_k)\}\ge\alpha$. So, we can develop a two-way contingency table with fuzzy categories in the following way.

Definition 4.1. In a two-way contingency table with fuzzy categories $T=\{\widetilde{t}_1,\widetilde{t}_2,\ldots,\widetilde{t}_r\}$ and $S=\{\widetilde{s}_1,\widetilde{s}_2,\ldots,\widetilde{s}_c\}$, the fuzzy frequencies $\widetilde{f}_{ij}$, $i=1,2,\ldots,r$, $j=1,2,\ldots,c$, are defined to be the fuzzy sets, with the degree of membership at $f\in\{0,1,\ldots,N\}$ as follows

$$\mu_{\widetilde{f}_{ij}}(f)=\sup\left\{\alpha\in[0,1]:\ \sum_{k=1}^{N}\mathbb{I}(\min\{\mu_{\widetilde{t}_i}(x_k),\mu_{\widetilde{s}_j}(y_k)\}\ge\alpha)=f\right\},$$

where, $\mathbb{I}$ is the indicator function,

$$\mathbb{I}(\rho)=\left\{\begin{array}{ll}1 & \text{if } \rho \text{ is true,}\\ 0 & \text{if } \rho \text{ is false.}\end{array}\right.$$

In the following, we denote by $\widetilde{f}_{ij}[\alpha]$, the α-cuts of the fuzzy frequency $\widetilde{f}_{ij}$, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, c$. In addition, the general element of the set $\widetilde{f}_{ij}[\alpha]$ will be denoted by z_{ij}.

4.2. Fuzzy Goodman–Kruskal statistic

Now, we are going to extend a well-known measure of association to the fuzzy environment in order to evaluate the relationship between underlying variables for contingency tables with fuzzy categories.

Definition 4.2. Consider the assumptions in Definition 4.1. The fuzzy Goodman–Kruskal statistic is defined to be a fuzzy set $\widetilde{G}$ with the following α-cuts

$$\widetilde{G}[\alpha] = [(\widetilde{G})^L_\alpha, (\widetilde{G})^U_\alpha],$$

where,

$$(\widetilde{G})^L_\alpha = \inf\{G(z_{11}, z_{12}, \ldots, z_{rc}) : z_{ij} \in \widetilde{f}_{ij}[\alpha]\},$$

$$(\widetilde{G})^U_\alpha = \sup\{G(z_{11}, z_{12}, \ldots, z_{rc}) : z_{ij} \in \widetilde{f}_{ij}[\alpha]\},$$

and G is defined in Section 2.

Remark 4.3. It is easy to verify that each family of closed intervals $\widetilde{G}[\alpha]$, $\alpha \in (0, 1]$ constitute a fuzzy number on $[-1, 1]$.

Remark 4.4. It should be mentioned that Hryniewicz [15] extended the concept of Goodman–Kruskal statistic to a two -way contingency table, too. He considered the case in which some data are not precise and the categories are crisp. But, in the present work, we consider the case in which data available are crisp and the categories are fuzzy sets.

Remark 4.5. If the fuzzy categories reduce to crisp categories, then the fuzzy frequencies and fuzzy Goodman–Kruskal statistic $\widetilde{G}$ reduce to the classical frequencies and classical Goodman–Kruskal statistic G, respectively.

4.3. Fuzzy p-value

As a consequence of the Goodman–Kruskal statistic being imprecise, the result of a p-value may become imprecise. However, this imprecisely is merely a consequence of the ambiguity of the categories, which is propagated in the calculations. So, by extending the classical p-value, we can develop a concept of fuzzy p-value for evaluating the hypothesis of independence in a two-way contingency table with fuzzy categories, as follows.

Definition 4.6. In the problem of testing independence in a two-way contingency table with fuzzy categories, the fuzzy p-value is defined to be a fuzzy set $\widetilde{p}$-value with the following α-cuts

$$\widetilde{p} - \text{value}[\alpha] = [(\widetilde{p} - \text{value})^L_\alpha, (\widetilde{p} - \text{value})^U_\alpha],$$

where,

$$(\widetilde{p}-\text{value})_{\alpha}^{L}=2\left[1-\Phi\left(\sup\left\{\left|\frac{G(z_{11},z_{12},\ldots,z_{rc})}{\widehat{\sigma}_{G}(z_{11},z_{12},\ldots,z_{rc})}\right|:z_{ij}\in\widetilde{f}_{ij}[\alpha]\right\}\right)\right],$$

$$(\widetilde{p}-\text{value})_{\alpha}^{U}=2\left[1-\Phi\left(\inf\left\{\left|\frac{G(z_{11},z_{12},\ldots,z_{rc})}{\widehat{\sigma}_{G}(z_{11},z_{12},\ldots,z_{rc})}\right|:z_{ij}\in\widetilde{f}_{ij}[\alpha]\right\}\right)\right],$$

in which, G and $\widehat{\sigma}_G$ are defined in Section 2.

Remark 4.7. Based on the Representation Theorem, one can conclude that the sequence of the closed intervals $\widetilde{p}-\text{value}[\alpha]$, $\alpha\in(0,1]$, constitute a fuzzy number on $[0,1]$.

Remark 4.8. If the fuzzy categories reduce to the crisp categories then the fuzzy p-value reduces to the classical p-value.

4.4. Method of decision making

Finally, a decision is made by comparing the observed fuzzy p-value and the given significance level.

Case I) The crisp significance level

When the level of significance is a crisp number as δ, we can define a fuzzy test $\widetilde{\varphi}_{\delta}[(\,x_1,y_1),\ldots,(x_N,y_N)]$ on $\{0,1\}$ as follows

$$\widetilde{\varphi}_{\delta}[(x_1,y_1),\ldots,(x_N,y_N)]=\left\{\frac{1}{\widetilde{\varphi}(1)},\frac{0}{\widetilde{\varphi}(0)}\right\},$$

where $\widetilde{\varphi}(1)=\sup_{p<\delta}\mu_{\widetilde{p}-\text{value}}(p)$ and $\widetilde{\varphi}(0)=\sup_{p\geq\delta}\mu_{\widetilde{p}-\text{value}}(p)$. The quantity $\mu_{\widetilde{\varphi_\delta}}(1)$ may interpret as possibility that the null hypothesis would be rejected. In a similar fashion, the quantity $\mu_{\widetilde{\varphi_\delta}}(0)$ may be interpreted as possibility that H_0 would not be rejected (see also [3]).

Case II) The fuzzy significance level

Since the p-value is defined as a fuzzy set, it is natural to consider the significance level as a fuzzy set, too. In fact, a fuzzy significance level is considered as a fuzzy set on (0,1), [13]. In such a situation, therefore, we need a method for comparing the observed fuzzy p-value and the given fuzzy significance level. There are several ways to carry out this comparison, see, for example, [3, 4, 24, 25]. An especially applicable class of ranking methods between fuzzy numbers are necessity and possibility indices (for more details, see [16]). Here, we recall a definition of a necessity index of strict dominance (NSD index) between fuzzy numbers, suggested by Dubois and Prade [4].

Definition 4.9. For two fuzzy numbers $\widetilde{A}$ and $\widetilde{B}$, we evaluate the degree of necessity to which the relation $\widetilde{A}\succ\widetilde{B}$ is fulfilled by

$$Nec\,(\widetilde{A}\succ\widetilde{B})=1-\sup_{x,y;x\leq y}\min\{\mu_{\widetilde{A}}(x),\mu_{\widetilde{B}}(y)\}.$$

In addition, the degree of possibility to which the relation $\widetilde{A}\preceq\widetilde{B}$ is fulfilled, is defined to be $Pos\,(\widetilde{A}\preceq\widetilde{B})=1-Nec\,(\widetilde{A}\succ\widetilde{B})$.

Table 2. The data set in Example 5.1.

No.	In.	Sat.	No.	In.	Sat.	No.	In.	Sat.	No.	In.	Sat.
1	300	10	2	350	15	3	400	20	4	450	25
5	250	10	6	800	40	7	400	55	8	450	60
9	2500	68	10	350	75	11	850	85	12	1200	10
13	1600	20	14	1800	25	15	900	40	16	1200	52
17	1400	54	18	1600	55	19	1700	56	20	1800	57
21	1900	58	22	2400	66	23	1300	72	24	1500	75
25	1600	77	26	1800	78	27	2600	63	28	1600	92
29	1700	94	30	2500	87	31	3200	25	32	4300	44
33	3100	52	34	3200	55	35	3300	60	36	3500	62
37	3600	64	38	4600	66	39	3200	72	40	3400	74
41	3500	76	42	3700	77	43	3800	78	44	3300	79
45	3600	73	46	4700	86	47	3200	94	48	3500	96
49	3600	95	50	4400	86	51	4700	27	52	5300	53
53	4700	68	54	5200	72	55	5500	74	56	5700	76
57	5600	77	58	5800	78	59	4800	83	60	5300	92
61	5500	95	62	5700	98	63	4900	88	64	700	67
65	400	77	-	-	-	-	-	-	-	-	-

The fuzzy relation *Nec* (NSD index) is antisymmetric and transitive. In fact, it provides a fuzzy partial ordering on $\mathbb{F}(\mathbb{R})$ [4].

We use NSD index because of its appropriate properties and its natural interpretation and effectiveness in applied statistical problems (see [3, 12, 16]).

Finally, one can expect that if the observed fuzzy p-value is less than the given fuzzy significance level $\widetilde{\delta}$, then H_0 (the hypothesis of independence) is rejected, otherwise H_0 is accepted. So, a suitable method for testing independence, can be defined as follows

Definition 4.10. Consider the problem of testing the null hypothesis of independence in a two-way contingency table with fuzzy categories. We define the test of independence as a fuzzy set $\widetilde{\varphi}_{\widetilde{\delta}}\,[(x_1, y_1), \ldots, (x_N, y_N)]$ on $\{0, 1\}$ as follows

$$\widetilde{\varphi}_{\widetilde{\delta}}[(x_1, y_1), \ldots, (x_N, y_N)] = \left\{ \frac{1}{\widetilde{\varphi}(1)}, \frac{0}{\widetilde{\varphi}(0)} \right\},$$

where

$$\widetilde{\varphi}(1) = Nec\,(\widetilde{\delta} \succ \widetilde{p} - \text{value}), \quad \widetilde{\varphi}(0) = 1 - \widetilde{\varphi}(1).$$

In such test, $\widetilde{\varphi}(1)$ is called the necessity degree that H_0 is rejected and $\widetilde{\varphi}(0)$ is called the possibility degree that H_0 is accepted.

5. NUMERICAL EXAMPLE

In this section, a practical example is provided to clarify the proposed method.

Example 5.1. (see also [1], p. 57) A study designed to investigate the relationship between income and job satisfaction among cabmans. Four categories are considered for income level: "low", "moderate", "high", and "very high" and four categories are considered for job satisfaction: "little satisfied", "moderately satisfied", "more or less satisfied", and "satisfied". A random sample of 65 cabmans reported their monthly average income (in $) and their job satisfaction (scaled between 0-100). The results of the collected data are given in Table 2. In this example, therefore, we deal with a two-way contingency table 4×4 with the fuzzy categories $T = \{\widetilde{t}_1 =$ "low", $\widetilde{t}_2 =$ "moderate", $\widetilde{t}_3 =$ "high" , $\widetilde{t}_4 =$ "very high"$\}$ for income, and $S = \{\widetilde{s}_1 =$ "little satisfied" , $\widetilde{s}_2 =$ "moderately satisfied", $\widetilde{s}_2 =$ "more or less satisfied", $\widetilde{s}_4 =$ "satisfied"$\}$ for job satisfaction. The membership functions of the linguistic terms are shown in Figures 1 and 2. To construct the contingency table, first we have to obtain the fuzzy frequency of each cell. For instance, assume that the income and job satisfaction of a person are $ 2500 and 45, respectively. For this case, we have $\mu_{\widetilde{t}_1}(2500) = 0$, $\mu_{\widetilde{t}_2}(2500) = 0.5$, $\mu_{\widetilde{t}_3}(2500) = 0.5$, $\mu_{\widetilde{t}_4}(2500) = 0$, and $\mu_{\widetilde{s}_1}(45) = 0.25$, $\mu_{\widetilde{s}_2}(45) = 0.75$, $\mu_{\widetilde{s}_3}(45) = 0$, $\mu_{\widetilde{s}_4}(45) = 0$. By employing Definition 4.1, the contingency table is obtained as shown in Table 3, where

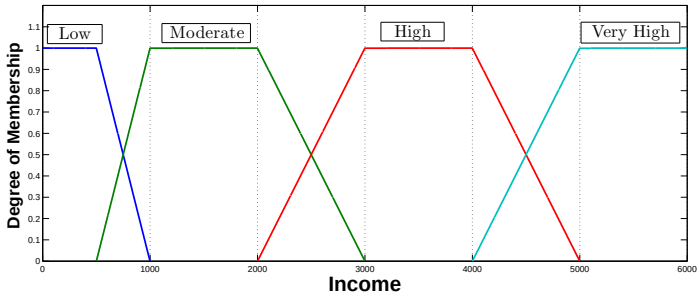

Fig. 1. Fuzzy categories for income in Example 5.1.

Fig. 2. Fuzzy categories for job satisfaction in Example 5.1.

Table 3. The two-way contingency table in Example 5.1.

	Job Satisfaction (S)			
Income (T)	Little	Moderate	More or Less	Satisfied
Low	$\widetilde{f}_{11}$	$\widetilde{f}_{12}$	$\widetilde{f}_{13}$	$\widetilde{f}_{14}$
Moderate	$\widetilde{f}_{21}$	$\widetilde{f}_{22}$	$\widetilde{f}_{23}$	$\widetilde{f}_{24}$
High	$\widetilde{f}_{31}$	$\widetilde{f}_{32}$	$\widetilde{f}_{33}$	$\widetilde{f}_{34}$
Very High	$\widetilde{f}_{41}$	$\widetilde{f}_{42}$	$\widetilde{f}_{43}$	$\widetilde{f}_{44}$

$\tilde{f}_{11} = \left\{\frac{0.2}{7}, \frac{0.4}{6}, \frac{1}{5}\right\}$, $\tilde{f}_{12} = \left\{\frac{0.2}{5}, \frac{0.4}{4}, \frac{0.6}{3}, \frac{1}{2}\right\}$, $\tilde{f}_{13} = \left\{\frac{0.3}{4}, \frac{0.4}{3}, \frac{1}{2}\right\}$,
$\tilde{f}_{14} = \left\{\frac{0.3}{1}, \frac{1}{0}\right\}$, $\tilde{f}_{21} = \left\{\frac{0.5}{5}, \frac{1}{3}\right\}$, $\tilde{f}_{22} = \left\{\frac{0.4}{12}, \frac{0.5}{9}, \frac{0.6}{7}, \frac{1}{6}\right\}$,
$\tilde{f}_{23} = \left\{\frac{0.2}{9}, \frac{0.3}{8}, \frac{0.4}{7}, \frac{0.5}{6}, \frac{1}{4}\right\}$, $\tilde{f}_{24} = \left\{\frac{0.5}{4}, \frac{1}{2}\right\}$, $\tilde{f}_{31} = \left\{\frac{0.3}{3}, \frac{1}{1}\right\}$,
$\tilde{f}_{32} = \left\{\frac{0.3}{11}, \frac{0.4}{10}, \frac{0.69}{6}, \frac{1}{5}\right\}$, $\tilde{f}_{33} = \left\{\frac{0.1}{16}, \frac{0.2}{15}, \frac{0.3}{12}, \frac{0.4}{9}, \frac{0.5}{8}, \frac{1}{7}\right\}$, $\tilde{f}_{34} = \left\{\frac{0.1}{8}, \frac{0.2}{7}, \frac{0.3}{6}, \frac{0.5}{5}, \frac{0.6}{4}, \frac{1}{3}\right\}$,
$\tilde{f}_{41} = \left\{\frac{0.3}{2}, \frac{0.69}{1}, \frac{1}{0}\right\}$, $\tilde{f}_{42} = \left\{\frac{0.3}{4}, \frac{0.4}{3}, \frac{0.6}{2}, \frac{1}{1}\right\}$, $\tilde{f}_{43} = \left\{\frac{0.2}{11}, \frac{0.4}{9}, \frac{0.6}{7}, \frac{0.69}{6}, \frac{1}{5}\right\}$,
$\tilde{f}_{44} = \left\{\frac{0.3}{7}, \frac{0.4}{6}, \frac{0.6}{5}, \frac{0.8}{4}, \frac{1}{3}\right\}$.

Now, suppose that we wish to test the null hypothesis that the income and job satisfaction are independent at significance level "about 0.05" which is represented by the triangular fuzzy number $\widetilde{\delta} = (0.02, 0.05, 0.08)_T$ (Figure 3). Based on the procedures in Section 4, one can obtain the membership function of the fuzzy p-value introduced in Definition 4.6. For example, at level $\alpha = 0.4$, we obtained $\inf\{|\frac{G(z_{11},z_{12},\ldots,z_{44})}{\widetilde{\sigma}_G(z_{11},z_{12},\ldots,z_{44})}| : z_{ij} \in \widetilde{f}_{ij}[0.4]\} \simeq 1.99$ and $\sup\{|\frac{G(z_{11},z_{12},\ldots,z_{44})}{\widetilde{\sigma}_G(z_{11},z_{12},\ldots,z_{44})}| : z_{ij} \in \widetilde{f}_{ij}[0.4]\} \simeq 5.39$, hence, $(\widetilde{p}-\text{value})^L_{0.4} = 1-\Phi(5.39) = 0.7\times 10^{-9}$ and $(\widetilde{p}-\text{value})^U_{0.4} = 1-\Phi(1.99) \simeq 0.04$ (i. e., $\widetilde{p}-\text{value}[0.4] \simeq [0.7\times 10^{-9},\ 0.04]$). By employing this procedure for all α in $[0,1]$, the membership function of fuzzy p-value can be obtained using the Resolution Identity given by

$$\mu_{\widetilde{p}-\text{value}}(x) = \sup_{\alpha\in[0,1]} \alpha \mathbb{I}_{[x\in\widetilde{p}-\text{value}[\alpha]]},\ x \in [0,1].$$

To compute the membership function of the fuzzy p-value, we used the computational procedures included usual accelerator programming with MATLAB software (optimization over a set of alternatives [23]). The membership function of the fuzzy p-value is obtained as shown in Figure 3 (which can be interpreted as: "about 0.129×10^{-5}"). Using Definition 4.9, we obtained $\widetilde{\varphi}(1) = Nec\,(\widetilde{\delta} \succ \widetilde{p}-\text{value}) = 0.6$

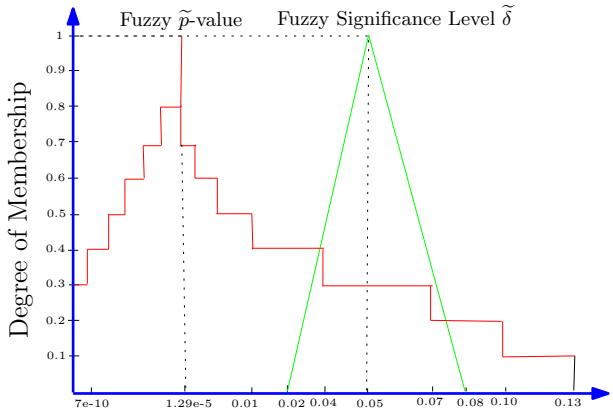

Fig. 3. Fuzzy p-value and fuzzy significance level $\widetilde{\delta}$ in Example 5.1.

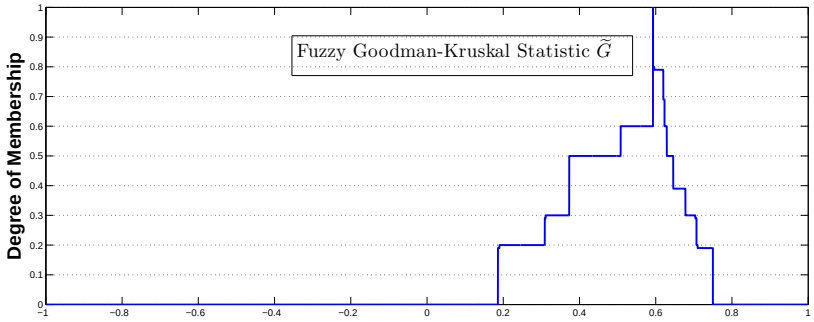

Fig. 4. Fuzzy Goodman–Kruskal statistic $\widetilde{G}$ in Example 5.1.

and $\widetilde{\varphi}(0) = 1 - \widetilde{\varphi}(1) = 0.4$. Therefore, we reject the null hypothesis H_0 with a necessity degree of 0.6 and we accept it with a possibility degree of 0.4.

To obtain the fuzzy Goodman–Kruskal statistic, we need to compute the $\widetilde{G}[\alpha]$, for every $\alpha \in [0,1]$ (Definition 4.2). For example, calculations show that $(\widetilde{G})^L_{0.4} = \inf\{G(z_{11}, z_{12}, \ldots, z_{44}) : z_{ij} \in \widetilde{f}_{ij}[0.4]\} \simeq 0.29$ and $(\widetilde{G})^U_{0.4} = \sup\{G(z_{11}, z_{12}, \ldots, z_{44}) : z_{ij} \in \widetilde{f}_{ij}[\alpha]\} \simeq 0.69$. Finally, the fuzzy Goodman–Kruskal statistic is obtained as shown in Figure 4, which is a representation of "about 0.58".

6. CONCLUSION

We introduced a fuzzy version of the Goodman–Kruskal γ measure of association for a two-way contingency table when the observations are crisp but the categories are described by fuzzy sets. We also developed a method for testing hypothesis of independence in a such two-way contingency table, when level of significance is a crisp or a fuzzy number. For this purpose, we introduced and employed a concept of

fuzzy p-value. To evaluate the independence hypothesis, we use a common index to compare the fuzzy p-value and the given fuzzy significance level. The proposed test, contrary to the classical crisp test, does not lead to a binary decision (acceptance or rejection of the hypothesis) but to a fuzzy decision. In fact, a user must decide whether to reject or to accept the hypothesis of interest actually, but a possibilistic-based index would support his/her decision.

The proposed method can be extended to other measures of associations which are commonly used for two-way contingency tables. The problem of analyzing measures of association in contingency tables for the case where the data available are fuzzy, is a potential subject for future research.

ACKNOWLEDGMENTS

The authors thank referees for their helpful suggestions and comments. The first author is partially supported by the Fuzzy Systems and Its Applications Center of Excellence, Shahid Bahonar University of Kerman, Iran.

(Received April 4, 2010)

REFERENCES

[1] A. Agresti: Categorical Data Analysis. Second Edition. J. Wiley, New York 2002.

[2] M. B. Brown and J. K. Benedetti: Sampling behavior of tests for correlation in two-way contingency tables. J. Amer. Statist. Assoc. *72* (1977), 309–315.

[3] T. Denoeux and M. H. Masson and P. H. Herbert: Non-parametric rank-based statistics and significance tests for fuzzy data. Fuzzy Sets and Systems *153* (2005), 1–28.

[4] D. Dubois and H. Prade: Ranking of fuzzy numbers in the setting of possibility theory. Inform. Sci. *30* (1983), 183–224.

[5] M. M. Engelgau, T. J. Thompson, W. H. Herman, J. P. Boyle, R. E. Aubert, S. J. Kenny, A. Badran, E. S. Sous, and M. A. Ali: Comparison of fasting and 2-hour glucose and HbA1c levels for diagnosing diabetes: diagnostic critera and performance revisited. Diabetes Care *20* (1997), 785–791.

[6] J. D. Gibbons: Nonparametric Measures of Association. Sage Publication, Newbury Park 1993.

[7] L. A. Goodman and W. H. Kruskal: Measures of association for cross classifications. J. Amer. Statist. Assoc. *49* (1954), 732–764.

[8] L. A. Goodman and W. H. Kruskal: Measures of Association for Cross Classifications. Springer, New York 1979.

[9] P. Grzegorzewski: Statistical inference about the median from vague data. Control Cybernet. *27* (1998), 447–464.

[10] P. Grzegorzewski: Distribution-free tests for vague data. In: Soft Methodology and Random Information Systems (M. Lopez-Diaz et al., eds.), Springer, Heidelberg 2004, pp. 495–502.

[11] P. Grzegorzewski: Two-sample median test for vague data. In: Proc. 4th Conf. European Society for Fuzzy Logic and Technology-Eusflat, Barcelona 2005, pp. 621–626.

[12] P. Grzegorzewski: K-sample median test for vague data. Internat. J. Intelligent Systems *24* (2009), 529–539.

[13] M. Holena: Fuzzy hypotheses testing in a framework of fuzzy logic. Fuzzy Sets and Systems *145* (2004), 229–252.

[14] O. Hryniewicz: Selection of variables for systems analysis, application of a fuzzy statistical test for independence. Proc. IPMU, Perugia *3* (2004), 2197–2204.

[15] O. Hryniewicz: Goodman-Kruskal γ measure of dependence for fuzzy ordered categorical data. Comput. Statist. Data Anal. *51* (2006), 323–334.

[16] O. Hryniewicz: Possibilistic decisions and fuzzy statistical tests. Fuzzy Sets and Systems *157* (2006), 2665–2673.

[17] C. Kahranam and C. F. Bozdag, and D. Ruan: Fuzzy sets approaches to statistical parametric and non-parametric tests. Internat. J. Intelligent Systems *19* (2004), 1069–1078.

[18] R. Kruse and K. D. Meyer: Statistics with Vague Data. Reidel Publishing, New York 1987.

[19] K. H. Lee: First Course on Fuzzy Theory and Applications. Springer, Heidelberg 2005.

[20] M. Mareš: Fuzzy data in statistics. Kybernetika *43* (2007), 491–502.

[21] S. Pourahmad, S. M. T. Ayatollahi and S. M. Taheri: Fuzzy logistic regression, a new possibilistic model and its application in clinical diagnosis. Iranian J. Fuzzy Systems, to appear.

[22] B. P. Tabaei and W. H. Herman: A multivariate logistic regression equation to screen for diabetes. Diabetes Care *25* (2002), 1999–2003.

[23] P. Venkataraman: Applied Optimization with MATLAB Programming. J. Wiley, New York 2002.

[24] X. Wang and E. Kerre: Reasonable properties for the ordering of fuzzy quantities (II). Fuzzy Sets and Systems *118* (2001), 387–405.

[25] Y. Yoan: Criteria for evaluating fuzzy ranking methods. Fuzzy Sets and Systems *43* (1991), 139–157.

[26] R. Viertl: Statistical Methods for Non-Precise Data. CRC Press, Boca Raton 1996.

S. M. Taheri, Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 84156-83111, and Department of Statistics, School of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad. Iran.

e-mail: taheri@cc.iut.ac.ir

G. Hesamian, Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 84156-83111. Iran.

e-mail: g.hesamian@math.iut.ac.ir