STATISTICA, anno LXXVI, n. 4, 2016

FUZZY INDICES OF RISK FOR A DEEPER EVALUATION OF EXPOSURE-AFFECTION STUDIES

Maurizio Brizzi [1]
Dipartimento di Scienze Statistiche, Università di Bologna, Bologna, Italia

1. Introduction

Fuzzy logic is based on the possibility to assign the same unit to different groups with a certain degree of pertinence. For example, a spring morning temperature may be considered as "warm" at 60% and "cold" at 40% . A fuzzy approach has been proposed in several fields of statistics, by a huge number of Authors, including Frühwirth-Schnatter (1992), Zadeh (1995), Arnold (1996), Körner (2000), Taheri (2003), Buckley (2005), Falsafain *et al.* (2008), Colubi (2009), Falsafain, Taheri (2011). A detailed review of methods and contributions dealing with statistics and fuzziness has been given by Viertl (2011). In this study the main aim is to propose a fuzzy version of the indices of risk commonly used within exposure-affection studies.

Suppose to focus the attention on the relationship between a specific factor of risk (smoke or drinking habits, environmental pollution, electromagnetic radiations etc.) and a critical event, which could be caused by the factor itself. The critical event is usually the affection by a certain kind of disease (cancer, heart attack, behavioral troubles or other medical or psychological consequences). Supposing to have selected and observed a sample of n statistical units, a simple and widely used way to represent such observations is a 2x2 contingency table, having four cells, each for any possible event (exposed and affected, exposed but not affected, affected but not exposed, neither exposed nor affected). Labeling "Yes" with 1 and "No" with 2, the number of occurrences of each kind of event can be denoted with $n_{hj} (h = 1, 2; j = 1, 2)$. For every cell the theoretical independence frequency can be easily computed; for the first cell the resulting frequency is the following one:

$$n^*_{11} = \frac{(n_{11} + n_{12}) \cdot (n_{11} + n_{21})}{n} \tag{1}$$

while the other theoretical frequencies can be calculated simply by difference, keeping the marginal frequencies as constant. Comparing the observed cell frequencies with the corresponding theoretical ones, four cell components can be defined as follows:

$$c_{hj} = \frac{n_{hj}}{n^*_{hj}}, \quad h = 1, 2; \quad j = 1, 2 \tag{2}$$

[1] Corresponding Author e-mail: maurizio.brizzi@unibo.it

TABLE 1
Contingency table with partial exposure and affection

	Major event	Minor event	no event	Row total
Totally exposed	n_{11}	n_{12}	n_{13}	n_{10}
Partially exposed	n_{21}	n_{22}	n_{23}	n_{20}
Not exposed	n_{31}	n_{32}	n_{33}	n_{30}
Column total	n_{01}	n_{02}	n_{03}	n

The main diagonal cell components c_{11}, c_{22} are directly proportional to the effect of risk factor, and if their value overtakes one there is an increased risk for exposed people. Instead, the cell components c_{12}, c_{21} are inversely proportional to the factor effect. The most known and widely used indices of risk in such a context are the Rate Ratio (RR) and the Odds Ratio (OR), whose statistical properties have been thoroughly analyzed by several Authors, such as Rudas (1998) and Agresti (2007). Both indices can be directly derived from the cell frequencies, and they can be also written as a function of cell components:

$$RR = \frac{n_{11}/(n_{11}+n_{12})}{n_{21}/(n_{21}+n_{22})} = \frac{c_{11}}{c_{21}} \tag{3}$$

$$OR = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{c_{11}\cdot c_{22}}{c_{12}\cdot c_{21}} \tag{4}$$

Looking at (3) and (4), it is evident that the index RR represents the information given by the two cells of the first column, while the index OR summarizes the information of the whole set of four cells. A third index of risk, called Diagonal Ratio (DR) has been proposed by Brizzi (2002, 2004), resuming the effect of the main diagonal cells:

$$DR = c_{11}\cdot c_{22} \tag{5}$$

The three indices DR, RR, OR always lie on the same side with respect to the value 1: if one of them is larger (equal, smaller) than one, all these three indices are larger (equal, smaller) than one. In particular, if the three indices are larger than one, the factor risk may possibly have some effects on the occurrence of the critical event.

2. Fuzzy indices with partial exposure and affection

Suppose now that is possible to distinguish between a total exposure and a partial exposure, which is very common to happen when considering the great majority of factor risks (no smoker, moderate smoker and hard smoker is just a simple example), and that there are two levels of critical event, which could be "minor affection" and "major affection". In this context, the contingency table becomes a 3x3 table, like the following: Considering, under a very simplified fuzzy hypothesis, the partially exposed individuals as "50% exposed and 50% not exposed", and the minor event as "50% event and 50% no event", it is possible to build a new 2x2 contingency table, rescaling the number of observations after

this fuzzy assignment of categories:

$$n_{11}^{'} = n_{11} + \frac{1}{2}n_{12} + \frac{1}{2}n_{21} + \frac{1}{4}n_{22} \tag{6}$$

$$n_{12}^{'} = n_{13} + \frac{1}{2}n_{12} + \frac{1}{2}n_{23} + \frac{1}{4}n_{22} \tag{7}$$

$$n_{21}^{'} = n_{31} + \frac{1}{2}n_{21} + \frac{1}{2}n_{32} + \frac{1}{4}n_{22} \tag{8}$$

$$n_{22}^{'} = n_{33} + \frac{1}{2}n_{23} + \frac{1}{2}n_{32} + \frac{1}{4}n_{22} \tag{9}$$

Once defined this rescaled 2x2 contingency table, it is possible (and useful) to recalculate the above mentioned indices of risk using these "fuzzyfied" frequencies, thus defining the new indices $DR^{'}$, $RR^{'}$ and $OR^{'}$. The result can be interpreted as usual, but here partial exposures and minor events have been taken into account, according to this fuzzy assignment of categories.

Evidently, the above described method can be generalized by assigning to "partial exposure" a certain degree of exposure $\alpha(0 < \alpha < 1)$, not necessarily equal to 1/2, and assigning to the "minor event" its own degree of importance $\beta(0 < \beta < 1)$. Using these degrees of pertinence,α and β, it is possible to calculate the following rescaled frequencies:

$$n_{11}^{'} = n_{11} + \alpha \cdot n_{12} + \beta \cdot n_{21} + \alpha\beta \cdot n_{22} \tag{10}$$

$$n_{12}^{'} = n_{13} + \alpha \cdot n_{23} + (1-\beta) \cdot n_{12} + \alpha(1-\beta) \cdot n_{22} \tag{11}$$

$$n_{21}^{'} = n_{31} + (1-\alpha) \cdot n_{21} + \beta \cdot n_{32} + (1-\alpha)\beta \cdot n_{22} \tag{12}$$

$$n_{22}^{'} = n_{33} + (1-\alpha) \cdot n_{23} + (1-\beta) \cdot n_{32} + (1-\alpha)(1-\beta) \cdot n_{22} \tag{13}$$

As well as before, the indices of risk can be applied to the rescaled frequencies $n_{11}^{'}$, $n_{12}^{'}$, $n_{21}^{'}$, $n_{22}^{'}$; the resulting indices $DR^{'}$, $RR^{'}$ and $OR^{'}$ have the usual interpretation but they include the information given by partial results. The degrees α and β whose values have a primary role in this procedure, have to be decided by using all the information available for the phenomenon; some expert opinion, if available, would be a very useful tool for this kind of choice.

3. Simulation study

Evidently, when a new index is proposed, it is essential to investigate its sample distribution under some particular hypotheses. Here, the sample distribution of $DR^{'}$, $RR^{'}$ and $OR^{'}$ has been simulated by the program GAUSS, for some different sample sizes (from $n = 30$ to $n = 150$), under the simplest choice of degrees of pertinence: $\alpha = \beta = 0.5$. The marginal frequencies were fixed this way:

$$n_{10} = 0.2 \cdot n; \qquad n_{20} = 0.3 \cdot n; \qquad n_{30} = 0.5 \cdot n \tag{14}$$

This simulation has been performed under the hypothesis that risk factor and critical event are totally independent, and for every unit, regardless of the level of exposure, there is a probability 0.1 for "major event", 0.3 for "minor event" and 0.6 for "no event". Under the above described constraints, 200,000 samples

TABLE 2
Sample distribution of the fuzzy risk indices DR', RR', OR' for some values of n

Index	Sample size(n)	Average A(n)	Std.dev SD(n)	Skewness Sk(n)	Kurtosis K(n)	Median Me(n)	Centile 2.5	Centile 97.5
DR',	30	1.0092	0.3022	0.351	3.332	0.9971	0.499	1.616
	50	1.0082	0.2588	0.286	3.287	0.9980	0.533	1.549
	100	1.0044	0.1797	0.197	3.132	0.9990	0.668	1.373
	150	1.0028	0.1455	0.147	3.069	0.9992	0.727	1.298
RR'	30	1.0473	0.4290	1.260	6.189	0.9962	0.434	2.004
	50	1.0419	0.3639	1.030	5.749	0.9974	0.469	1.880
	100	1.0201	0.2423	0.610	3.901	0.9988	0.608	1.557
	150	1.0129	0.1936	0.456	3.475	0.9990	0.673	1.432
OR'	30	1.1024	0.6128	1.857	9.043	0.9947	0.349	2.564
	50	1.0903	0.5156	1.533	7.989	0.9963	0.385	2.352
	100	1.0431	0.3326	0.931	4.745	0.9983	0.528	1.818
	150	1.0277	0.2630	0.711	3.965	0.9987	0.598	1.622

have been simulated, rescaling the frequencies and calculating the fuzzy indices of risk for every sample. In Table 2 some essential statistical features of the resulting sample distribution have been reported, for some different sample sizes. It is quite clear that, for every index, the average tends to one as the sample size increases, and the standard deviation is inversely proportional to n, as well as skewness and kurtosis. The median of the sample distribution is very near to one even for small values of n. It can be noticed that the sample distribution of the index DR' is the less skewed and leptokurtic, thus being the most near to be normal, and this can be an important property when applying procedures of statistical inference. Using the results given in Table 2, it is even possible to define a two-sided statistical test for the null hypothesis H_0 of perfect independence between risk factor and critical event, with a standard significance level of 0.05. Indeed, the tail centiles (2.5 and 97.5) can be taken as limit values for rejecting the null hypothesis.

Some analytical attempts have been done for interpolating the statistics reported in Table 2 as functions of n; in particular, it seems to hold an almost perfect linearity between n and the natural logarithm of average, standard deviation, skewness coefficient, kurtosis and even tail centiles, for every simulated distribution considered (the resulting values of the Bravais-Pearson correlation coefficient r are very close to 1). This allows to extend the results to any value of n between 30 and 150, possibly even outside this interval. These are, specifically, the interpolating functions for average, standard deviation and tail centiles:

Average: $A(n)$
$DR' \to \ln A(n) = 0.0146 - 0.0010\sqrt{n} \quad [r = -0.981].$
$DR' \to \ln A(n) = 0.0751 - 0.0053\sqrt{n} \quad [r = -0.979].$
$DR' \to \ln A(n) = 0.1582 - 0.0111\sqrt{n} \quad [r = -0.980].$

Standard deviation: $SD(n)$
$DR' \to \ln SD(n) = -0.5917 - 0.1105\sqrt{n} \quad [r = -0.998].$
$DR' \to \ln SD(n) = -0.1850 - 0.1207\sqrt{n} \quad [r = -0.997].$
$DR' \to \ln SD(n) = -0.2150 - 0.1287\sqrt{n} \quad [r = -0.996].$

Lower Centile (2.5): $LC(n)$
$DR^{'} \to \ln LC(n) = -1.0166 + 0.0588\sqrt{n} \quad [r = +0.987].$
$DR^{'} \to \ln LC(n) = -1.2090 + 0.0684\sqrt{n} \quad [r = +0.988].$
$DR^{'} \to \ln LC(n) = -1.5102 + 0.0837\sqrt{n} \quad [r = +0.989].$

Upper Centile (97.5): $UC(n)$
$DR^{'} \to \ln UC(n) = +0.6642 - 0.0337\sqrt{n} \quad [r = -0.993].$
$DR^{'} \to \ln UC(n) = +0.9790 - 0.0518\sqrt{n} \quad [r = -0.992].$
$DR^{'} \to \ln UC(n) = +1.3288 - 0.0707\sqrt{n} \quad [r = -0.992].$

4. Generalization of the indices

This method, based on a fuzzy rescaling of the cell frequencies, can be generalized to experimental contexts with a number whatsoever of intermediate levels of exposure between "Totally exposed" (maximum) and "Not exposed" (minimum), and a number whatsoever of intermediate degrees of affection, between "Completely affected" and "Not affected". Suppose to have at disposal a complete scale of evaluation, discrete or continuous, allowing to evaluate the level of exposure α_k $(0 < \alpha_k < 1)$ of the observed individual (or statistical unit) u_k, and to have another scale of values for quantifying the corresponding degree of affection $\beta_k (0 < \beta_k < 1)$. After determining the level of exposure and degree of affection of every observed unit, the sample generates a set of couples of values (α_k, β_k), all belonging to the closed interval [0, 1]. It is then possible to define a 2x2 contingency table whose cell values have been rescaled by using such couples:

$$n_{11}^{'} = \sum_{k=1}^{n} \alpha_k \cdot \beta_k \tag{15}$$

$$n_{12}^{'} = \sum_{k=1}^{n} \alpha_k \cdot (1 - \beta_k) \tag{16}$$

$$n_{21}^{'} = \sum_{k=1}^{n} (1 - \alpha_k) \cdot \beta_k \tag{17}$$

$$n_{22}^{'} = \sum_{k=1}^{n} (1 - \alpha_k) \cdot (1 - \beta_k) \tag{18}$$

The corresponding theoretical cell frequencies are the following:

$$n_{11}^{*} = \frac{\sum_{k=1}^{n} \alpha_k \cdot \sum_{k=1}^{n} \beta_k}{n} \tag{19}$$

$$n_{12}^{*} = \frac{\sum_{k=1}^{n} \alpha_k \cdot \sum_{k=1}^{n} (1 - \beta_k)}{n} \tag{20}$$

$$n_{21}^{*} = \frac{\sum_{k=1}^{n} (1 - \alpha_k) \cdot \sum_{k=1}^{n} \beta_k}{n} \tag{21}$$

$$n_{22}^{*} = \frac{\sum_{k=1}^{n} (1 - \alpha_k) \cdot \sum_{k=1}^{n} (1 - \beta_k)}{n} \tag{22}$$

TABLE 3
Physical activity and perceived health problems in people over 65

Level of activity	Health problems				
	Major	Minor (more)	Minor (one)	No	Total
None	452	606	234	232	1524
Low	87	353	133	118	691
Moderate	89	515	324	237	1165
Intensive	70	438	353	305	1166
	698	1912	1044	892	4546

Dividing each fuzzyfied frequency $n_{hj}^{'}(h = 1, 2; j = 1, 2)$ by the corresponding theoretical frequency n_{hj}^{*}, as done in (2) it is possible to calculate the new cell components $c_{hj}^{'}$:

$$c_{hj}^{'} = \frac{n_{hj}^{'}}{n_{hj}^{*}}, \quad h = 1, 2; \quad j = 1, 2. \tag{23}$$

The fuzzy indices of risk can be then easily calculated:

$$DR' = c_{11}^{'} \cdot c_{22}^{'} \tag{24}$$

$$RR' = c_{11}^{'}/c_{21}^{'} \tag{25}$$

$$OR' = (c_{11}^{'} \cdot c_{22}^{'})/(c_{12}^{'} \cdot c_{21}^{'}) \tag{26}$$

5. Application to an italian survey about self-perceived health

The fuzzy indices $DR^{'}$, $RR^{'}$ and $OR^{'}$ have been applied to the results of an Italian survey, described in Broccoli *et al.* (2005) and Cavrini *et al.* (2005). The sample is composed by 4,546 people aged over 65, male and female, living in the Italian province of Bologna. The factor of risk considered in this survey is the lack of a regular physical activity (four levels, from "no activity" to "intense activity"), while the effect variable is the self-assessed health status (four degrees, from "major problems" to "no problem"). The resulting 4x4 contingency table has been reported in Table 3: Following the generalization just described in the previous section, it is possible to evaluate the qualitative levels of activity and the categories of health problems, assigning the degree 1 to "no activity" (maximum risk exposure) and to "major problem(s)" (maximum affection), and reducing the degrees as the exposure or affection diminishes. The degrees of exposure chosen here are: 1 (no activity), 2/3 (low activity), 1/3 (moderate), 0 (intense). On the other side, the degrees of affection have been fixed this way: 1 (one or more major problems), 1/2 (two or more minor problems), 1/4 (just one minor problem), 0 (no perceived problem). Applying formulas (15) to (18), Table 2 can be rescaled, getting a new 2x2 contingency table, represented in Table 4, jointly with the corresponding independence frequencies.

TABLE 4
Rescaled contingency table for the relationship between physical activity and perceived health problems [corresponding theoretical independence frequencies]

Lack of activity	Health problems Yes	No	Total
Yes	1154 [1000]	1219 [1373]	2373
No	761 [915]	1412 [1258]	2173
	1915	2631	4546

Applying (23) to the rescaled frequencies, it is possible to calculate the cell components:

$c_{11}^{'} = 1154/1000 = 1.154; \quad c_{12}^{'} = 1219/1373 = 0.888;$

$c_{21}^{'} = 761/915 = 0.832; \quad c_{22}^{'} = 1412/1258 = 1.122$

The cell components indicate that there is an increased risk of health problems for people lacking of physical activity, since $c_{11}^{'}$ and $c_{22}^{'}$ are larger than one, while the other components are smaller. Now it is possible to determine the values of the three fuzzy indices of risk considered:

$DR^{'}$=1.154· 1.122=1.295
$RR^{'}$=1.154 / 0.832=1.387
$OR^{'}$=1.154 / 0.832 = 1.387

All these results confirm that physical activity is linked with a better self-perceived health. Evidently, it is worth to consider that self-perceived health status does not necessarily coincide with real health status; this survey was limited to the self-assessment.

6. Concluding remarks

The statistical procedure outlined in this paper possibly opens some new paths in exposure-affection evaluation, but it also leaves some open questions. Surely, a primary problem refers to the assessment of the degrees of pertinence (α and β values), which are essential for a correct interpretation of the results. Only in some selected situations it can be possible to define a perfectly univocal scale of evaluation, and it is most likely to need some expert opinion. It would be useful, when possible, to consult different experts, each proposing his proper assessment, and to join the resulting information (some suitable methods have been propose by Cooke (1991). The determination of the adequate sample size is another crucial point, especially when defining a wide spectrum of modalities of exposure and affection, since it is necessary to have a sufficient sample information for each cell. The effect of sample size on the sample distribution of the indices has been pointed out by sample simulation. Finally, there is an open problem regarding the joint interpretation of the three indices $DR^{'}$, $RR^{'}$ and $OR^{'}$, since

each index gives, from a proper point of view, a measure of the relationship between risk factor and critical event. Therefore, every index is worth of a specific interpretation, even if the Odds Ratio, whose calculation involves all the cell components, can probably be considered, also in this fuzzy version, as the most complete and meaningful among the indices of risk.

References

A. Agresti (2007). *An introduction to categorical data analysis.* Chapter 2. John Wiley & Sons, New York.

B. F. Arnold (1996). *An approach to fuzzy hypothesis testing.* Metrika, 44, pp. 119–126.

M. Brizzi (2002). *The relationship between the relevance quotient and the indices of risk in a 2x2 exposure-affection table.* Methodology and Statistics, Ljubljana 2002, pp. 25–28.

M. Brizzi (2004). *Indices of risk based on relevance quotients, with an application to self-assessed health status.* Atti della XLII Riunione Scientifica SIS, Bari. Contributi spontanei, pp. 345–348.

S. Broccoli, G. Cavrini, M. Zoli (2005). *Il modello di regressione quantile nellanalisi delle determinanti della qualit di vita in una popolazione anziana.* Statistica, LXV, 4, pp. 419–436.

J. J. Buckley (2005). *Fuzzy statistics: hypothesis testing.* Soft Computing, 9, pp. 512–518

G. Cavrini, B. Pacelli, A. Mattivi, G. Bianchi, P. Pandolfi, M. Zoli (2005). *Benefits of physical activities on the perceived health in elderly people.* Proceedings of the 7th European Meeting of EuroQol, Oslo.

A. Colubi (2009). *Statistical inference about the means of fuzzy random variables: Applications to the analysis of fuzzy- and real-valued data.* Fuzzy sets and systems, 160, pp. 344–356.

R. M. Cooke (1991). *Experts in uncertainty. Opinion and subjective probability in science.* Oxford University Press.

A. Falsafain, S. M. Taheri, M. Mashinchi (2008). *Fuzzy estimation of parameters in statistical models.* International Journal of Computational and Mathematical Science, 2, pp. 79-85.

A. Falsafain, S. M. Taheri (2011). *On Buckleys approach to fuzzy estimation.* Soft Computing, 15, pp. 345-349.

S. Frhwirth-Schnatter (1992). *On statistical inference for fuzzy data with applications to descriptive statistics.* Fuzzy sets and systems, 50, pp. 143-165.

R. Krner (2000). *An asymptotic test for the expectation of fuzzy random variables.* Journal of Statistical Planning and Inference, 83, pp. 331-346.

T. Rudas (1998). *Odds Ratios in the Analysis of Contingency Tables*. SAGE Publishing Editor.

S. M. Taheri (2003). *Trends in fuzzy statistics*. Austrian Journal of Statistics, 32, pp. 239–257.

R. Viertl (2011). *Statistical methods for fuzzy data*. John Wiley & Sons, New York.

L. A. Zadeh (1995). *Probability theory and fuzzy logic are complementary rather than competitive*. Technometrics, 37, pp. 271–277.

Summary

In this paper a fuzzy version of three indices of risk (Diagonal Ratio, Rate ratio and Odds ratio), usually applied to exposure-affection studies, has been proposed and developed, considering the presence of a partial level of exposure and/or affection. These fuzzy indices are calculated after rescaling the cell frequencies according to fuzzy degrees of pertinence of partial modalities. A simulation study has then been performed, under the hypothesis of absence of effect of the risk factor, and some exploratory statistics have been reported, corresponding to different sample sizes; a transformed linear interpolation method has been described for extending simulation results. The rescaling method has been generalized, supposing that every observation has its proper level of exposure and affection. Finally, the fuzzy indices have been applied to an Italian survey, dealing with the relationship between physical activity and self-perceived health status of more than 4,500 people over 65, living in the province of Bologna.

Keywords: Exposure-affection study; Indices of risk; Diagonal ratio, Fuzzy logic; Monte Carlo simulation