Statistical Methodology 27 (2015) 132–145

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Directionally collapsible parameterizations of multivariate binary distributions

Tamás Rudas

Department of Statistics, Faculty of Social Sciences, Eötvös Loránd University, Budapest, Hungary

ARTICLE INFO

Article history:
Received 20 July 2014
Received in revised form
4 July 2015
Accepted 4 July 2015
Available online 17 July 2015

Keywords:
Directional collapsibility
Odds ratio
Simpson's paradox
Parameterization of binary distributions
Variation independence from marginal distributions

ABSTRACT

Odds ratios and log-linear parameters are not collapsible, which means that including a variable into the analysis or omitting one from it, may change the strength of association among the remaining variables. Even the direction of association may be reversed, a fact that is often discussed under the name of Simpson's paradox. A parameter of association is directionally collapsible, if this reversal cannot occur. The paper investigates the existence of parameters of association which are directionally collapsible. It is shown, that subject to two simple assumptions, no parameter of association, which depends only on the conditional distributions, like the odds ratio does, can be directionally collapsible. The main result is that every directionally collapsible parameter of association gives the same direction of association as a linear contrast of the cell probabilities does. The implication for dealing with Simpson's paradox is that there is exactly one way to associate direction with the association in any table, so that the paradox never occurs.

1. Introduction

This paper studies the relationships between certain properties which parameters of associations for binary distribution may have. Goodman and Kruskal [10] gave an overview of bivariate parameters of association and they argued that no single concept of association may be used in all research problems. Interest since then has turned towards the multivariate case and, although there have been alternative suggestions, see, e.g., [1,13], applications and theoretical work in the last fifty years

E-mail address: rudas@tarki.hu.

http://dx.doi.org/10.1016/j.stamet.2015.07.001
1572-3127/

have concentrated around the odds ratio and quantities derived from it, mostly because of their relevance in log-linear and other graphical Markov models, see, e.g, [6,14]. The multivariate version of the odds ratio was first considered in [3], see also [5], and [12] for a review of related approaches. However, not every analyst is entirely satisfied with odds ratios (or their logarithms), as parameters of association. First, the standard error of the sample odds ratio, as an estimator, depends not only on the true value of the odds ratio, but is a monotone function of the sum of the reciprocals of the cell probabilities, resulting in high variability of estimates. Second, lack of collapsibility is often cited as an undesirable property, see, e.g., [23,22,21]. The fact that even the direction of association may change after collapsing (e.g., taking the new drug may be associated with recovery for both male and female patients, but disregarding sex, taking the old drug is associated with recovery) is seen as paradoxical by many, as shown by the widespread literature on 'Simpson's paradox'. In addition to well-known occurrences of Simpson's paradox in sociology, education and the health sciences, it is also being discussed in genetics [7] and in physics [15].

As opposed to the vast majority of this literature, Simpson's paradox is not considered here as a special, perhaps negative, feature of the data for which it occurs, rather it is considered as a characteristic of the parameter of association applied, namely the odds ratio, that conditional and marginal associations may have opposing directions (cf. [22,18]). Directional collapsibility means, that such a reversal cannot occur.

The direction of association is readily interpreted for $k = 2$. If one variable is treatment, the other is response to it, then the direction of association tells whether the treatment is beneficial or detrimental to the response. If the two variables are treated on an equal footing, that is, none of them is assumed to be a response to the other, then the direction of association tells whether concordant or discordant types of observations are more likely. For more than 2 variables, when one is response to the others, if each treatment is beneficial when applied individually, the direction of association may tell whether applying all treatments has additional benefit or whether it is not better or even worse than applying the treatments individually. When the variables are treated on an equal footing, one possible interpretation is given in (6) and in the discussion following it. However, just like there are several parameters of association, there are also several meanings of association.

The paper investigates the possibility of finding directionally collapsible parameters of association, which also provide a parameterization of multivariate binary distributions. The main results are obtained under two simple assumptions made for parameters of association, which are described and motivated in Section 2. These two properties are possessed not only by the odds ratio, but also by a simple contrast of the cell probabilities defined in (5). It is also shown in Section 2, that both the odds ratios and the contrasts, associated with all marginal distributions, constitute a parameterization of the joint distribution.

The main results of the paper are given in Section 3. Variation independence of the odds ratio from lower dimensional marginal distributions, formulated here as dependence on the conditional distributions only, which is a very desirable property in other contexts (see, e.g., [17]), turns out to imply the lack of directional collapsibility. More precisely, any parameter of association which depends only on the conditional distributions, assigns the same direction of association to every distribution as the log odds ratio does, and, therefore, is not directionally collapsible. On the other hand, a parameter of association is directionally collapsible, if and only if it assigns the same direction of association to every distribution, as the contrast of the cell probabilities does.

One is then left with the following simple situation. If the two properties described in Section 2 are assumed, then all parameters of association which depend on the conditional distributions only, deem the direction of association as the odds ratio does, and are not directionally collapsible. Further, all directionally collapsible parameters of association assign the same direction to association as the contrast does, and the latter also provides a parameterization of the distribution.

Section 3 gives an illustrative example of using the contrast to determine the direction of association in a way such that Simpson's paradox is avoided, even if the odds ratio exhibits the paradox. Section 4 concludes the paper with a brief discussion of the potential use and limitations of the contrast as a parameter of association and the implications for dealing with Simpson's paradox. Those analysts who are interested in the direction of association only, and find the contrast being overly simple, failing to properly describe their concept of association, cannot avoid Simpson's