Measures of association. Ordinal variables. Symmetric Tables

MSc Further Statistical Methods, Lecture 3
Hilary Term 2007

Steffen Lauritzen, University of Oxford; January 24, 2007

Measures of association

If (conditional) independence among a pair of variables does not hold, it becomes of interest to *quantify* and *describe* the dependence.

When variables are nominal, there is no direct analogue of covariance or correlation and one must use other measures of association.

We consider the *relative risk* and the *odds-ratio*.

For ordinal variables there are analogues of the correlation coefficient. We shall consider *Kruskal's γ-coefficient*.

Relative risk

Consider 2×2-table with probabilities

	B	
A	1	2
1	p_{11}	p_{12}
2	p_{21}	p_{22}

The *relative risk* ($\rho = RR$) compares
$P(A = 1 \,|\, B = 1) = p_{1\,|\,1} = p_{11}/(p_{11} + p_{21})$ with
$P(A = 1 \,|\, B = 2) = p_{1\,|\,2} = p_{12}/(p_{12} + p_{22})$:

$$\rho = \frac{p_{11}}{p_{12}} \frac{p_{12} + p_{22}}{p_{11} + p_{21}}.$$

Example

The empirical counterpart of the relative risk is

$$\hat{\rho} = \frac{n_{11}}{n_{12}} \frac{n_{12} + n_{22}}{n_{11} + n_{21}}$$

	Sex	
Admitted	Male	Female
Yes	1198	557
No	1493	1278

Here

$$\hat{\rho} = \frac{1198}{557} \frac{557 + 1278}{1198 + 1493} = 1.47$$

so it appears that chances for a male to be admitted is about 47% higher than those for females.

Odds–ratio

The relative risk is an asymmetric measure of association between A and B. This may sometimes be inconvenient, so an alternative is the *odds-ratio* θ.

The (conditional) *odds* for $A = 1$ given $B = 1$ are

$$\omega(A = 1 \,|\, B = 1) = \omega_{11} = \frac{P(A = 1 \,|\, B = 1)}{P(A = 2 \,|\, B = 1)} = \frac{p_{11}}{p_{21}}$$

and similarly for $B = 2$. The odds-ratio is thus

$$\theta = \frac{\omega_{11}}{\omega_{12}} = \frac{(p_{11}/p_{21})}{p_{12}/p_{22}} = \frac{p_{11}p_{22}}{p_{12}p_{21}},$$

which is fully symmetric in A and B and in the labels 1 and 2. Thus it does not change if we relabel the variables or its states.

The odds-ratio is also known as the *cross-product ratio* and its empirical counterpart is

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}},$$

which for the admission example gives

$$\hat{\theta} = \frac{1198 \times 1278}{557 \times 1493} = 1.84.$$

One can easily show that

$$A \perp\!\!\!\perp B \iff \theta = 1$$

and a value of θ greater than one corresponds to positive association (as in the admission example) whereas $\theta < 1$ corresponds to negative association.

Conditional odds-ratios

More generally, if A and B have more than two states, the odds-ratio is defined for two pairs of states (i, i^*) and (j, j^*) as

$$\theta_{ii^*jj^*} = \frac{p_{ij}p_{i^*j^*}}{p_{ij^*}p_{i^*j}}$$

and *$A \perp\!\!\!\perp B$ if and only if all such ratios are equal to one*.

Conditioning on the values of a third variable $C = k$ we similarly have conditional independence $A \perp\!\!\!\perp B \mid C$ if and only if

$$\theta_{ii^*jj^*\,|\,k} = \frac{p_{ijk}p_{i^*j^*k}}{p_{ij^*k}p_{i^*jk}} = 1$$

for all combinations of the indices.

No second-order interaction

If the distribution satisfies the restriction of a log-linear model with no second-order interaction, i.e. if

$$p_{ijk} = a_{ij} b_{jk} c_{ik}$$

then

$$\theta_{ii^*jj^* \mid k} = \frac{a_{ij} b_{jk} c_{ik} a_{i^*j^*} b_{j^*k} c_{i^*k}}{a_{ij^*} b_{j^*k} c_{ik} a_{i^*j} b_{jk} c_{i^*k}} = \frac{a_{ij} a_{ij^*}}{a_{ij^*} a_{i^*j}}$$

so *the conditional odds-ratio is constant in k*.

This does not imply absence of a Simpson paradox! For the marginal distribution of I, J is

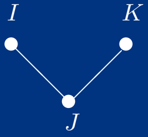
$$p_{ij+} = a_{ij} \sum_k b_{jk} c_{ik} = a_{ij} \tilde{b}_{ij}.$$

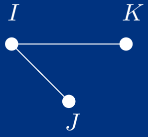
For the IJ odds-ratio to be the same in the marginal table as in the condition it must additionally hold that $\tilde{b}$ satisfies

$$\tilde{b}_{ij} = \alpha_i \beta_j.$$

This holds if either $I \perp\!\!\!\perp K \mid J$ or $J \perp\!\!\!\perp K \mid I$.

Thus, *a Simpson paradox concerning association between I and J is avoided if one of the following graphical models hold,* and typically not otherwise.





Example

	Overall	Department I	II	III	IV	V	VI
odds-ratio	1.84	0.35	0.8	1.13	0.92	1.22	0.83

The empirical odds-ratios for the admission data indicate a strong example of Simpson's paradox.

For department I, Sex and admission is strongly negatively associated. For other departments the association is moderate and of changing sign.

But overall, the association is strong and positive!

Two ordinal variables

Income	Job satisfaction			
	Very diss.	Little diss.	Mod. sat.	Very sat.
$< 15,000$	1	3	10	6
$15,000$–$25,000$	2	3	10	7
$25,000$–$40,000$	1	6	14	12
$> 40,000$	0	1	9	11

For ordinal variables we consider concordant and discordant pairs: A pair $(i_1, j_1), (i_2, j_2)$ is *concordant*

$$i_1 < i_2 \text{ and } j_1 < j_2$$

it is *discordant* if it is the other way around

$$i_1 < i_2 \text{ and } j_1 > j_2,$$

and otherwise it is *tied*.

Kruskal's gamma

Kruskal's γ-coefficient is defined as

$$\gamma = \frac{p_c - p_d}{p_c + p_d},$$

where p_c and p_d are the probability that a random pair of individuals is a concordant or discordant pair.

Clearly, $-1 \leq \gamma \leq 1$ and $\gamma = 0$ for independent variables, so γ is an analogue of the correlation.

As for the correlation, *the variables can be dependent and still have* $\gamma = 0$.

Also $\gamma = 1$ *if and only* $p_{ij} = 0$ *for* $j < i$ and similarly for $\gamma = -1$.

The empirical analogue of Kruskal's γ is

$$\hat{\gamma} = \frac{n_c - n_d}{n_c + n_d} = \frac{1331 - 841}{1331 + 841} = 0.221$$

in the example. So there is a mild (but significant) positive relation between income and job satisfaction.

A test using $|\hat{\gamma}|$ as test statistic can be made using Monte-Carlo p-values (not implemented in MIM).

MIM features a variety of alternative test statistic for exploiting ordinality.

These include the Wilcoxon statistic, the Kruskal–Wallis statistic and the Jonckheere–Terpstra statistic. See Edwards (2002), Chapter 5 for detailed description of these.

Wilcoxon test

			Response		
Centre	Status	Treatment	Poor	Moderate	Excellent
1	1	Active	3	20	5
		Placebo	11	14	8
	2	Active	3	14	12
		Placebo	6	13	5
2	1	Active	12	12	0
		Placebo	11	10	0
	2	active	3	9	4
		Placebo	6	9	3

Multicentre analgesic trial. Here are four variables C: Centre, S: Status, T: Treatment, and R: Response.

Wilcoxon test-statistic compares distribution of *ranks* between two distributions. Ranks are well-defined for ordinal data.

Several categories

Drug regimen	Response		
	None	Partial	Complete
1	2	0	0
2	1	1	0
3	3	0	0
4	2	2	0
5	1	1	4

Two variables D: Drug regimen, R: response. The *Kruskal-Wallis* test statistic measure deviations from independence in direction of at *least one distribution stochastically larger* than the others.

Kruskal-Wallis test specializes to Wilcoxon test for binary variables.

Two ordinal variables

	Job satisfaction			
Income	Very diss.	Little diss.	Mod. sat.	Very sat.
$< 15,000$	1	3	10	6
$15,000$–$25,000$	2	3	10	7
$25,000$–$40,000$	1	6	14	12
$> 40,000$	0	1	9	11

Two ordinal variables: J: Job satisfaction, I: Income. *Jonckheere-Terpstra* test measures deviations from independence in direction of *all distributions being stochastically ordered.*

The Jonckheere–Terpstra test specializes to the Wilcoxon test if one of the two ordinal variables are binary.

Square tables

In some cases, the variables A and B represent 'the same thing' and quite different hypotheses become relevant, for example that of *marginal homogeneity*

$$p_{i+} = p_{+i}.$$

	After		
Before	Approve	Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

Attitude towards UK prime minister. Opinion poll data from Agresti, Ch. 10.

A panel of 1600 persons were asked at two points in time whether they approved of the policy of the current PM. The interesting question is whether the opinion has changed. If it has not, we say there is *marginal homogeneity*

$$p_{i+} = p_{+i}, \text{ for all } i. \tag{1}$$

In 2×2 case this is equivalent to having $\delta = 0$ where

$$\begin{aligned} \delta &= p_{1+} - p_{+1} \\ &= (p_{11} + p_{12}) - (p_{11} + p_{21}) = p_{12} - p_{21} \end{aligned}$$

so

$$p_{1+} = p_{+2} \iff p_{12} = p_{21},$$

i.e. *marginal homogeneity is equivalent to symmetry,* where the hypothesis of symmetry is given as

$$p_{ij} = p_{ji}. \tag{2}$$

The empirical counterpart of δ is

$$\hat{\delta} = \frac{n_{12} - n_{21}}{n}.$$

Under the assumption of homogeneity, the variance of $\hat{\delta}$ can be calculated as

$$\mathbf{V}(n\hat{\delta}) = 2np_{12} = 2np_{21} = 2np.$$

Under the hypothesis

$$\hat{p} = \frac{n_{12} + n_{21}}{2n},$$

so

$$\chi^2 = \frac{n\hat{\delta}^2}{2n\hat{p}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

is for large n approximately χ^2 distributed with 1 degree of freedom.

In the example, we get

$$\chi^2 = \frac{(86 - 150)^2}{86 + 150} = 17.4$$

which is highly significant.

More than two states

The test for symmetry of A and B as expressed in (2) generalizes immediately to several states as

$$\chi^2 = \sum_i \sum_{j>i} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

which is approximately χ^2 distributed with $I(I-1)/2$ degrees of freedom.

Clearly, *marginal symmetry implies marginal homogeneity*.

However, *the converse is false in the multi-state case*.

Testing for marginal homogeneity is more complicated then, see Agresti, Ch. 10.