Book of the Short Papers

Editors: Francesco Maria Chelli, Mariateresa Ciommi, Salvatore Ingrassia, Francesca Mariani, Maria Cristina Recchioni

SIS Società Italiana di Statistica

UNIVERSITÀ POLITECNICA DELLE MARCHE

Dipartimento di Scienze Economiche e Sociali DISES

LIUC Università Cattaneo

LIUC Università Cattaneo | BUSINESS ANALYTICS AND DATA SCIENCE HUB

Pearson

CHAIRS

Salvatore Ingrassia (Chair of the Program Committee) - *Università degli Studi di Catania*
Maria Cristina Recchioni (Chair of the Local Organizing Committee) - *Università Politecnica delle Marche*

PROGRAM COMMITTEE

Salvatore Ingrassia (Chair), Elena Ambrosetti, Antonio Balzanella, Matilde Bini, Annalisa Busetta, Fabio Centofanti, Francesco M. Chelli, Simone Di Zio, Sabrina Giordano, Rosaria Ignaccolo, Filomena Maggino, Stefania Mignani, Lucia Paci, Monica Palma, Emilia Rocco.

LOCAL ORGANIZING COMMITTEE

Maria Cristina Recchioni (Chair), Chiara Capogrossi, Mariateresa Ciommi, Barbara Ermini, Chiara Gigliarano, Riccardo Lucchetti, Francesca Mariani, Gloria Polinesi, Giuseppe Ricciardo Lamonica, Barbara Zagaglia.

ORGANIZERS OF INVITED SESSIONS

Pierfrancesco Alaimo Di Loro, Laura Anderlucci, Luigi Augugliaro, Ilaria Benedetti, Rossella Berni, Mario Bolzan, Silvia Cagnone, Michela Cameletti, Federico Camerlenghi, Gabriella Campolo, Christian Capezza, Carlo Cavicchia, Mariateresa Ciommi, Guido Consonni, Giuseppe Ricciardo Lamonica, Regina Liu, Daniela Marella, Francesca Mariani, Matteo Mazziotta, Stefano Mazzuco, Raya Muttarak, Livia Elisa Ortensi, Edoardo Otranto, Ilaria Prosdocimi, Pasquale Sarnacchiaro, Manuela Stranges, Claudia Tarantola, Isabella Sulis, Roberta Varriale, Rosanna Verde.

FURTHER PEPOPLE OF LOCAL ORGANIZING COMMITTEE

Elisa D'Adamo, Christian Ferretti, Giada Gabbianelli, Elvina Merkaj, Luca Pedini, Alessandro Pionati, Marco Tedeschi, Francesco Valentini, Rostand Arland Yebetchou Tchounkeu

Technical support: Matteo Mercuri, Maila Ragni, Daniele Ripanti

PUBLISHED BY PEARSON
WWW.PEARSON.COM
ISBN 9788891935618AAVV

How to increase the power of the test in sparse contingency tables: a simulation study

Federica Nicolussi[a] and Manuela Cazzaro[b]

[a]Politecnico di Milano; federica.nicolussi@polimi.it
[b]Università di Milano-Bicocca; manuela.cazzaro@unimib.it

Abstract

When analyzing categorical or ordinal data one often comes across sparse contingency tables. One of the biggest problems with this situation is due to the low power of tests for independencies between variables. In this paper, we propose a procedure, based on context-specific independencies, to increase the power of tests. The idea is to focus on sub-tables where the number of null cells is relatively low. In addition, including these kinds of independencies in the Union-Intersection procedure can provide comforting results. These results are shown in a simulative study of different scenarios.

Keywords: Context-specific independencies, sub-tables, union-intersection principle

1. Introduction

The study of relationships between categorical or ordinal variables is often reduced to frequency analysis in contingency tables (see (1)). However, as the number of variables or the number of allowable categories for each variable increases, the corresponding contingency tables become sparse (full of null cells). The main problem when dealing with sparse tables is the low power of the classical statistics and an inaccurate type I error. Many works in the literature have dealt with this topic. For instance, in (7), a study on sparsity, it is showed that Fisher's exact test and the asymptotic X^2 Pearson's test give contradictory results for high levels of sparseness. In (12), it is studied the goodness of fit of the X^2 test, likelihood ratio test G^2, and Cressie-Read statistics. In (6) it is showed that the Gaussian approximation of the likelihood ratio statistic G^2 is more accurate than the χ^2 approximation in sparse contingency tables. Further, the problematic likelihood ratio test's asymptotic properties in sparse tables are studied in (4) and (5).
In the case of sparseness, often, at the exploratory stage, many categories are merged because they are poorly tested in the observed sample. However, this procedure can distort the results obtained by inflating the frequencies of some categories. Moreover, even in the best situations, categorically aggregating leads to a loss of informativeness. In this paper, we propose to take advantage of the study of conditional independencies defined in sub-tables that identify a dense portion of the contingency table related to all the variables under consideration. An example of these relationships is context-specific independencies (CSI) (see (8) and (9)) that study the relationship between two groups of variables in conditional tables. Generally speaking, context-specific independencies are conditional independencies holding for particular values of the variables in the conditioning set. Next, we broaden this concept by considering portions of variables for which independence applies. Subsection 2.1 is devoted to defining these independence relations and the parametric model needed to determine them. Subsection 2.2, on the other hand, explains how the Union-Intersection procedure (see (10) and (11)) can be applied to this type of independence to

extend the previously identified relationships, where possible, over the entire contingency table. Section 3 is devoted to the study of simulations to support the theory presented. Conclusions are reserved for Section 4.

2. Methodology

In the following subsections, we define a new type of independencies defined on contingency subtables. We also see how their use can be employed to have greater power in the likelihood ratio test than the standard test on the whole contingency table.

2.1 Context-specific independencies and their extensions

Let us consider a vector of random variables $X_V = (X_j)_{j \in V}$ where each variable X_j takes value i_j in a set of finite categories $\mathcal{I}_j = (1, \ldots, i_j, \ldots, I_j)$. Let $|\cdot|$ be the cardinality of a set. The contingency table of the $|V|$ variables is defined by $\mathcal{I}_V = \times_{j \in V} \mathcal{I}_j$ where each cell is defined as $\boldsymbol{i}_V = (i_j, j \in V)$. The strictly positive probability associated with any cell $\boldsymbol{i}_V$ is denoted with $\pi(\boldsymbol{i}_V)$. The vector of the probability of the whole contingency table is represented by $\boldsymbol{\pi}$, obtained by stacking each $\pi(\boldsymbol{i}_V)$ in the lexicographical order. Similarly, by considering a subset of variables $X_{\mathcal{M}}$ with $\mathcal{M} \subseteq V$ which generates the marginal $\mathcal{M}$-contingency table $\mathcal{I}_{\mathcal{M}} = \times_{j \in \mathcal{M}} \mathcal{I}_j$, the marginal probability of the generic cell $\boldsymbol{i}_{\mathcal{M}}$ is $\pi(\boldsymbol{i}_{\mathcal{M}})$, obtained by summing with respect to the variables in $X_{V \setminus \mathcal{M}}$. The whole set of these marginal probabilities defined on the $\mathcal{M}$-contingency table $\mathcal{I}_{\mathcal{M}}$, is represented by the vector $\boldsymbol{\pi}_{\mathcal{M}}$.
Given three incompatible subsets of variables X_A, X_B and X_C, a CSI is a independence statement like

$$X_A \perp\!\!\!\perp X_B | (X_C = \boldsymbol{i}'_C), \qquad \boldsymbol{i}'_C \in \mathcal{K}_C, \tag{1}$$

where $\boldsymbol{i}'_C$ is the vector of certain level(s) of the variable(s) in X_C, such that $X_j = i'_j$ for all $j \in C$, and it takes value in the list of levels $\mathcal{K}_C \subseteq \mathcal{I}_C$ for which the independence in formula (1) holds. Here, the table obtained as a cartesian product of $\mathcal{K}_C$ and $\mathcal{I}_{AB}$ is the sub-table where the independence is defined. The following formula is a generalization of the previous CSI where also the first two arguments of the independence statement are constrained to a subtable. Hereafter, we refer to this relationship as sub-CSI:

$$(X_A = \boldsymbol{i}'_A) \perp\!\!\!\perp (X_B = \boldsymbol{i}'_B) | (X_C = \boldsymbol{i}'_C), \qquad (\boldsymbol{i}'_A, \boldsymbol{i}'_B, \boldsymbol{i}'_C) \in \mathcal{K}_A \times \mathcal{K}_B \times \mathcal{K}_C \tag{2}$$

or in short $\boldsymbol{i}'_A \perp\!\!\!\perp \boldsymbol{i}'_B | \boldsymbol{i}'_C$, with $(\boldsymbol{i}'_A, \boldsymbol{i}'_B, \boldsymbol{i}'_C) \in \mathcal{K}_A \times \mathcal{K}_B \times \mathcal{K}_C$. Trivially, by replacing $\mathcal{K}_A$ with $\mathcal{I}_A$ and $\mathcal{K}_B$ with $\mathcal{I}_B$ in formula (2), we easily obtain the CSI in formula (1).
Although the class of sub-CSIs may seem difficult to interpret and of little use, a first advantage lies in the fact that the definition of these statements corresponds to linear constraints on log-linear parameters. For a more comprehensive treatment of the phenomenon, below we take advantage of marginal models, see e.g. (2) which impose constraints on marginal distributions of the tables in order to test different independence hypotheses. More specifically, we focus on hierarchical multinomial marginal (HMM) models, see (8) and (9) for interesting applications. In HMM models, the elements of $\boldsymbol{\eta}$ are the parameters based on different types of logits and defined on marginal distributions. The whole parametrization can be expressed in matrix form as

$$\boldsymbol{\eta} = C \log(M \boldsymbol{\pi}) \tag{3}$$

where C is a contrasts matrix and M is a 1's and 0's matrix which elements provide a suitable sum of probabilities. In general, the vector of parameters associated with the variables in $X_{\mathcal{L}}$ and defined in the marginal table $\mathcal{I}_{\mathcal{M}}$, $\boldsymbol{\eta}^{\mathcal{M}}_{\mathcal{L}} = \left\{\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}})\right\}_{\boldsymbol{i}_{\mathcal{L}} \in (\mathcal{I}_{\mathcal{L}} - \mathbf{1})}$ where $\mathbf{1}$ represents the first (reference) cell used for the *baseline* codification of the parameters. The above parameters are contrasts of logarithms of sums of probabilities.

Theorem 1. *Let us consider a set of variables X_V, with probability distribution $\mathcal{P}$ parametrized through the parameters in formula (3), where the baseline criterion is used. Then, the probability distribution*

$\mathcal{P}$ obeys the sub-CSI in formula (2) if and only if the following constraints on the HMM parameters are satisfied:

$$\sum_{c \subseteq C} \eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = 0 \qquad \textit{where } \mathcal{L} = a \cup b \cup c \textit{ and } \boldsymbol{i}_{\mathcal{L}} \in \mathcal{K}_a \times \mathcal{K}_b \times \mathcal{K}_c, \tag{4}$$

where $a \cap A \neq \emptyset$ and $b \cap B \neq \emptyset$, $\mathcal{K}_a \subsetneq (\mathcal{I}_a - \mathbf{1}_a)$ and $\mathcal{K}_b \subsetneq (\mathcal{I}_b - \mathbf{1}_b)$.

The following example shows how to apply formula (4).

Example 1 In order to test $\boldsymbol{i}'_A \perp\!\!\!\perp \boldsymbol{i}'_B | \boldsymbol{i}'_C$, we need to impose $\eta_{AB}^{ABC}(i'_A, i'_B) + \eta_{ABC}^{ABC}(i'_A, i'_B, i'_C) = 0$. If the constrain is satisfied for $i'_A i'_B \in \mathcal{I}_{ab} - \mathbf{1}$ (all the categories of the variables A and B except the reference category) then the global conditional hypothesis $H_0 : A \perp\!\!\!\perp B | C$ is satisfied.

Before proceeding, two clarifications should be made. First, sub-CSI in formulation (2) differs from CSI in formulation (1) if $\mathcal{K}_{\mathcal{A}}$ and $\mathcal{K}_B$ are proper subsets of $\mathcal{I}_A - \mathbf{1}_A$ and $\mathcal{I}_B - \mathbf{1}_B$. In fact, by construction, the parameter sets associated with the reference cell are worth zero. So if we impose the constraints in formula (4) for all cells except the reference cell, we automatically have that it is also satisfied for the reference cell. The second clarification always concerns the reference categories of the variables. Since the value of the parameters is zero at these categories, it is difficult to discriminate whether the independence statement (2) also involves the reference cell. However, the choice of the reference cell, which by default is defined as the first cell of the contingency table, can be defined as desired without changing the truthfulness of the constraints.

2.2 Union-Intersection principle

The Union-Intersection (UI) principle has been proposed by (10) and (11). In nutshell, the UI test states that we can express a (global) null hypothesis H_0 as the intersection of k several component hypotheses H_{0_i} and a (global) alternative as the union of the k component alternatives $\overline{H}_{0_i}$:

$$H_0 : \bigcap_{i=1}^{k} H_{0_i} \quad H_1 : \bigcup_{i=1}^{k} \overline{H}_{0_i}. \tag{5}$$

We reject the global null hypothesis if any of the tests on the component hypotheses lead to rejection, and we retain the global null hypothesis if none of the component tests leads to rejection.
Note that to test any H_{0_i} it is possible to choose a different level α_i. When H_0 holds it is desirable that the rejection probability of the UI test is not greater than a fixed level α^*. This is ensured if the levels α_i of the component tests are such that $\sum_{i=1}^{k} \alpha_i = \alpha^*$. Bonferroni's correction is a popular choice to achieve this: $\alpha_i = \frac{\alpha^*}{k}$. Generally, the rejection probability of the UI test is not less than the rejection probability of any component test. Thus, when H_0 does not hold, the rejection probability provides the power of the test and, in this case, the global test's power is greater than or equal to that of the component test with the highest power.
In our context, the Union-Intersection principle may offer advantages when the component hypotheses are defined on lower dimensional sub-tables less affected by sparsity than the whole table on which H_0 is defined. Reasoning in terms of sub-CSIs, we can divide hypothesis testing for conditional independence as a set of hypotheses about sub-CSIs, such that the unified support of the variables in all individual tests covers the entire support. For greater clarity, consider the following example.

Example 2 Let us suppose to have a contingency table involving three variables, X_1, X_2 and X_3. Let the suitable $(\mathcal{K}_1 \times \mathcal{K}_2 \times \mathcal{K}_3)$ be dense, while the remaining sub-tables, defined by $(\bar{\mathcal{K}}_1 \times \mathcal{I}_2 \times \mathcal{I}_3)$, $(\mathcal{K}_1 \times \bar{\mathcal{K}}_2 \times \mathcal{I}_3)$, and $(\mathcal{K}_1 \times \mathcal{K}_2 \times \bar{\mathcal{K}}_3)$ are more or less sparse. Here the symbol $\bar{\mathcal{K}}$ denotes the complementary set of $\mathcal{K}$. We can think of H_0 as the conditional independence statement $X_1 \perp\!\!\!\perp X_2 | X_3$. The component hypotheses H_{0_i} can be seen as sub-CSI independence statements $\boldsymbol{i}'_1 \perp\!\!\!\perp \boldsymbol{i}'_2 | \boldsymbol{i}'_3$, where the list of admissible values for $\boldsymbol{i}_{\mathcal{L}} = (\boldsymbol{i}'_1, \boldsymbol{i}'_2, \boldsymbol{i}'_3)$ discriminates between the hypothesis. In detail, we have

that $\boldsymbol{i}_{\mathcal{L}} \in (\mathcal{K}_1 \times \mathcal{K}_2 \times \mathcal{K}_3)$ for H_{0_1}, $\boldsymbol{i}_{\mathcal{L}} \in (\bar{\mathcal{K}}_1 \times \mathcal{I}_2 \times \mathcal{I}_3)$ for H_{0_2}, $\boldsymbol{i}_{\mathcal{L}} \in (\mathcal{K}_1 \times \bar{\mathcal{K}}_2 \times \mathcal{I}_3)$ for H_{0_3} and $\boldsymbol{i}_{\mathcal{L}} \in (\mathcal{K}_1 \times \mathcal{K}_2 \times \bar{\mathcal{K}}_3)$ for H_{0_4}.

The truthfulness of the union of all these hypotheses implies the truthfulness of the global one. The idea is to split the global hypothesis into several sub-CSIs. If we retain all of them then the global conditional hypothesis is satisfied. One way to proceed could be as follows. An exploratory survey is carried out in the whole contingency table to see where the table is most sparse. This would identify blocks in the global table: the dense block, the sparse block, and the middle ground tables. One of the component hypotheses H_{0_i} is constructed in the densest sub-table. If in general some of the H_{0_i} on the sparse sub-tables leads to the rejection of the global hypothesis but the H_{0_i} hypothesis on the dense sub-table is instead in favour of H_0, this provides interesting information on the sub-CSI between the variables involved in the contingency table.

3. Simulation study of the power

In this section, we want to show some preliminary results obtained through simulations as evidence to support the proposed methodology. We considered a simple case where we have 3 variables: $X_1 \in (1, \ldots, 5)$, $X_2 \in (1, \ldots, 5)$ and $X_3 \in (1, \ldots, 5)$. We want to investigate the global hypothesis $H_0 : X_1 \perp\!\!\!\perp X_2 | X_3$ for $\boldsymbol{X} = (X_1, X_2, X_3)$ taking values in $\mathcal{I}_{123} = (1, \ldots, 5)^3$, against the alternative.
In order to perform the Union-Intersection procedure, we divided the contingency table into 4 sub-tables.

- $\mathcal{K}^I = (1, 2) \times (1, 2) \times (1, 2)$, 8 cells;
- $\mathcal{K}^{II} = (3, 4, 5) \times (1, 2, 3, 4, 5) \times (1, 2, 3, 4, 5)$, 75 cells;
- $\mathcal{K}^{III} = (1, 2) \times (3, 4, 5) \times (1, 2, 3, 4, 5)$, 30 cells;
- $\mathcal{K}^{IV} = (1, 2) \times (1, 2) \times (3, 4, 5)$, 12 cells.

Then, we define 4 component hypotheses H_{0_i} to test H_0 in the Union-Intersection procedure as sub-CSI. The null component hypotheses are H_{0_1}: $i_1 \perp\!\!\!\perp i_2 | i_3$ for $\boldsymbol{i} \in \mathcal{K}^I$, H_{0_2}: $i_1 \perp\!\!\!\perp i_2 | i_3$ for $\boldsymbol{i} \in \mathcal{K}^{II}$, H_{0_3}: $i_1 \perp\!\!\!\perp i_2 | i_3$ for $\boldsymbol{i} \in \mathcal{K}^{III}$ and H_{0_4}: $i_1 \perp\!\!\!\perp i_2 | i_3$ for $\boldsymbol{i} \in \mathcal{K}^{IV}$. Since the degree of freedom of the χ^2 for the distribution of the likelihood test is not accurate when the sparseness occurs, we simulated $m = 10000$ samples and we evaluated the MC distribution of the G^2 statistics of the likelihood ratio test as follows.
First, we build a probability distribution P_0 where the independence in H_0 holds. We use that distribution to simulate $m = 10000$ frequency tables under H_0. In particular, we impose that the sub-table A must be dense with 50 observations on 8 cells; the remaining three sub-tables are sparse with $57 \sim 75/4 * 3$ observations for B, $20 \sim 30/3 * 2$ observations for C and $6 \sim 12/2$ observations for D.
Then, we evaluate the statistic G^2 of the LR test in the full table and in each sub-tables of each sample, obtaining the MC distribution under H_0 for $\boldsymbol{G}^2_{\mathcal{I}_{123}}$, $\boldsymbol{G}^2_{\mathcal{K}^I}$, $\boldsymbol{G}^2_{\mathcal{K}^{II}}$, $\boldsymbol{G}^2_{\mathcal{K}^{III}}$, and $\boldsymbol{G}^2_{\mathcal{K}^{IV}}$, respectively.
On the previous distributions, we evaluated the simulated critical values for the tests as the quantile of order $1 - \alpha = 0.95$ ($X^2_{H_0,\alpha}$), for the global hypothesis and as the quantile of order $1 - \alpha_1$, $1 - \alpha_2$, $1 - \alpha_3$, and $1 - \alpha_4$ for the component hypothesis $X^2_{H_{0_1},\alpha_1}$, $X^2_{H_{0_2},\alpha_2}$, $X^2_{H_{0_3},\alpha_3}$, and $X^2_{H_{0_4},\alpha_4}$. The only constraint on the significance level of the component hypotheses is that their sum must be equal to α. The analysis is led for different values of significant levels in order to investigate how it is possible to gain more power.
Further, we proceed in the same way with the generation of $m = 10000$ samples from an alternative distribution, where the global hypothesis does not hold.
In order to cover all possible scenarios we build different probability distributions under H_1:

- P_1 where only the sub-CSI H_{0_1} holds (the dense table);
- P_2 where the sub-CSI H_{0_1} does not hold but the others (in the sparse tables) hold.
- P_3 where all the sub-CSIs do not hold.

In any of the above scenarios, we evaluated the rejection rate as an estimator of the test power. The results were collected in the following subsection.
All the analyses were carried out with the software R and the package `hmmm` (3) for the estimation of parameters.

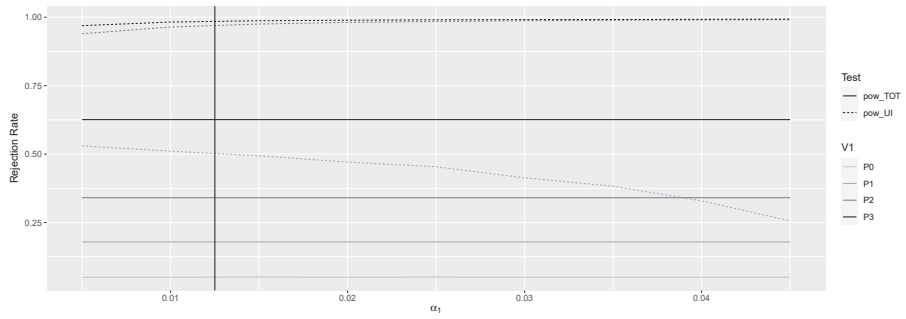

Figure 1: Simulated rejection rate distribution (dashed line) related to the UI procedure by increasing the size α_1 of test H_{0_1} using the alternative distribution P_3. In solid lines are the corresponding (constant) simulated rejection rates for the global test.

3.1 Results description

Table 1 reports the rejection rate in all the scenarios detailed above and also for the case where the alternative distribution satisfies H_0, in order to provide information on the significant level of the Union-Intersection test. It can be seen from the results in Table 1 that the test based on the UI procedure is always more powerful than the test based on the global hypothesis. In particular, the power of the test through the UI procedure gains a lot of power when the null hypothesis does not apply in the dense sub-table (P_2 and P_3). However, due to the segmentation of the entire contingency table, even in the case of P_1 the UI procedure outperforms the test conducted on the entire contingency table. Looking at the row referring to the P_3 distribution, where in each sub-table independence does not hold, it is evident how the power increases in the dense situation.

Table 1: Rejection rates for all possible scenarios evaluated on the distributions P_1, P_2, P_3 and P_0, respectively. The shaded cells show the simulated levels of the tests and non-shaded cells show the simulated power.

n	m	$r(H_{0_1}, \alpha_1)$	$r(H_{0_2}, \alpha_2)$	$r(H_{0_3}, \alpha_3)$	$r(H_{0_4}, \alpha_4)$	$r(H_0, \boldsymbol{\alpha})_{T_{UI}}$	$r(H_0, \alpha^*)_{T_0}$
$P_1 : i_1 \not\perp\!\!\!\perp i_2 \vert i_3$ for $\boldsymbol{i} \in \mathcal{K}^{II}$ and $\boldsymbol{i} \in \mathcal{K}^{III}$ and $\boldsymbol{i} \in \mathcal{K}^{IV}$						$\rightarrow \boldsymbol{H_0}$ ***rejected***	
133	10000	0.0133	0.3691	0.1400	0.0657	0.4996	0.1792
$P_2 : i_1 \not\perp\!\!\!\perp i_2 \vert i_3$ for $\boldsymbol{i} \in \mathcal{K}^{I}$						$\rightarrow \boldsymbol{H_0}$ ***rejected***	
133	10000	0.9698	0.0114	0.0112	0.0164	0.9706	0.3406
$P_3 : i_1 \not\perp\!\!\!\perp i_2 \vert i_3$ for $\boldsymbol{i} \in \mathcal{K}^{I}$ and $\boldsymbol{i} \in \mathcal{K}^{II}$ and $\boldsymbol{i} \in \mathcal{K}^{III}$ and $\boldsymbol{i} \in \mathcal{K}^{IV}$ $\rightarrow \boldsymbol{H_0}$ ***rejected***							
133	10000	0.9690	0.3689	0.1400	0.0657	0.9844	0.6259
$P_0 : i_1 \perp\!\!\!\perp i_2 \vert i_3$ for $\boldsymbol{i} \in \mathcal{K}^{I}$ and $\boldsymbol{i} \in \mathcal{K}^{II}$ and $\boldsymbol{i} \in \mathcal{K}^{III}$ and $\boldsymbol{i} \in \mathcal{K}^{IV}$ $\rightarrow \boldsymbol{H_0}$ ***not rejected***							
133	10000	0.0125	0.0125	0.0125	0.0151	0.0516	0.0500

n: number of observations; m: number of simulated elements in the MC distributions; $r(H, \alpha)$: rejection rate of the component tests H_{0_1}, H_{0_2}, H_{0_3}, H_{0_4}, and H_0, with test size equal to $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.0125$, $\boldsymbol{\alpha}^T = (0.0125, 0.0125, 0.0125, 0.0125)$, and $\alpha^* = 0.05$.

Further, Figure 1 shows how the power of the UI test changes by varying the significant levels of the component hypothesis and by varying the different degree of connection between the two conditional distribution of X_1 and X_2 given X_3. Note that the scenario reported in Table 1 corresponds to the vertical line in $\alpha = 0.125$. Looking at the solid lines, Figure 1 reports the rejection rates of the different tests. The red line belongs to the global test, the other lines represent the (greater) power of the other tests. The greater power is related to the test evaluated on the P_3 distribution where none of the sub-CSIs holds. Concerning the dashed lines, it is worthwhile to note that the power of the tests evaluated on P_2, P_3 distributions is increasing and always increases the power of the overall test. In contrast, this trend is decreasing with respect to the P_1 distribution, where independence does not hold in the densest table and holds in all others. Moreover, for the different levels of α_1, the red dashed line traces the corresponding value of the classical test.

4. Conclusion and further research

This study proposes a method to increase the power of tests performed in sparse contingency tables. The idea is based on the logic of the Union Intersection procedure to decompose the null hypothesis H_0. The original proposal is to consider a set of context-specific hypotheses. This type of hypothesis focuses on sub-spaces of variables. By identifying the least sparse ones, a discrete increase in power can be achieved. The work is still preliminary, but early results from simulations give promising results in terms of power. Clearly, it is necessary to extend the simulations already carried out for different levels of n, m and different scenarios among α_i, for $i = 1, \ldots, k$.

References

[1] Agresti, A.: Categorical Data Analysis - Third Edition. Wiley, Hoboken (2013)

[2] Bergsma, W. P.,& Rudas, T.: Marginal models for categorical data. Ann. Stat. **30(1)**, 140-159 (2002).

[3] Colombi, R., Giordano, S., Cazzaro, M.: hmmm: An R Package for Hierarchical Multinomial Marginal Models. J. Stat. Softw. **59**, 1–25 (2014)

[4] Dale, J.R.: Asymptotic normality of goodness-of-fit statistics for sparse product multinomials. J. R. Stat. Soc. Series B Stat. Methodol. **48**, 48–59 (1986)

[5] Fienberg, S.E., Rinaldo, A.: Maximum likelihood estimation in log-linear models. Ann. Stat. **40**, 996–1023 (2012)

[6] Koehler, K.J.: Goodness-of-fit tests for log-linear models in sparse contingency tables. J. Am. Stat. Assoc. **81**, 483–493 (1986)

[7] Mehta, C.R., Patel, N.R.: A network algorithm for performing Fisher's exact test in r×c contingency tables. J. Am. Stat. Assoc. **78**, 427–434 (1983)

[8] Nicolussi, F., Cazzaro, M.: Context-specific independencies in hierarchical multinomial marginal models. Stat. Methods Appt. **29**, 767–786 (2020)

[9] Nicolussi, F., Cazzaro, M.: Context-Specific Independencies in Stratified Chain Regression Graphical Models. Bernoulli **27**, 2091–2116 (2021)

[10] Roy, S.N.: On a heuristic method of test construction and its use in multivariate analysis. Ann. Math. Stat. **24**, 220–238 (1953)

[11] Roy, S.N., Mitra, S.K.: An introduction to some non-parametric generalizations of analysis of variance and multivariate analysis. Biometrika. **43**, 361–376 (1956)

[12] Rudas, T.: A Monte Carlo comparison of the small sample behaviour of the Pearson, the likelihood ratio and the Cressie-Read statistics. J. Stat. Comput. Simul. **24**, 107–120 (1986)