Computational Statistics and Data Analysis 68 (2013) 52–65

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A reformulation of the aggregate association index using the odds ratio

Eric J. Beh *, Duy Tran, Irene L. Hudson

School of Mathematical and Physical Sciences, University of Newcastle, Australia

ARTICLE INFO

Article history:
Received 29 August 2012
Received in revised form 6 June 2013
Accepted 6 June 2013
Available online 14 June 2013

Keywords:
2×2 contingency table
Aggregate association index (AAI)
Chi-squared statistic
Fisher's criminal twin data
New Zealand voting data (1883–1919)

ABSTRACT

Since its inception in the 1950s the odds ratio has become one of the most simple and popular measures available for analysing the association between two dichotomous variables. Since the direction and magnitude of the association can be captured in such a simple measure, its impact has been felt throughout much of scientific research, in particular in epidemiology and clinical trials. Despite this, its applicability for analysing aggregate data has rarely been considered. In this paper we shall express a new measure of association (the aggregate association index, or AAI), in terms of the classic odds ratio. The advantage of doing so is that we are able to explore the use of the odds ratio in a context for which it was not originally intended, and that is for the analysis of a 2×2 table where only the aggregate data is known.

1. Introduction

The odds ratio remains one of the most simple, influential and diversely used measures of association available to the analyst. Due to the widespread use of logistic regression, the odds ratio is widely used in many fields of medical and social science research. It is commonly used in survey research, in epidemiology (Rothman, 2002; Rothman et al., 2008) and to express the results of some clinical trials, such as in case-control studies (Miettinen, 1976). The odds ratio underpins the field of meta-analysis (Cheung et al., 2012). Meta analysis is a statistical method used to compare and combine effect sizes from a pool of relevant empirical studies. It is now a standard approach to synthesise research findings in many disciplines, including medical and healthcare research, and climate change research (Hudson, 2010) and increasingly in genome-wide studies (Nakaoka and Inoue, 2009; Kraft et al., 2009; Schurink et al., 2012) and drug discovery (Hudson et al., 2012). The odds ratio is often used as an alternative to the relative risk measure (Zhang and Yu, 1998; Montreuil et al., 2005; Schmidt and Kohlmann, 2008; Viera, 2008) in many applications where it is important to measure the strength, and direction, of the association between two dichotomous variables from a 2×2 table. Despite its popularity, using the odds ratio in cases where only the margins of the 2×2 table are available has rarely been considered. One exception to this is Plackett (1977) who showed that the margins do not provide enough information to make inferences about the cell values. There are a host of techniques that lie within the ecological inference literature that one may consider for inferring cell values, or some function of them; none of them, however, consider the odds ratio. For example, King's (1997) groundbreaking parametric and non-parametric approaches may be considered. King (1997) also describes the ecological inference problem at length. Other strategies include Goodman's (1953) ecological regression, Freedman et al.'s (1991) neighbourhood model, Chamber and Steel's (2001) semi-parametric approach, Steel et al.'s (2004) homogeneous model and Wakefield's (2004) Bayesian extension of this model. A comprehensive review of these ecological inference techniques, and their application to early New Zealand gender and voter turnout data was given by Hudson et al. (2010). Wakefield et al. (2011) further discuss

* Corresponding author. Tel.: +61 2 4921 5113.
E-mail address: eric.beh@newcastle.edu.au (E.J. Beh).

0167-9473/$ – see front matter Crown Copyright © 2013 Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.06.009

Table 1
Notation for a 2×2 contingency table.

	Column 1	Column 2	Total
Row 1	n_{11}	n_{12}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

strategies for determining individual level information based only on aggregate data and Imai et al. (2011) provide an R package, `eco`, for performing ecological inference.

While these techniques all have their merits, they are all applicable only to the case where multiple, or stratified, 2×2 tables are simultaneously considered; they cannot used for analysing the aggregate data of a single 2×2 table. They also involve the estimation of simple transformations of the cell frequencies, rather than the general association structure of the data. Therefore, all of these techniques are also subject to a variety of untestable assumptions which are problematic for evaluating their effectiveness. It is therefore appropriate to consider a strategy that does not rest on assumptions that are not testable while being applicable to a single 2×2 table. Thus, such a strategy should not estimate the $(1, 1)$'th cell frequency, or some transformation of it, but rather examine the structure of the association between two dichotomous variables based only on the marginal information. This issue has a long history and dates back as far as Fisher (1935, p. 48) who focused on the case where one may "blot out the contents of the table". We shall consider this same issue but examine it by incorporating the classic odds ratio into a new index of association called the aggregate association index, also simply referred to as the AAI, proposed by Beh (2008, 2010). It shall be shown that considering the odds ratio offers a simple alternative to considering the AAI and whose calculation is as easy as that considered in Beh (2008). This will be achieved in the following 5 sections. Section 2 provides a review of Beh's (2008, 2010) AAI and Section 3 demonstrates how the odds ratio can be incorporated into this measure. The direction of the association structure, when only the marginal information is available is explored in Section 4 while Section 5 considers the application of the odds ratio—AAI link using two data sets. The first is the classic twin data of Fisher (1935) and was considered by Beh (2008, 2010) in his development of the AAI. The second 2×2 table is based on the study conducted by Hudson et al. (2010) and considers data obtained from the 1893 national election held in New Zealand. This data describes the gendered voting rates and is important because it was in this election that New Zealand became the first self-governing country in the world where women could vote at the national level. Some final comments are made in Section 6.

It must be pointed out that it is not the aim of this paper to formulate strategies for making inferences about the magnitude of the odds ratio given only the marginal information. Instead we shall use the properties of the odds ratio and the AAI, given only the marginal information of a 2×2 contingency table, to explore the association structure of the variables.

2. Aggregate association index

Consider a single two-way contingency table where both variables are dichotomous. Suppose that n individuals/units are classified into this table such that the number classified into the (i, j)th cell is denoted by n_{ij} and the proportion of those in this cell $p_{ij} = n_{ij}/n$ for $i = 1, 2$ and $j = 1, 2$. Denote the proportion of the sample classified into the i'th row and j'th column $p_{i\bullet} = p_{i1} + p_{i2}$ and $p_{\bullet j} = p_{1j} + p_{2j}$ respectively. Table 1 provides a description of the notation used in this paper.

Typically, measuring the extent to which the row and column variables are associated is achieved by considering the Pearson chi-squared statistic calculated from the counts and margins of a contingency table. For a 2×2 table of the form described by Table 1, this statistic is

$$X^2 = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}.$$

The direction of the association may be determined by considering Pearson's (1900, p. 12) estimate of his tetrachoric correlation. Such an estimate, and one of the most popular measures of correlation for 2×2 contingency tables due to its relative simplicity is

$$r = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}}$$

so that $X^2 = nr^2$.

Another very common measure of association for a 2×2 table is the odds ratio (Cornfield, 1951):

$$\theta = \frac{n_{11}n_{22}}{n_{21}n_{12}}. \tag{1}$$

We shall consider the odds ratio in more detail in Section 3.

Suppose, for now, that the cell values of Table 1 are known. Define $P_1 = n_{11}/n_{1\bullet}$; this is the conditional proportion of an individual/unit being classified into "Column 1" given that they are classified in "Row 1". For the analysis of marginal