MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 6: Bivariate Correspondence Analysis - part II

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Chi-square Statistics

Independence

The independence between variables x and y can be tested using chi-square statistic. The null hypothesis of the test is

$$H_o : p_{jk} = p_{j.}p_{.k}, \text{ for all } j, k$$

and the test statistic is given by

$$\chi^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(n_{jk} - n_{jk}^*)^2}{n_{jk}^*}.$$

Independence

Under random sampling, the n_{jk} follow multinomial distribution with parameters $n, p_{11}, ..., p_{JK}$ and $E[n_{jk}] = np_{jk}$. In the test statistics above, the np_{jk}, under the null, are estimated by n^*_{jk}. When the sample size n is large, the test statistic has, under the null hypothesis, approximately chi-square distribution with $(K-1)(J-1)$ degrees of freedom. Thus the null hypothesis (independence between variables x and y) is rejected at the level α if

$$\chi^2 > \chi^2_{(K-1)(J-1),1-\alpha}.$$

Links

Chi-square distribution

Multinomial distribution

Chi-square Distances

Chi-square Distance

When the data is in the form of frequency distribution, the distance between the rows (or columns) is measured using weighted euclidian distances. The distance between two rows j_1 and j_2 is given by

$$d^2(j_1, j_2) = \sum_{k=1}^{K} \frac{1}{f_{.k}} (\frac{f_{j_1 k}}{f_{j_1 .}} - \frac{f_{j_2 k}}{f_{j_2 .}})^2.$$

The euclidian distance gives the same weight to each column. The χ^2 distance gives the same relative importance to each column proportionally to the average frequency. The division of each squared term by the expected frequency is variance standardizing and compensates for the larger variance in high frequencies and the smaller variance in low frequencies. If no such standardization were performed, the differences between larger proportions would tend to be large and thus dominate the distance calculation, while the differences between the smaller proportions would tend to be swamped. The weighting factors are used to equalize these differences.

Chi-square Distance

The distance between two columns k_1 and k_2 is given by

$$d^2(k_1, k_2) = \sum_{j=1}^{J} \frac{1}{f_{j.}} (\frac{f_{jk_1}}{f_{.k_1}} - \frac{f_{jk_2}}{f_{.k_2}})^2.$$

Decomposition of the Chi-square Statistic

Decomposition of the Chi-square Statistic

Let $Z \in \mathbb{R}^{J \times K}$, where

$$Z_{jk} = \frac{f_{jk} - f_{j.}f_{.k}}{\sqrt{f_{j.}f_{.k}}}.$$

Clearly

$$\sum_{j=1}^{J}(f_{jk} - f_{j.}f_{.k}) = \sum_{j=1}^{J} f_{jk} - \sum_{j=1}^{J} f_{j.}f_{.k} = f_{.k} - f_{.k}\sum_{j=1}^{J} f_{j.} = f_{.k} - f_{.k} = 0.$$

Similarly,

$$\sum_{k=1}^{K}(f_{jk} - f_{j.}f_{.k}) = 0.$$

Thus, the matrix Z gives scaled and centered relative frequencies of the variables. Moreover, the variables are scaled such that the elements $Z_{jk} = \frac{f_{jk} - f_{j.}f_{.k}}{\sqrt{f_{j.}f_{.k}}} = \frac{f_{jk} - f^*_{jk}}{\sqrt{f^*_{jk}}}$ are the terms that are squared and summed in the chi-square statistic that is used for testing the independence of the variables.

Decomposition of the Chi-square Statistic

A large positive value Z_{jk} indicates a large contribution to the chi-square statistic. This indicates a positive association between row j and column k. (More observations than expected under independence.) A large negative value Z_{jk} also indicates a large contribution to the chi-square statistic, but this indicates a negative association between row j and column k. (Less observations than expected under independence.) Values near zero indicate no contribution to the test statistic. (The number of observations is equal to the expected number under independence.)

Let

$$V = Z^T Z$$

and let

$$W = ZZ^T.$$

Now the chi-square statistic

$$\chi^2 = n(trace(V)) = n(trace(W)).$$

PCA on the Row Profiles

PCA on the Row Profiles

Principal component analysis is based on maximizing euclidian distances. In the context of frequency distributions, the proper distance between the variables is the chi-square distance. Thus, for frequency distributions, PCA has to be applied to modified data.

Whereas traditional PCA relies on euclidian distances, correspondence analysis is based on chi-square distances.

PCA on the Row Profiles

The chi-square distances between two row profiles can be given as

$$d^2(j_1, j_2) = \sum_{k=1}^{K} \frac{1}{f_{.k}} \left(\frac{f_{j_1 k}}{f_{j_1 .}} - \frac{f_{j_2 k}}{f_{j_2 .}}\right)^2$$

$$= \sum_{k=1}^{K} \left(\frac{f_{j_1 k}}{f_{j_1 .}\sqrt{f_{.k}}} - \frac{f_{j_2 k}}{f_{j_2 .}\sqrt{f_{.k}}}\right)^2.$$

Thus, if the row profiles are scaled, the usual euclidian metric can be used on the new scaled data.

PCA on the Row Profiles

Let $R \in \mathbb{R}^{J \times K}$, where

$$R_{jk} = \frac{f_{jk}}{f_{j.}\sqrt{f_{.k}}} - \sqrt{f_{.k}}$$

The matrix R contains the scaled and shifted row profiles. The shifting is such that the weighted sum

$$\sum_{j=1}^{J} f_{j.} \frac{f_{jk}}{f_{j.}\sqrt{f_{.k}}} = \sqrt{f_{.k}}.$$

Let R_j denote the jth row of R. Performing PCA on the row profiles equals to finding orthonormal vectors (directions) u_i such that projection $P_i(\cdot)$ onto u_i maximizes the weighted sum of the euclidian distances,

$$\sum_{j=1}^{J} f_{j.} d^2(0, P_i(R_j)),$$

under the constraint that u_i is orthogonal to all u_l, $1 \leq l < i$.

PCA on the Row Profiles

The problem is again a problem of maximization under constraint, and similarly as in the usual PCA, the solution is given by the eigenvalues and the eigenvectors of the matrix

$$V = \sum_{j=1}^{J} f_{j.} R_j^T R_j$$

Some matrix algebra is needed to show that the matrix

$$V = \sum_{j=1}^{J} f_{j.} R_j^T R_j = Z^T Z.$$

PCA on the Row Profiles

Let λ_i denote the ith largest eigenvalue of the matrix V and let u_i denote the corresponding unit length eigenvector. Let $u_{i,k}$ denote the kth element of u_i. The value (score) of the row profile j (associated with modality A_j) on the ith principal component is given by

$$\phi_{i,j} = \sum_{k=1}^{K} u_{i,k} R_{jk}.$$

It can be proven that ϕ_i is centered such that

$$\sum_{j=1}^{J} f_{j.} \phi_{i,j} = 0,$$

and that the variance of ϕ_i is λ_i.

Contribution of the Modalities

The contribution of the modality A_j on construction of the axis u_i is given by

$$\frac{f_{j.}(\phi_{i,j})^2}{\lambda_i}.$$

Quality of the Representation

The quality of the representation of the centered row profile R_j by the principal axis i is measured by the squared cosine of angle between the vector OR_j and u_i :

$$cos^2(\alpha) = \left(\frac{< OR_j, u_i >}{||OR_j|| \cdot ||u_i||}\right)^2 = \frac{(\phi_{i,j})^2}{||OR_j||^2}.$$

If the value is close to 1, the quality of the representation is good.

Note that the formula above does not contain the weight f_j, and thus one modality can be:

- Close to the axis u_i and and therefore be well represented (well explained).
- Due to a low weight f_j, it can have a low contribution to the axis.

PCA on the Column Profiles

PCA on the Column Profiles

Performing PCA on the column profiles does not differ from performing PCA on the row profiles. The solution is given by the eigenvalues and the eigenvectors of the matrix $W = ZZ^T$.

PCA on the Column Profiles

Let $C \in \mathbb{R}^{J \times K}$, where

$$C_{jk} = \frac{f_{jk}}{f_{.k}\sqrt{f_{j.}}} - \sqrt{f_{j.}}$$

The matrix C contains scaled and shifted column profiles. Let C_k denote the kth column of C. Performing PCA on the column profiles equals to finding orthonormal vectors (directions) v_h such that projection $P_h(\cdot)$ onto v_h maximizes the weighted sum of the euclidian distances,

$$\sum_{k=1}^{K} f_{.k} d^2(0, P_h(C_k)),$$

under the constraint that v_h is orthogonal to all v_l, $1 \leq l < h$. The solution is given by the eigenvalues and the eigenvectors of the matrix $W = ZZ^T$.

PCA on the Column Profiles

Let λ_h denote the hth largest eigenvalue of the matrix W and let v_h denote the corresponding unit length eigenvector. Let $v_{h,k}$ denote the kth element of v_h. The value (score) of the column profile k (associated with modality B_k) on the hth principal component is given by

$$\psi_{h,k} = \sum_{j=1}^{J} v_{h,j} C_{jk}.$$

It can be proven that ψ_h is centered such that

$$\sum_{k=1}^{K} f_{.k} \psi_{h,k} = 0,$$

and that the variance of ψ_h is λ_h.

Contribution of the Modalities

The contribution of the modality B_k on construction of the axis v_h is given by

$$\frac{f_{.k}(\psi_{h,k})^2}{\lambda_h}.$$

Quality of the Representation

The quality of the representation of the centered column profile C_k by the principal axis h is measured by the squared cosine of angle between the vector OC_k and v_h.

$$cos^2(\beta) = \left(\frac{< OC_k, v_h >}{||OC_k|| \cdot ||v_h||}\right)^2 = \frac{(\psi_{h,k})^2}{||OC_k||^2}.$$

If the value is close to 1, the quality of the representation is good.

Association Between the Profiles

Association Between the Profiles

It can be shown that the matrices V and W have the same nonzero eigenvalues. Moreover, the eigenvectors u_i can be given in terms of v_i and vice versa:

$$u_i = \frac{1}{\sqrt{\lambda_i}} Z^T v_i$$

and

$$v_i = \frac{1}{\sqrt{\lambda_i}} Z u_i.$$

Association Between the Profiles

Let $H = rank(V) = rank(W)$. The coolest thing in correspondence analysis is that the attraction-repulsion indices d_{jk} can be given in terms of ϕ and ψ as follows

$$d_{jk} = 1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,j} \psi_{h,k}.$$

Association Between the Profiles

The components are often standardized defining

$$\hat{\psi}_{h,k} = \frac{1}{\sqrt{\lambda_h}} \psi_{h,k}$$

and

$$\hat{\phi}_{h,j} = \frac{1}{\sqrt{\lambda_1}} \phi_{h,j}.$$

Then

$$d_{jk} = 1 + \sqrt{\lambda_1} \sum_{h=1}^{H} \hat{\phi}_{h,j} \hat{\psi}_{h,k}.$$

The attraction-repulsion index d_{jk} is now larger than 1 if and only if the smallest angle between $(\hat{\phi}_{1,j}, ..., \hat{\phi}_{H,j})$ and $(\hat{\psi}_{1,k}, ..., \hat{\psi}_{H,k})$ is less than 90°.

If the row profile j and the column profile k are well represented by the first two principal components, then the attraction-repulsion index

$$d_{jk} \approx 1 + \sqrt{\lambda_1} \sum_{h=1}^{2} \hat{\phi}_{h,j} \hat{\psi}_{h,k}.$$

We can therefore say that the modalities A_j and B_k are attracted to each if the angle between $(\hat{\phi}_{1,j}, \hat{\phi}_{2,j})$ and $(\hat{\psi}_{1,k}, \hat{\psi}_{2,k})$ is less than 90° and they repulse each other if the angle between $(\hat{\phi}_{1,j}, \hat{\phi}_{2,j})$ and $(\hat{\psi}_{1,k}, \hat{\psi}_{2,k})$ is larger than 90°. In this case, one can simply observe the angle from the (double) biplot of the first two components of $\hat{\phi}$ and $\hat{\psi}$.

Example of Correspondence Analysis

Correspondence analysis using the data presented in lecture five. Variable x Education is divided to categories A_1 Primary School, A_2 High School, and A_3 University, and variable y Salary is divided to categories B_1 low, B_2 average, and B_3 high.

	B_1	B_2	B_3	
A_1	150	40	10	200
A_2	190	350	60	600
A_3	10	110	80	200
	350	500	150	1000

Table: Contingency table

Example of Correspondence Analysis

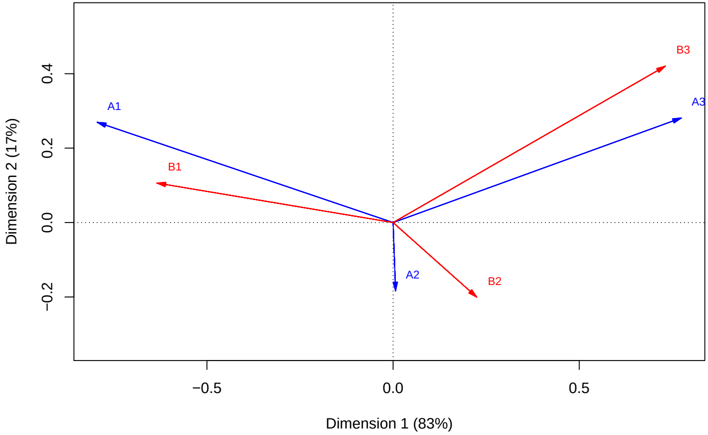

Figure: Salary and education (A1=Primary School education, A2=High School education, A3=University level education, B1=low salary, B2=average salary, B3=high salary)

Next Week

Next week we will talk about multiple correspondence analysis (MCA).

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

References III

L. Simar, An Introduction to Multivariate Data Analysis, Université Catholique de Louvain Press, 2008.