Multi-way Contingency Tables with Uniform Margins

# Multi-way contingency tables with uniform margins

Elisa Perrone[1], Roberto Fontana[2], and Fabio Rapallo[3]

[1] Department of Mathematics and Computer Science, Eindhoven University of Technology, Groene Loper 3, 5612 AE Eindhoven, The Netherlands,
e.perrone@tue.nl

[2] Department of Mathematical Sciences, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy,
roberto.fontana@polito.it

[3] Dipartimento DIEC, Università di Genova, via Vivaldi 5, 16126 Genova, Italy,
fabio.rapallo@unige.it

**Abstract.** We study the problem of transforming a multi-way contingency table into an equivalent table with uniform margins and same dependence structure. Such a problem relates to recent developments in copula modeling for discrete random vectors. Here, we focus on three-way binary tables and show that, even in such a simple case, the situation is quite different than for two-way tables. Many more constraints are needed to ensure a unique solution to the problem. Therefore, the uniqueness of the transformed table is subject to arbitrary choices of the practitioner. We illustrate the theory through some examples, and conclude with a discussion on the topic and future research directions.

**Keywords:** Categorical data analysis, odds ratios, multivariate Bernoulli, Iterative Proportional Fitting.

## 1 Preliminaries

Tabular data in the form of contingency tables appear in many applications, such as health care, biology, and social science. Such data format have extensively been investigated in statistics with the primary goal of developing tools to extract information about the relations between the variables (see [7]). In [5], the author draws interesting connections between the analysis of bivariate contingency tables and *copulas*. We recall that the key concept behind copula theory is the separation between the marginal effect and the dependence structure. Such a separation allows for *ad hoc* modeling tools for dependence and is possible by transforming the original joint probability distribution into one with uniform margins on $[0,1]$. This transformation can be obtained through the *Probability Integral Transform* (PIT). The transformed distribution is the copula associated with the original distribution and is uniquely defined if the margins are continuous [8]. In case of discrete random variables, the marginal distributions cannot be transformed into uniform distributions through the PIT.

In fact, the PIT only identifies the copula on a subdomain. As a consequence, there are infinite copula models that would fit such a subdomain. The question if we can adjust the copula theory to benefit from a similar idea of obtaining a margin-free model in the discrete context has been investigated in [5,6]. There and in the references therein, the authors explore the concept of transforming a given two-way contingency table into a new one with uniform margins. Such a transformation makes it easier to interpret the underlying association of the table which might be hidden when margins are heavily unbalanced, as discussed below.

For $2 \times 2$ contingency tables, the transformation entails converting the original table into a new one where all marginal probabilities equal 1/2, i.e., they are uniform on the subdomain. Table 1 reports a classical example from [9] where the data represents smallpox patients at Sheffield Hospital classified based on vaccination status [yes/no] and recovery [yes/no]. The odds ratio, which is a standard tool to measure association in contingency table analysis, is substantial as it equals 19.47. However, due to marginals that are far from uniform, the association may not be readily apparent in the original data. Nevertheless, the association becomes evident when the data is transformed to have marginal probabilities of 1/2 while maintaining the same odds ratio of 19.47. Figure 1 shows the effect of the transformation on the visualization of the data.

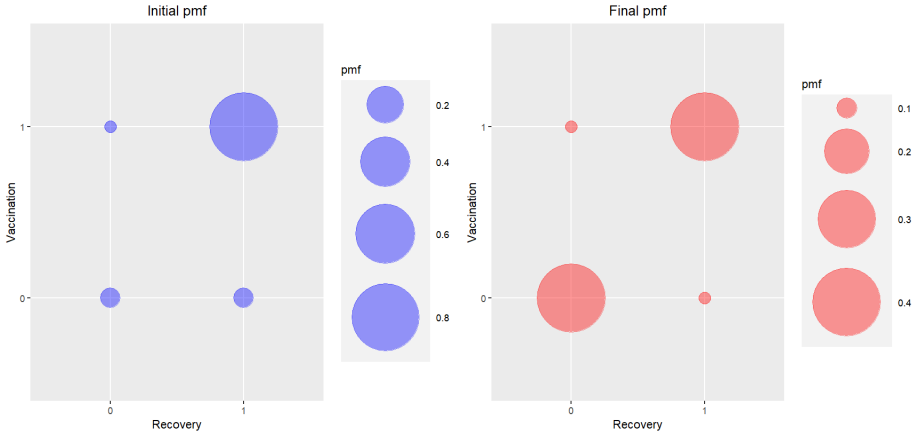
In this work, we investigate how to extend the idea of identifying an equivalent contingency table with uniform margins when we have dimension greater than two. In particular, we focus on the simplest case of $2\times2\times2$ contingency tables. In Sect. 2, we introduce the notation and show our main results. Some interesting examples are illustrated in Sect. 3, and we conclude with open questions for future research on the topic in Sect. 4.

| Vaccination ($X_1$) | Recovery ($X_2$) | $\tilde{n}$ | $\tilde{p}$ |
|---|---|---|---|
| no | no | 274 | 0.06 |
| no | yes | 278 | 0.06 |
| yes | no | 200 | 0.04 |
| yes | yes | 3951 | 0.84 |

Table 1: Sheffield smallpox epidemic reported in [9].

## 2 Multi-way contingency tables and odds ratios

We here outline the mathematical framework introduced by Geenens in [5] relevant for our work. We summarize the findings in two dimensions before moving the discussion to three dimensions.




(a) Original data (b) Transformed data

Fig. 1: Bubble plot of original data (left) and transformed data (right)

**Two-way binary tables.** Given a $2 \times 2$ contingency table $(\tilde{n}_{ij}, i, j = 0,1)$ let $N = \tilde{n}_{00} + \tilde{n}_{11} + \tilde{n}_{01} + \tilde{n}_{10}$ be the grand total and $\omega_\star = \frac{\tilde{n}_{00}\tilde{n}_{11}}{\tilde{n}_{01}\tilde{n}_{10}}$ the odds ratio. For our goal, we consider the table written in terms of the relative frequencies $\tilde{p}_{ij} = \frac{\tilde{n}_{ij}}{N}$ instead of the counts $\tilde{n}_{ij}$. As explained in the motivating example of Sect. 1, the goal is to find a new table $(p_{ij} \geq 0, i, j = 0,1)$ with marginals equal to $1/2$ that maintains the same odds ratio $\omega_\star$ of the original table. Such a table can be obtained by solving the system of equations in Eq. (1)

$$\begin{cases} \dfrac{p_{00}p_{11}}{p_{01}p_{10}} = \omega_\star \\ p_{00} + p_{01} + p_{10} + p_{11} = 1 \\ p_{00} + p_{01} - p_{10} - p_{11} = 0 \\ p_{00} - p_{01} + p_{10} - p_{11} = 0 \end{cases}, \tag{1}$$

whose unique solution only depends on the odds ratio and is as follows

$$\begin{cases} p_{00} = p_{11} = \dfrac{\sqrt{\omega_\star}}{2(1+\sqrt{\omega_\star})} \\ p_{01} = p_{10} = \dfrac{1}{2(1+\sqrt{\omega_\star})} \end{cases}. \tag{2}$$

Using the data in Table 1, we obtain $p_{00} = p_{11} \approx 0.41$ and $p_{01} = p_{10} \approx 0.09$.

**Three-way binary tables.** We now consider three classification variables $X_1, X_2, X_3$, with relative frequencies denoted by $p_{i,j,k}$, for $i, j, k \in \{0,1\}$. A possible generalization of the approach presented for two-way tables to three-way tables

goes through the definition of the *3-dimensional odds ratio* $\omega_3$ given as

$$\omega_3 = \frac{p_{000}p_{011}p_{101}p_{110}}{p_{001}p_{010}p_{100}p_{111}}. \tag{3}$$

Additionally, one can also define the so-called *conditional odds ratios*:

$$\omega_2(X_i, X_j|X_k = a),\ i \neq j \neq k \tag{4}$$

where $\omega_2(X_i, X_j|X_k = a)$ is the 2-dimensional odds ratio computed for $X_i$ and $X_j$ referring to the sub-table defined by $X_k = a$, $a \in \{0,1\}$. Since $k$ is uniquely determined by $i$ and $j$, we streamline the notation by writing $\omega_{ij|a}$ instead of $\omega_2(X_i, X_j|X_k = a)$. Interestingly, the 3-dimensional odds ratio and the conditional odds ratios are linked by the following chain of equalities (see [7]):

$$\omega_3 = \frac{\omega_{12|0}}{\omega_{12|1}} = \frac{\omega_{23|0}}{\omega_{23|1}} = \frac{\omega_{13|0}}{\omega_{13|1}}. \tag{5}$$

We are now ready to propose a possible generalization of the problem presented for two-way binary tables. In particular, we consider a $2\times2\times2$ contingency table $(\tilde{n}_{ijk}, i, j, k = 0, 1)$, with grand total $N$, relative frequencies $\tilde{p}_{ijk} = \tilde{n}_{ijk}/N, i, j, k \in \{0,1\}$, 3-dimensional odds ratio $\omega_\star$ (defined as in Eq. (3)), and conditional odds ratios $\omega^\star_{ij|a}$ (defined as in Eq. (4)). The goal is to find the relative frequencies of a new table $(p_{ijk} \geq 0, i, j, k \in \{0,1\})$ with uniform marginals equal to 1/2 that maintains the same odds ratio $\omega_\star$ as the original table. Basically, we transform the problem of Eq. (1) into the system of equations reported below:

$$\begin{cases} \dfrac{p_{000}p_{011}p_{101}p_{110}}{p_{001}p_{010}p_{100}p_{111}} = \omega_\star \\ p_{000} + p_{001} + p_{010} + p_{011} + p_{100} + p_{101} + p_{110} + p_{111} = 1 \\ p_{000} + p_{001} + p_{010} + p_{011} - p_{100} - p_{101} - p_{110} - p_{111} = 0 \\ p_{000} + p_{001} - p_{010} - p_{011} + p_{100} + p_{101} - p_{110} - p_{111} = 0 \\ p_{000} - p_{001} + p_{010} - p_{011} + p_{100} - p_{101} + p_{110} - p_{111} = 0 \end{cases}. \tag{6}$$

Differently from the case of two-way tables, the solution of the system in Eq. (6) is not unique. One possible solution is given by the generalization of the 2-dimensional solution Eq. (2) as follows

$$\mathbf{p}_\star = \begin{cases} p_{000} = p_{011} = p_{101} = p_{110} = \dfrac{\sqrt[4]{\omega_\star}}{4(1 + \sqrt[4]{\omega_\star})} \\ p_{111} = p_{100} = p_{010} = p_{001} = \dfrac{1}{4(1 + \sqrt[4]{\omega_\star})} \end{cases}. \tag{7}$$

However, there are in general many more solutions suggesting that this straightforward extension is not enough to guarantee a prescribed dependence structure in the new table. A way to ensure a unique solution is to add more constraints to the problem. In particular, a natural choice would be to require that the

transformed table $p_{ijk}$ preserves also the conditional odds ratios of the original table. A new system of equations that includes the additional constraints is given in Eq. (8). We notice that the relationships highlighted in Eq. (5) allows us to only use the conditional odds ratios $\omega^{\star}_{23|0}$,$\omega^{\star}_{13|0}$, and $\omega^{\star}_{12|0}$, since the others can be derived by combining them with the 3-dimensional odds ratio. The problem presented in Eq. (8) has a unique solution that can be computed using symbolic mathematical software like Mathematica

$$\begin{cases} \dfrac{p_{000}p_{011}p_{101}p_{110}}{p_{001}p_{010}p_{100}p_{111}} = \omega_{\star} \\ \dfrac{p_{000}p_{011}}{p_{001}p_{010}} = \omega^{\star}_{23|0} \\ \dfrac{p_{000}p_{101}}{p_{001}p_{100}} = \omega^{\star}_{13|0} \\ \dfrac{p_{000}p_{110}}{p_{010}p_{100}} = \omega^{\star}_{12|0} \\ p_{000} + p_{001} + p_{010} + p_{011} + p_{100} + p_{101} + p_{110} + p_{111} = 1 \\ p_{000} + p_{001} + p_{010} + p_{011} - p_{100} - p_{101} - p_{110} - p_{111} = 0 \\ p_{000} + p_{001} - p_{010} - p_{011} + p_{100} + p_{101} - p_{110} - p_{111} = 0 \\ p_{000} - p_{001} + p_{010} - p_{011} + p_{100} - p_{101} + p_{110} - p_{111} = 0 \end{cases} . \tag{8}$$

## 3 Examples

In this section, we present two examples to illustrate the results reported in Sect. 2. First, we analyze the artificial data of Table 2, taken from [1]. The cross-classification involves three binary variables, which are the response to a treatment (success or failure, $X_1$), the treatment (drug A or drug B, $X_2$), and the clinic (1 or 2, $X_3$). The frequencies in Table 2 are expected counts from a model with conditional odds ratio $\omega^{\star}_{12|0}$ (treatment and response given the clinic) equal to 1. The other conditional odds-ratios are given by $\omega^{\star}_{23|0} = 6$, $\omega^{\star}_{13|0} = 0.167$, and the 3-dimensional odds-ratio $\omega_{\star} = 1$. Through Mathematica, it was possible to find the solution $p$ to the problem of Eq. (8). The solution is reported in Table 2. Again using Mathematica, we were also able to find the closed form solution to the more general problem of Eq. (6) with the only constraint (besides uniform margins) that the 3-dimensional odds ratio equals 1. The solution has three free variables and is reported in the following Eq. (9):

$$\begin{cases} 0 < p_{000} < 1/2 \\ 0 < p_{001} < 1/2(1 - 2p_{000}) \\ 0 < p_{010} < 1/2(1 - 2p_{000} - 2p_{001}) \\ p_{011} = 1/2(1 - 2p_{000} - 2p_{001} - 2p_{010}) \\ p_{100} = p_{011} \\ p_{101} = 1/2(2p_{010} + 2p_{011} - 2p_{100}) \\ p_{110} = 1/2(2p_{001} + 2p_{011} - 2p_{100}) \\ p_{111} = p_{000} + p_{001} + p_{010} + p_{011} - p_{100} - p_{101} - p_{110} \end{cases} . \tag{9}$$

We now consider another example that is based on real data reported in Table 3. Such dataset has been analyzed in [3] to illustrate interactions in three-way tables. In this example, thirty patients suffering from lymphocytic lymphoma who responded to a course of combination chemotherapy are classified by sex ($X_2$) and cell type ($X_3$). With a log-linear analysis, the author concludes that the cell type is related to both sex and outcome of the therapy ($X_1$). By using Mathematica, we could find the solution of the constrained problem with same odds ratio structure of the original data but uniform margins. The new table entries are showed in Table 3.

In these examples the difference between the original table and the transformed one is less evident because the three marginal distributions are more balanced than those in Table 1. We can conclude that the more unbalanced the margins, the more advantageous the transformation, as it provides more insights on the dependence structure. We also notice that the solution derived with our method coincides with the one obtained by applying the standard Iterative Proportional Fitting Procedure (IPFP) [2,7]. The connection between the two approaches will be subject of further studies.

| | | Response ($X_1$) | |
|---|---|---|---|
| Clinic ($X_3$) | Treatment ($X_2$) | Success | Failure |
| 1 | A | 18 | 12 |
| | | (0.252) | (0.103) |
| | B | 12 | 8 |
| | | (0.103) | (0.042) |
| 2 | A | 2 | 8 |
| | | (0.042) | (0.103) |
| | B | 8 | 32 |
| | | (0.103) | (0.252) |

Table 2: Response vs Treatment and Clinic, from [1]. Counts and solution to problem Eq. (8) (in brackets).

## 4 Conclusions and discussion

In this work, we present possible extensions of [5] to multi-way contingency tables. The primary goal of our investigation is to derive a unique solution to the problem of searching for a new table with preserved dependence structure and uniform margins. As discussed in Sect. 2 and Sect. 3, there are infinite tables with uniform margins and a fixed 3-dimensional odds ratio. However, we show that the uniqueness is guaranteed when also imposing constraints on the conditional odds-ratios. We notice that this is not the only possible choice to ensure uniqueness. For example, we could add constraints on the uniformity

| | | Outcome ($X_1$) | |
|---|---|---|---|
| Cell type ($X_3$) | Sex ($X_2$) | No Response | Response |
| Nodular | Male | 1 | 4 |
| | | (0.024) | (0.133) |
| | Female | 2 | 6 |
| | | (0.065) | (0.278) |
| Diffuse | Male | 12 | 1 |
| | | (0.305) | ( 0.040) |
| | Female | 3 | 1 |
| | | (0.105) | (0.050) |

Table 3: Response of Lymphoma patients to Combination Chemotherapy: Distribution by Sex and Cell type, from [3]. Counts and solution to problem Eq. (8) (in brackets).

of the 2-dimensional sections of the multi-way table. Such a choice results in searching for the probabilities $p_{ijk} \geq 0,\ i,j,k = 0,1$ that satisfy the system of equations reported in the following Eq (10):

$$\begin{cases} \dfrac{p_{000}p_{011}p_{101}p_{110}}{p_{001}p_{010}p_{100}p_{111}} = \omega_\star \\ p_{000}+p_{001}+p_{010}+p_{011}+p_{100}+p_{101}+p_{110}+p_{111} = 1 \\ p_{000}+p_{001}+p_{010}+p_{011}-p_{100}-p_{101}-p_{110}-p_{111} = 0 \\ p_{000}+p_{001}-p_{010}-p_{011}+p_{100}+p_{101}-p_{110}-p_{111} = 0 \\ p_{000}-p_{001}+p_{010}-p_{011}+p_{100}-p_{101}+p_{110}-p_{111} = 0 \\ p_{000}+p_{001}-p_{010}-p_{011} = 0 \\ p_{010}+p_{011}-p_{100}-p_{101} = 0 \\ p_{100}+p_{101}-p_{110}-p_{111} = 0 \\ p_{000}-p_{001}+p_{010}-p_{011} = 0 \\ p_{001}+p_{011}-p_{100}-p_{110} = 0 \\ p_{100}-p_{101}+p_{110}-p_{111} = 0 \\ p_{000}-p_{001}+p_{100}-p_{101} = 0 \\ p_{001}-p_{010}+p_{101}-p_{110} = 0 \\ p_{010}-p_{011}+p_{110}-p_{111} = 0 \end{cases} . \tag{10}$$

Interestingly, the unique solution to this problem is the generalization of Eq (2) given by the following entries of the three-way table

$$\begin{cases} p_{000} = p_{011} = p_{101} = p_{110} = \dfrac{\sqrt[4]{\omega_\star}}{4(1+\sqrt[4]{\omega_\star})} \\ p_{111} = p_{100} = p_{010} = p_{001} = \dfrac{1}{4(1+\sqrt[4]{\omega_\star})} \end{cases} .$$

Although the solution is unique, this approach results in a constrained dependence of $(X_1, X_2)$, $(X_1, X_3)$, and $(X_2, X_3)$ which is then modified from the original table structure. Therefore, it is not in line with the idea of maintaining the dependence structure of the original table. Unique solutions could also be obtained under other types of stochastic constraints, such as imposing exchangeability. We will investigate this aspect in the future. Additionally, we plan to extend the theory to arbitrary three-way tables. We believe that this extension is always possible if there are no zeros in the table and the (conditional) odds ratios can be computed. If zero entries are present, there might be cases where there is no table with uniform margins that also maintains the odds-ratio structure. Extending Theorem 6.1 of [5] to multi-way tables is needed to ensure the existence of such a table. In future work, we plan to analyze this aspect and build on previous results on the impact of structural zeros for two-way tables presented in [4]. Finally, extending the approach introduced here to arbitrary multi-way contingency tables in comparison with existing methods, i.e., IPFP, is also an interesting direction for further research. Such a case is non trivial: as the dimension grows, the number of conditional odds ratios and the degree of the involved equations grow exponentially. A possible way to ensure the feasibility of the problem is to pair odds-ratio constraints with stochastic constraints that naturally reduce the space of solutions. We will explore this idea in a follow-up paper.

**Acknowledgements** We thank the anonymous reviewer for their comments on a previous version of the manuscript. Additionally, Roberto Fontana and Fabio Rapallo are members of the GNAMPA-INdAM group.

## References

1. A. Agresti. *Categorical data analysis. Third edition.* John Wiley and Sons, 2012.
2. J. Barthélemy and T. Suesse. mipfp: An R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. *Journal of Statistical Software*, 86(2): 1–20, 2018.
3. Y. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis.* Springer, 2007.
4. R. Fontana, E. Perrone, and F. Rapallo. Contingency tables with structural zeros and discrete copulas. In *Book of the short papers - SIS 2023, June 21-23, 2023, Ancona (Italy)*, pages 713 – 718. Pearson, 2023.
5. G. Geenens. Copula modeling for discrete random vectors. *Dependence Modeling*, 8(1):417–440, 2020.
6. I. Kojadinovic and T. Martini. *Copula-like inference for discrete bivariate distributions with rectangular supports.* arXiv:2307.04225, 2024.
7. T. Rudas. *Lectures on categorical data analysis.* Springer, 2018.
8. A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de Paris*, 8:229–231, 1959.
9. G Udny Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652, 1912.