

Indices of risk based on relevance quotients, with an application to self-assessed health status

*Indici di Rischio Basati su Quozienti di Rilevanza, con
Applicazione alla Percezione del Proprio Stato di Salute*

Maurizio Brizzi

Dipartimento di Scienze statistiche “Paolo Fortunati”, Università di Bologna,
Via delle Belle Arti 41, 40126 Bologna, e-mail: brizzi@stat.unibo.it

Riassunto: Applicando il quoziente di rilevanza nello studio dell’associazione tra un fattore di rischio E e un evento critico C, si perviene a una scomposizione degli indici RR e OR che evidenzia il contributo di ogni casella di dati. Viene proposta una rappresentazione grafica e vengono definiti due indici di rischio. La distribuzione campionaria esatta degli indici proposti viene studiata e descritta sotto l’ipotesi H_0 di indipendenza tra C ed E, e ne viene valutata la potenza per un’ipotesi alternativa H_1 . Infine, viene proposta un’applicazione relativa a uno studio sulla differente percezione del proprio stato di salute tra i due sessi: i maschi risultano essere molto più ottimisti.

Keywords: Relevance quotient, Rate ratio, Odds ratio, Geometric ratio, Linear ratio, Self-assessed health status.

1. Introduction

Let E be a factor of risk (presence/absence of a certain condition) and C a critical event (occurrence of disease or malformation, death, or any else). If we are studying an eventual association between E and C, the simplest representation of a set of data is the 2×2 association table, indicating the observed frequencies, according to exposure to the risk factor (yes or not), and to the occurrence of the critical event (yes or not), as represented in Table 1.

Table 1: Table of association between exposure (E) and critical event (C)

	C: Yes	No	Row Total
E: Yes	n_{11}	n_{12}	n_{10}
No	n_{21}	n_{22}	n_{20}
Column Total	n_{01}	n_{02}	n

The most commonly used risk indices are rate ratio (RR) and odds ratio (OR):

$$RR = \frac{n_{11}/n_{10}}{n_{21}/n_{20}} = \frac{n_{11} \cdot n_{20}}{n_{21} \cdot n_{10}}; \quad OR = \frac{n_{11} \cdot n_{22}}{n_{21} \cdot n_{12}}. \quad (1)$$

The main properties of these indices are reported in many textbooks (one of the most known is Agresti, 1990). A thorough monographical analysis of the Odds Ratio has been given by Rudas (1997); for recent developments, see Cole et al. (2003). We are

intended to analyse here the link with the *relevance quotient*, a symmetric measure of concordance between events, whose properties have been developed by Costantini (1979). Given two events, A and B, their relevance quotient $Q_{A,B}$ is defined as follows:

$$Q_{A,B} = \frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A) \cdot P(B)} \quad (2)$$

$Q_{A,B}$ is equal to one when A and B are independent, and to zero when they are disjoint; it is greater (less) than one when the events are positively (negatively) correlated. The relevance quotient may be estimated by means of the *empirical relevance quotient*, denoted by $q_{A,B}$, in which probabilities are estimated by relative frequencies.

2. Analysis of cell components

If we consider the table of independence associated to Table 1, whose cells are

$$\tilde{n}_{ik} = \frac{n_{i0} \cdot n_{0k}}{n} \quad (i=1,2; \quad k=1,2),$$

we can determine four empirical relevance quotients, one for each couple of events: $q_{ik} = \frac{n_{ik}}{\tilde{n}_{ik}} \quad (i,k=1,2)$. Main diagonal cells represent a

direct association, the other cells an inverse association. Let us now consider the set $\{c_{ik}; i,k=1,2\}$, with $c_{11}=q_{11}$, $c_{22}=q_{22}$, $c_{12}=1/q_{12}$, $c_{21}=1/q_{21}$. Each of these *cell components* represents the contribution given by a single cell about the association between E and C. It easy to verify that *RR* and *OR* may be written as products of cell components:

$$RR = c_{11} \cdot c_{21}; \quad OR = c_{11} \cdot c_{21} \cdot c_{12} \cdot c_{22} \quad (3)$$

The *OR* is then readable as the overall result of cell information, while the *RR* is the result of the left column cells. If we take the natural logarithm of *RR* and *OR*, we can decompose them into additive components (instead of multiplicative ones).

Let $c_{(h)}$, $h=1,2,3,4$ be the cell components, ranked in decreasing order. Starting with the main component $c_{(1)}$ and multiplying by the next one, we can then define a sequence of

indices of risk (called *cumulative cell ratios*), whose generic term is: $R_h = \prod_{g=1}^h c_{(g)}$. The

h -th index resumes the information coming from h cells; evidently, $R_4 = OR$. If E and C result to be perfectly independent, these indices are all equal to one; if there is a positive association, they are ranked in increasing order. When the probability of being affected is less than 0.5 in both groups, as often happens in epidemiological studies, the main cell components are the left column ones; therefore $R_2 = RR$. Taking the natural logarithm $L_h = \ln(R_h)$ the sequence becomes additive; it is possible to give a graphical representation of the association between E and C, by plotting the set of points $\{t_h, R_h\}$ or $\{t_h, L_h\}$, where t_h are equally spaced values, starting with $t_0 = 0$, $R_0 = 1$, $L_0 = 0$.

We propose here two indices of association, based the sequence of L_h 's. First of all, we consider the average of the indices $\{L_h\}$, which is also the logarithm of the geometric mean of the cumulative cell ratios $\{R_h\}$. We call this index *geometric ratio (GR)*:

$$GR = \frac{1}{4}(L_1 + L_2 + L_3 + L_4) = \ln \left(\sqrt[4]{R_1 \cdot R_2 \cdot R_3 \cdot R_4} \right) \quad (4)$$

Then, we can consider, as a synthesis of the indices $\{L_h\}$, the linear coefficient of regression of the set of points: $\{t_h, L_h\}$, described above, where $t_h = h/4$, which may be seen as the “weight” of the corresponding index (e.g. L_2 represents a half of the cells). Evidently, the resulting coefficient, called here *linear ratio*, is a linear combination of the basic indices $\{L_h\}$; after some easy calculations, we derive:

$$LR = 0,4 (L_3 - L_1 + 2 L_4) \quad (5)$$

Both GR and LR are equal to zero when there is perfect independence between C and E; they become positive (negative) when the factor and the critical event are positively (negatively) correlated. These indices may be used as a test statistic for checking the null hypothesis $H_0: P(C|E) = P(C|\bar{E})$ vs. $H_1: P(C|E) > P(C|\bar{E})$.

We have studied all the possible 2x2 tables and calculated their probabilities under these constraints: $n_{10} = n_{20} = n/2$, $P(C|E) = P(C|\bar{E}) = 0.20$ (null hypothesis holds). Then, we studied the exact sample distribution of the indices GR and LR . In Table 2 we reported some statistical features of the distribution of GR and LR , as well as the corresponding features of the log odds ratio (LOR):

Table 2: Distribution of risk indices: mean, standard deviation, right tail centiles

	n=100			n=200		
	GR	LR	LOR	GR	LR	LOR
Mean value	0.038	0.000	0.000	0.025	0.000	0.000
Std. deviation	0.377	0.519	0.533	0.256	0.357	0.364
95 th centile	0.689	0.849	0.871	0.469	0.593	0.603
98 th centile	0.902	1.107	1.145	0.599	0.731	0.748
99 th centile	1.049	1.234	1.290	0.683	0.837	0.858
99.5 th centile	1.165	1.383	1.443	0.763	0.930	0.956

We can notice that LR values are quite near to LOR ones (just a little lower). The tail centiles may be used as critical points in a statistical test; we have calculated the power for a specific alternative $H_1: P(C|E) = 0.25$, $P(C|\bar{E}) = 0.15$ (see Table 3).

Table 3: Power of GR and LR for a specific alternative H_1

n=100	GR	LR	n=200	GR	LR
$\alpha = 0.05$	0.3264	0.3278	$\alpha = 0.05$	0.5477	0.5482
$\alpha = 0.02$	0.1926	0.1882	$\alpha = 0.02$	0.3737	0.3806
$\alpha = 0.01$	0.1250	0.1197	$\alpha = 0.01$	0.2758	0.2801
$\alpha = 0.005$	0.0765	0.0770	$\alpha = 0.005$	0.1917	0.1958

3. Application to self-assessed health status

One of the most natural dichotomies is sex (Female vs. Male), and many bio-medical analyses have different results depending on this variable. In Table 4 we reported a set

of a health survey data (Cavrini *et al.*, 2004), regarding a sex- and age-representative random sample of 756 people, all living in the northern part of Bologna Province, classified by self-assessed health status. By means of a simple questionnaire, people were classified in two groups: “Feels some health problems” or “Does not feel problems”. The self-evaluation results are markedly different for females and males.

Table 4: *Sex and self-assessed health status (Northern Bologna Province)*

	Feels some health problems		Row total
	Yes	No	
Female	270 (231)	111 (150)	381
Male	189 (228)	186 (147)	375
Column total	459	297	756

Note: Entries are observed data; independence frequencies are in parenthesis.

Cell components: $c_{11} = 1.167$, $c_{12} = 1.348$, $c_{21} = 1.205$, $c_{22} = 1.263$

Cumulative cell ratios: $R_{(1)} = 1.348$; $R_{(2)} = 1.703$; $R_{(3)} = 2.051$; $R_{(4)} = 2.394$

Log cell ratios: $L_{(1)} = 0.299$; $L_{(2)} = 0.532$; $L_{(3)} = 0.718$; $L_{(4)} = 0.873$

Indices of risk: $RR = 1.406$; $OR = 2.394$; $LOR = 0.873$; $GR = 0.606$; $LR = 0.866$

Looking at these results, we can observe that there is a sensible difference between self-assessment of health status in Female and Male groups, showing the first one a worse evaluation. The stronger cell component is c_{12} , corresponding to “Female without problems”, which are 111 (expected 150). The Rate ratio is 1.406 ; therefore, female individuals have a 40.6% higher frequency of feeling some health troubles. The Geometric Ratio is 0.606 (average contribution of a single cell) and the Linear Ratio is 0.866 (average increase for a further cell). All the possible statistical tests derived from risk indices (LOR , GR and LR) show a strongly significant difference between Female and Male group ($p < 0.001$). It would be of great interest to analyse the psychological reason of this difference, but this point goes beyond the purposes of this paper.

References

- Agresti A. (1990) *Categorical data analysis*, Wiley, New York.
- Cavrini G., Collina N., Mezzetti F., Pandolfi P. (2004) *Qualità di vita percepita e stato di salute nell'Azienda Sanitaria Bologna Nord*. (Technical report, in press).
- Cole S.R., Allison P.D., Ananth C.V. (2003) Estimation of cumulative odds ratios, *Annals of Epidemiology*, 14, 172-178.
- Costantini D. (1979) The relevance quotient, *Erkenntnis*, 14, 149-157.
- Rudas T. (1997) *Odds Ratios in the analysis of contingency tables*, Sage, London.