Categorical Data Analysis

Eric J. Beh

Preliminary Quiz, 2021

This course is designed to be a "first" look at various techniques for the analysis of categorical data. It is assumed that you have an undergraduate level understanding of statistical inference but all that is generally required (from a categorical data analysis perspective) is that you are familiar with the goodness-of-fit test and the chi-squared test of independence. The course involves a mix of theoretical, computational and practical problems covering material that looks at the numerical and visual analysis and modelling of categorical data. The three questions in this quiz are meant to reflect some (not all) of the material that is covered in the course.

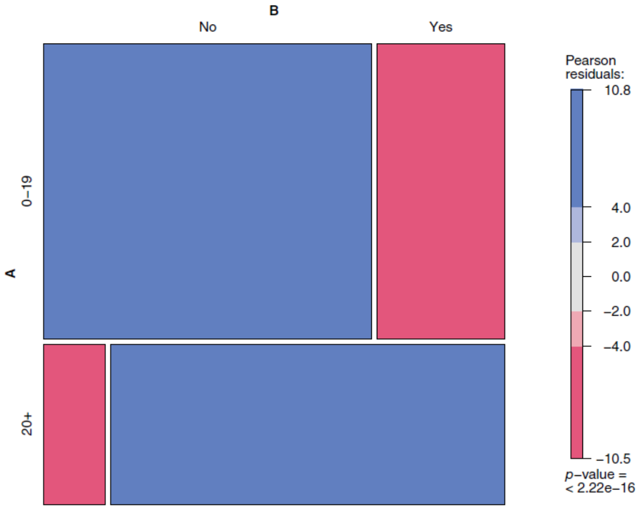
Question 1

Consider the following 2 × 2 contingency table that cross-classifies the years of occupational exposure to asbestos fibres and whether a worker has been diagnosed with asbestosis.

Occupational exposure (years)	Asbestosis		Total
	No	Yes	
0–19	522	203	725
20+	53	339	392
Total	575	542	1117

a) Calculate the odds ratio and provide an interpretation of its value.

b) Below is a mosaic plot that visually summarises the association between the two variables of our contingency table. Use the plot to provide an interpretation of this association.



c) Suppose we have a 2 × 2 contingency table where the notation is defined as follows:

	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

Show that the odds ratio may be expressed only in terms of the (1,1)'th cell frequency n_{11} and the marginal frequencies by

$$\theta = \frac{n_{11}\left(n + n_{11} - (n_{1\bullet} + n_{\bullet 1})\right)}{(n_{1\bullet} - n_{11})(n_{\bullet 1} - n_{11})}.$$

d) Show that n_{11} may be expressed as a function of the odds ratio and the marginal cell frequencies by

$$n_{11} = \frac{[n + (\theta - 1)(n_{1\bullet} + n_{\bullet 1})] \pm \sqrt{[n + (\theta - 1)(n_{1\bullet} + n_{\bullet 1})]^2 - 4\theta(\theta - 1)n_{1\bullet}n_{\bullet 1}}}{2(\theta - 1)}$$

Question 2

Suppose we have an I × J contingency table with sample size n where the (i, j)th relative frequency is denoted by p_{ij} and the i'th row and j'th column marginal relative frequencies are $p_{i\bullet}$ and $p_{\bullet j}$, respectively.

a) Pearson's classic chi-squared statistic can be expressed in the form

$$X^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij} - p_{i\bullet}p_{\bullet j}\right)^2}{p_{i\bullet}p_{\bullet j}}$$

Comment on the role the sample size plays in determining the magnitude of X^2 and what this means when testing the association between categorical variables.

b) Show that Pearson's chi-squared statistic in part a), X^2, can be expressed as

$$X^2 = n\left[\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right) - 1\right]$$

Note that the term $p_{ij}/(p_{i\bullet}p_{\bullet j})$ is sometimes referred to as the (i, j)th Pearson ratio and is just the ratio of the observed and expected cell proportion (under independence).

c) A family of statistics that are chi-squared random variables is defined by the *Cressie-Read divergence statistic* that takes the form

$$\mathrm{CR}(\lambda) = \frac{2n}{\lambda(\lambda+1)} \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \left[\left(\frac{p_{ij}}{p_{i\bullet} p_{\bullet j}} \right)^{\lambda} - 1 \right].$$

Show Pearson's chi-squared statistic, X^2, is a member of this family when $\lambda = 1$.

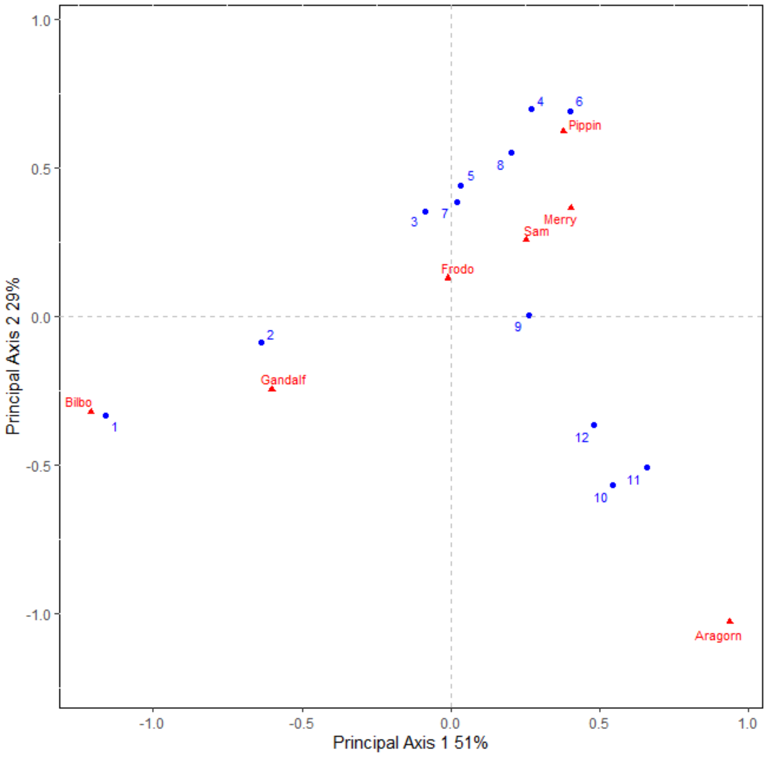
Question 3 is on the next page

Question 3

Correspondence analysis is a statistical technique designed to provide a visual inspection of the association between categorical variables. An analysis of the number of times the primary characters in book 1 (of 6) of the *Lord of the Rings* trilogy (thereby forming the first half of the *Fellowship of the Ring* book), written by JRR Tolkien, was performed to see what characters dominated which chapters. Below is a contingency table that summarises the number of times each character was mentioned by name in each of the 12 chapters.

Name	*Chapter in Book 1*												*Book 1 Total*
	1	2	3	4	5	6	7	8	9	10	11	12	
Frodo	58	88	104	46	52	27	23	28	48	49	39	68	630
Sam	4	27	41	29	11	18	1	11	7	15	25	26	215
Merry	6	4	6	6	34	29	4	13	2	14	21	13	152
Pippin	0	3	35	37	17	13	3	9	8	11	14	10	160
Bilbo	95	50	16	3	11	0	1	2	3	1	4	7	193
Aragorn	0	1	0	0	0	0	0	0	12	57	61	50	181
Gandalf	41	55	29	1	11	1	2	2	1	21	10	6	180
Total	204	228	231	122	136	88	34	65	81	168	174	180	1711

The correspondence analysis produced the following visual summary of the association:



In no more than half a page, describe what this plot says of the nature of the association between the variables of the above table.

Categorical Data Analysis

Eric J. Beh

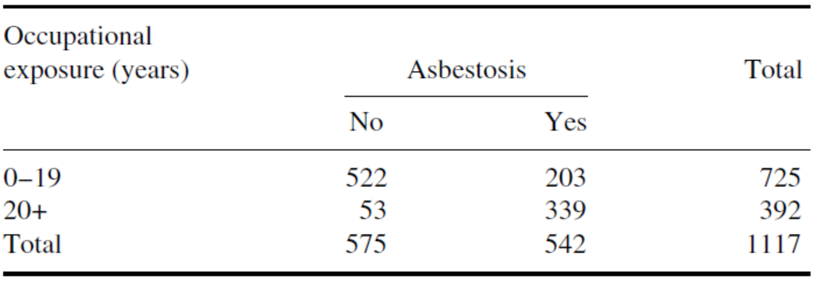
Preliminary Quiz, 2021

Solutions

This course is designed to be a "first" look at various techniques for the analysis of categorical data. It is assumed that you have an undergraduate level understanding of statistical inference but all that is generally required (from a categorical data analysis perspective) is that you are familiar with the goodness-of-fit test and the chi-squared test of independence. The course involves a mix of theoretical, computational and practical problems covering material that looks at the numerical and visual analysis and modelling of categorical data. This quiz is meant to reflect some (not all) of the material that is covered in the course.

Question 1

Consider the following 2 × 2 contingency table that cross-classifies the years of occupational exposure to asbestos fibres and whether a worker has been diagnosed with asbestosis.



Occupational exposure (years)	Asbestosis		Total
	No	Yes	
0–19	522	203	725
20+	53	339	392
Total	575	542	1117

a) Calculate the odds ratio and provide an interpretation of its value.

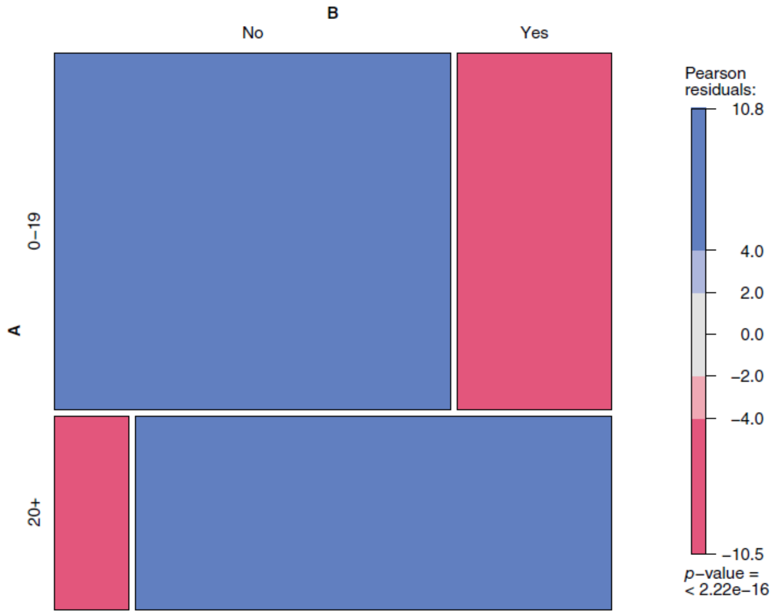
Solution

The odds ratio for this table is

$$\theta = \frac{522 \times 339}{53 \times 203} = 16.45$$

Therefore, a person who has been exposed to asbestos for at least 20 years is 16.45 times more likely to contract asbestosis than those who have been exposed to asbestos for less than 20 years.

b) Below is a mosaic plot that visually summarises the association between the two variables of our contingency table. Use the plot to provide an interpretation of this association.



Solution

The mosaic plot shows that the majority of those who are exposed to asbestos for less than 20 years are unlikely to be diagnosed with asbestosis, while those who have been exposed for at least 20 years are very likely to be diagnosed with the disease. In fact, $522/725 \times 100 = 72\%$ of those who are exposed to asbestos fibres for less than 20 years are not diagnosed with asbestosis while $53/392 \times 100 = 13.5\%$ of those who are exposed to the fibres for at least 20 years will not contract the disease.

c) Suppose we have a 2 × 2 contingency table where the notation is defined as follows:

	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

Show that the odds ratio may be expressed only in terms of the (1,1)'th cell frequency n_{11} and the marginal frequencies by

$$\theta = \frac{n_{11}\big(n + n_{11} - (n_{1\bullet} + n_{\bullet 1})\big)}{(n_{1\bullet} - n_{11})(n_{\bullet 1} - n_{11})}.$$

Solution

$$\begin{aligned}\theta &= \frac{n_{11}n_{22}}{n_{12}n_{21}} \\ &= \frac{n_{11}(n_{2\bullet} - n_{21})}{(n_{1\bullet} - n_{11})(n_{\bullet 1} - n_{11})} \\ &= \frac{n_{11}\big((n - n_{1\bullet}) - (n_{\bullet 1} - n_{11})\big)}{(n_{1\bullet} - n_{11})(n_{\bullet 1} - n_{11})} \\ &= \frac{n_{11}\big(n + n_{11} - (n_{1\bullet} + n_{\bullet 1})\big)}{(n_{1\bullet} - n_{11})(n_{\bullet 1} - n_{11})}\end{aligned}$$

. . . as required.

d) Show that n_{11} may be expressed as a function of the odds ratio and the marginal cell frequencies by

$$n_{11} = \frac{[n + (\theta - 1)(n_{1\bullet} + n_{\bullet 1})] \pm \sqrt{[n + (\theta - 1)(n_{1\bullet} + n_{\bullet 1})]^2 - 4\theta(\theta - 1)n_{1\bullet}n_{\bullet 1}}}{2(\theta - 1)}$$

Solution

From our result in part c),

$$\theta = \frac{n_{11}\big(n + n_{11} - (n_{1\bullet} + n_{\bullet 1})\big)}{(n_{1\bullet} - n_{11})(n_{\bullet 1} - n_{11})}.$$

Then, upon rearranging, we get

$$\begin{aligned}\theta(n_{1\bullet} - n_{11})(n_{\bullet 1} - n_{11}) &= n_{11}\big(n + n_{11} - (n_{1\bullet} + n_{\bullet 1})\big) \\ \theta(n_{1\bullet}n_{\bullet 1} - n_{1\bullet}n_{11} - n_{\bullet 1}n_{11} + n_{11}^2) &= nn_{11} + n_{11}^2 - n_{11}(n_{1\bullet} + n_{\bullet 1})\end{aligned}$$

Bringing everything over to the left-hand side of the equal sign gives

$$n_{11}^2\theta - n_{11}^2 - n_{11}\theta(n_{1\bullet} + n_{\bullet 1}) - n_{11}\big(n - (n_{1\bullet} + n_{\bullet 1})\big) + \theta n_{1\bullet}n_{\bullet 1} = 0$$

or, as a quadratic equation of n_{11},

$$n_{11}^2(\theta - 1) - n_{11}[n + (\theta - 1)(n_{1\bullet} + n_{\bullet 1})] + \theta n_{1\bullet}n_{\bullet 1} = 0$$

Therefore, solving this quadratic equation for n_{11} gives

$$n_{11} = \frac{[n + (\theta - 1)(n_{1\bullet} + n_{\bullet 1})] \pm \sqrt{[n + (\theta - 1)(n_{1\bullet} + n_{\bullet 1})]^2 - 4\theta(\theta - 1)n_{1\bullet}n_{\bullet 1}}}{2(\theta - 1)}$$

. . . as required.

Question 2

Suppose we have an I × J contingency table with sample size n where the (i, j)th relative frequency is denoted by p_{ij} and the i'th row and j'th column marginal relative frequencies are $p_{i\bullet}$ and $p_{\bullet j}$, respectively.

a) Pearson's classic chi-squared statistic can be expressed in the form

$$X^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij} - p_{i\bullet}p_{\bullet j}\right)^2}{p_{i\bullet}p_{\bullet j}}$$

Comment on the role the sample size plays in determining the magnitude of X^2 and what this means when testing the association between categorical variables.

Solution

This result shows that doubling the sample size (say) doubles the magnitude of the chi-squared statistic. Multiplying the sample size by a factor of 10 (again, say) will multiply the chi-squared statistic by 10 even if there is no change in the underlying association between the variables. This means that there will always be a sample size (even if the differences $p_{ij} - p_{i\bullet}p_{\bullet j}$ remain unchanged) that will result in a statistically significant association between the categorical variables.

b) Show that Pearson's chi-squared statistic in part a), X^2, can be expressed as

$$X^2 = n\left[\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right) - 1\right]$$

Note that the term $p_{ij}/\left(p_{i\bullet}p_{\bullet j}\right)$ is sometimes referred to as the (i, j)th Pearson ratio and is just the ratio of the observed and expected cell proportion (under independence).

Solution

$$\begin{aligned}
X^2 &= n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij} - p_{i\bullet}p_{\bullet j}\right)^2}{p_{i\bullet}p_{\bullet j}} \\
&= n\sum_{i=1}^{I}\sum_{j=1}^{J}\left[\frac{p_{ij}^2 - 2p_{ij}p_{i\bullet}p_{\bullet j} + p_{i\bullet}^2p_{\bullet j}^2}{p_{i\bullet}p_{\bullet j}}\right] \\
&= n\sum_{i=1}^{I}\sum_{j=1}^{J}\left[\frac{p_{ij}^2}{p_{i\bullet}p_{\bullet j}} - 2p_{ij} + p_{i\bullet}p_{\bullet j}\right]
\end{aligned}$$

$$= n\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^2}{p_{i\bullet}p_{\bullet j}} - 2 + 1\right]$$

$$= n\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^2}{p_{i\bullet}p_{\bullet j}} - 1\right]$$

$$= n\left[\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right) - 1\right]$$

. . . as required.

c) A family of statistics that are chi-squared random variables is defined by the *Cressie-Read divergence statistic* that takes the form

$$\mathrm{CR}(\lambda) = \frac{2n}{\lambda(\lambda+1)}\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\left[\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)^{\lambda} - 1\right].$$

Show Pearson's chi-squared statistic, X^2, is a member of this family when $\lambda = 1$.

Solution

$$\mathrm{CR}(1) = \frac{2n}{1(1+1)}\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\left[\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)^{1} - 1\right]$$

$$= n\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\left[\frac{p_{ij} - p_{i\bullet}p_{\bullet j}}{p_{i\bullet}p_{\bullet j}}\right]$$

$$= n\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^2}{p_{i\bullet}p_{\bullet j}} - 1\right]$$

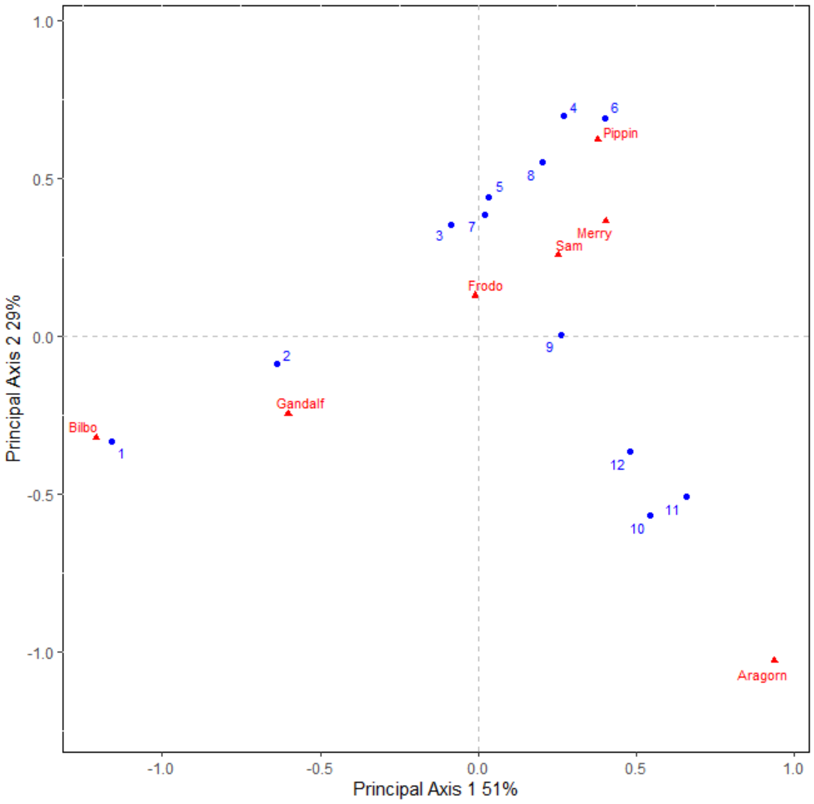
. . . and is just Pearson's chi-squared statistic in part a).

Question 3

Correspondence analysis is a statistical technique designed to provide a visual inspection of the association between categorical variables. An analysis of the number of times the primary characters in book 1 (of 6) of the *Lord of the Rings* trilogy (thereby forming the first half of the *Fellowship of the Ring* book), written by JRR Tolkien, was performed to see what characters dominated which chapters. Below is a contingency table that summarises the number of times each character was mentioned by name in each of the 12 chapters.

Name	*Chapter in Book 1*												*Book 1*
	1	2	3	4	5	6	7	8	9	10	11	12	*Total*
Frodo	58	88	104	46	52	27	23	28	48	49	39	68	630
Sam	4	27	41	29	11	18	1	11	7	15	25	26	215
Merry	6	4	6	6	34	29	4	13	2	14	21	13	152
Pippin	0	3	35	37	17	13	3	9	8	11	14	10	160
Bilbo	95	50	16	3	11	0	1	2	3	1	4	7	193
Aragorn	0	1	0	0	0	0	0	0	12	57	61	50	181
Gandalf	41	55	29	1	11	1	2	2	1	21	10	6	180
Total	204	228	231	122	136	88	34	65	81	168	174	180	1711

The correspondence analysis produced the following visual summary of the association:



In no more than half a page, describe what this plot says of the nature of the association between the variables of the above table.

Solution

This plot shows that the character Bilbo dominated chapter 1 while Gandalf dominated chapter 2. Aragorn doesn't dominate any one chapter but is strongly present in chapters 10, 11 and 12. Pippin is dominant in chapters 4 and 6 while Frodo, Sam and Merry are mentioned consistently in many of the middle chapters. This can be seen by observing the proximity of character points from the chapter points.

We note that Frodo is closest to the origin. In correspondence analysis, the origin is the point where all of the coordinates would be if there was complete independence in the contingency table. Therefore, since Frodo is closest to the origin this means that he is a central figure in virtually all of the chapters and so does not impact greatly on the association that exists between the variables (there is an association, and this can be confirmed by performing a chi-squared test of independence which yields a tiny p-value).

For all of you who have read the books (which you should do, they are great . . . much better than the movies . . . which are also great), these findings should come as no surprise!

Another important point is the quality of this two-dimensional display. The percentages along each dimension describe the percentage of the total association that is described (measured using Pearson's chi-squared statistic). Therefore, the 2D plot visually describes 51% + 29% = 80% of the association that is present in the contingency table. This is considered to be a very good visual summary of the association.

More details about the nature of the association can be determined by projecting out to a third dimension or even determining the statistical significance of each category on the association structure using confidence regions.

This is a very simple demonstration of correspondence analysis and it can be extended out for the analysis of three or more categorical variables, and the ordinal structure of the chapters can also be incorporated. An alternative visual display related to this plot, called a *biplot*, can be constructed which gives a better picture of the "distance" between a row point and a column point. More recently, it has been shown that the Cressie-Read divergence statistic can be used to generate a family of correspondence analysis solutions (depending on what measure is used to assess the statistical significance of the association) and such an approach can be used to determine the value of λ that gives the "best" (and "worst") 2D visual representation of the association. Traditionally correspondence analysis is performed where Pearson's chi-squared statistic ($\lambda = 1$) is used as the measure of association but $\lambda = 1$ does not always guarantee that this will give us the "best" view of the association between our categorical variables.