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 $\gamma$-coefficient.
Relative risk

Consider $2 \times 2$-table with probabilities

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>$p_{11}$</td>
<td>$p_{12}$</td>
</tr>
<tr>
<td>2</td>
<td>$p_{21}$</td>
<td>$p_{22}$</td>
</tr>
</tbody>
</table>

The *relative risk* ($\rho = RR$) compares

$$P(A = 1 \mid B = 1) = p_{1\mid 1} = \frac{p_{11}}{p_{11} + p_{21}}$$

with

$$P(A = 1 \mid B = 2) = p_{1\mid 2} = \frac{p_{12}}{p_{12} + p_{22}}$$

$$\rho = \frac{p_{11} p_{12} + p_{22}}{p_{12} p_{11} + p_{21}}.$$
The empirical counterpart of the relative risk is

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

<table>
<thead>
<tr>
<th>Admitted</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
</tr>
<tr>
<td>Yes</td>
<td>1198</td>
</tr>
<tr>
<td>No</td>
<td>1493</td>
</tr>
</tbody>
</table>

Here

\[ \hat{\rho} = \frac{1198 \times 557 + 1278}{557 \times 1198 + 1493} = 1.47 \]

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* $\theta$.

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

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

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 \( \theta \) 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

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

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

\[ \theta_{ii^*jj^* | k} = \frac{a_{ij}b_{jk}c_{ik}a_{i^*j^*b_{j^*k}c_{i^*k}}}{a_{ij^*b_{j^*k}c_{i^*k}a_{i^*j}b_{jk}c_{i^*k}}} = \frac{a_{ij}a_{ij^*}}{a_{ij^*a_{i^*j}}} \]

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.
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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 15,000</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>15,000–25,000</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>25,000–40,000</td>
<td>1</td>
<td>6</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>&gt; 40,000</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>11</td>
</tr>
</tbody>
</table>

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 $\gamma$-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 $\gamma$ 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 $\gamma$ 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

<table>
<thead>
<tr>
<th>Centre</th>
<th>Status</th>
<th>Treatment</th>
<th>Poor</th>
<th>Moderate</th>
<th>Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Active</td>
<td>3</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>11</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Active</td>
<td>3</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>6</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Active</td>
<td>12</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>11</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>active</td>
<td>3</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>6</td>
<td>9</td>
<td>3</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Drug regimen</th>
<th>Response</th>
<th>None</th>
<th>Partial</th>
<th>Complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

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: $J$: 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.
In some cases, the variables $A$ and $B$ represent ‘the same thing’ and quite different hypotheses become relevant, for example that of *marginal homogeneity*

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

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Approve</td>
<td>Disapprove</td>
<td></td>
</tr>
<tr>
<td>Approve</td>
<td>794</td>
<td>150</td>
<td>944</td>
</tr>
<tr>
<td>Disapprove</td>
<td>86</td>
<td>570</td>
<td>656</td>
</tr>
<tr>
<td>Total</td>
<td>880</td>
<td>720</td>
<td>1600</td>
</tr>
</tbody>
</table>

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*

\[ p_{i+} = p_{+i}, \text{ for all } i. \]  

In a 2 × 2 case this is equivalent to having \( \delta = 0 \) where

\[
\delta = p_{1+} - p_{+1} \\
= (p_{11} + p_{12}) - (p_{11} + p_{21}) = p_{12} - p_{21}
\]

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

\[ p_{ij} = p_{ji}. \] (2)

The empirical counterpart of $\delta$ is

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

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

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

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 \( \chi^2 \) 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 \( \chi^2 \) distributed with \( I(I-1)/2 \) degrees of freedom.

Clearly, \textit{marginal symmetry implies marginal homogeneity}.  

However, \textit{the converse is false in the multi-state case}.  

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