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

Pauliina Ilmonen

Contents

Correspondence Analysis

Correspondence Analysis

Correspondence analysis is a PCA-type method appropriate for analyzing categorical variables. The aim in bivariate correspondence analysis is to describe dependencies (correspondences) in a two-way contingency table.

Example, Education and Salary

In this lecture, we consider an example where we examine dependencies of categorical variables education and salary.

Frequency Tables

Contingency Tables

We consider a sample of size n described by two qualitative variables, x with categories $A_1, \ldots, A_J$ and y with categories $B_1, \ldots, B_K$. The number of individuals having the modality (category) A_j for the variable x and the modality B_k for the variable y is denoted by n_{jk}. Now the number of individuals having the modality A_j for the variable x is given by

$$n_{j.} = \sum_{k=1}^{K} n_{jk},$$

the number of individuals having the modality B_k for the variable y is given by

$$n_{.k} = \sum_{j=1}^{J} n_{jk},$$

and

$$n = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk}.$$

Contingency Tables

The data is often displayed as a two-way contingency table.

	B_1	B_2	$\cdots$	B_K	
A_1	n_{11}	n_{12}	$\cdots$	n_{1K}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2K}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	n_{J1}	n_{J2}	$\cdots$	n_{JK}	$n_{J.}$
	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.K}$	n

Table: Contingency table

Example, Education and Salary

We consider size 1000 sample of two categorical variables. 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.

Example, Education and Salary

We display the Education and Salary data as a two-way contingency table.

	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

- In this sample of 1000 observations, there are 150 individuals that have Primary School education and low salary.
- In this sample of 1000 observations, there are 10 individuals that have Primary School education and high salary.
- In this sample of 1000 observations, there are 110 individuals that have University education and average salary.
- ...

Contingency Tables

The value of the numbers n_{jk} is naturally relative to the total number of observations, n. Thus it is preferable to analyze the contingency table in the form of joint relative frequencies. From the contingency table, it is straightforward to compute the associated relative frequency table (F) where the elements of the contingency table are divided by the number of individuals n leading to $f_{jk} = \frac{n_{jk}}{n}$. The marginal relative frequencies are computed as

$$f_{j.} = \sum_{k=1}^{K} f_{jk}$$

and

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

Contingency Tables

	B_1	B_2	$\cdots$	B_K	
A_1	f_{11}	f_{12}	$\cdots$	f_{1K}	$f_{1.}$
A_2	f_{21}	f_{22}	$\cdots$	f_{2K}	$f_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	f_{J1}	f_{J2}	$\cdots$	f_{JK}	$f_{J.}$
	$f_{.1}$	$f_{.2}$	$\cdots$	$f_{.K}$	1

Table: Table of relative frequencies

Example, Education and Salary

	B_1	B_2	B_3	
A_1	0.15	0.04	0.01	0.20
A_2	0.19	0.35	0.06	0.60
A_3	0.01	0.11	0.08	0.20
	0.35	0.50	0.15	1

Table: Table of relative frequencies

- In this sample 15% of individuals have Primary School education and low salary.
- In this sample, 1% of individuals have Primary School education and high salary.
- In this sample, 11% of individuals have University education and average salary.
- ...

Frequencies

The frequency f_{jk} is the estimate of

$$p_{jk} = P(x \in A_j, y \in B_k),$$

and $f_{j.}$ and $f_{.k}$ are the estimates of

$$p_{j.} = P(x \in A_j),$$

and

$$p_{.k} = P(y \in B_k),$$

respectively.

Row Profiles

Tables of Conditional Frequencies

The proportion of individuals that belong to category B_k for the variable y among the individuals that have the modality A_j for the variable x form the so called table of row profiles. The conditional frequencies for fixed j and all k are

$$f_{k|j} = \frac{n_{jk}}{n_{j.}} = \frac{n_{jk}/n}{n_{j.}/n} = \frac{f_{jk}}{f_{j.}}.$$

The frequency $f_{k|j}$ is the estimate of

$$p_{k|j} = P(y \in B_k | x \in A_j).$$

Row Profiles

	B_1	B_2	$\cdots$	B_K	
A_1	$\frac{f_{11}}{f_{1.}}$	$\frac{f_{12}}{f_{1.}}$	$\cdots$	$\frac{f_{1K}}{f_{1.}}$	1
A_2	$\frac{f_{21}}{f_{2.}}$	$\frac{f_{22}}{f_{2.}}$	$\cdots$	$\frac{f_{2K}}{f_{2.}}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	$\frac{f_{J1}}{f_{J.}}$	$\frac{f_{J2}}{f_{J.}}$	$\cdots$	$\frac{f_{JK}}{f_{J.}}$	1

Table: Row profiles

Example, Education and Salary

	B_1	B_2	B_3	
A_1	0.75	0.20	0.05	1
A_2	0.32	0.58	0.10	1
A_3	0.05	0.55	0.40	1

Table: Row profiles

- In this sample 75% of the individuals that have Primary School education, have low salary.
- In this sample, 5% of the individuals that have Primary School education, have high salary.
- In this sample, 55% of the individuals that have University education, have average salary.
- ...

Column Profiles

Tables of Conditional Frequencies

The proportion of individuals that belong to category A_j for the variable x among the individuals that have the modality B_k for the variable y form the table of column profiles. The conditional frequencies for fixed k and all j are

$$f_{j|k} = \frac{n_{jk}}{n_{.k}} = \frac{n_{jk}/n}{n_{.k}/n} = \frac{f_{jk}}{f_{.k}}.$$

The frequency $f_{j|k}$ is the estimate of

$$p_{j|k} = P(x \in A_j | y \in B_k).$$

Column Profiles

	B_1	B_2	$\cdots$	B_K
A_1	$\frac{f_{11}}{f_{.1}}$	$\frac{f_{12}}{f_{.2}}$	$\cdots$	$\frac{f_{1K}}{f_{.K}}$
A_2	$\frac{f_{21}}{f_{.1}}$	$\frac{f_{22}}{f_{.2}}$	$\cdots$	$\frac{f_{2K}}{f_{.K}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	$\frac{f_{J1}}{f_{.1}}$	$\frac{f_{J2}}{f_{.2}}$	$\cdots$	$\frac{f_{JK}}{f_{.K}}$
	1	1	$\cdots$	1

Table: Column profiles

Example, Education and Salary

	B_1	B_2	B_3
A_1	0.43	0.08	0.07
A_2	0.54	0.70	0.40
A_3	0.03	0.22	0.53
	1	1	1

Table: Column profiles

- In this sample 43% of the individuals that have low salary, have Primary School education.
- In this sample, 7% of the individuals that have high salary, have Primary School education.
- In this sample, 22% of the individuals that have average salary, have University education.
- ...

Dependence and Independence

Independence

The variables x and y are independent if and only if for all j, k it holds that

$$P(x \in A_j, y \in B_k) = P(x \in A_j)P(y \in B_k),$$

$$P(x \in A_j | y \in B_k) = P(x \in A_j),$$

and

$$P(y \in B_k | x \in A_j) = P(y \in B_k).$$

These equalities can be estimated by

$$f_{jk} \approx f_{j.} f_{.k},$$

$$f_{j|k} = \frac{f_{jk}}{f_{.k}} \approx f_{j.},$$

and

$$f_{k|j} = \frac{f_{jk}}{f_{j.}} \approx f_{.k},$$

respectively.

Theoretical Relative Frequencies Under Independence

We can now define the theoretical relative frequencies and theoretical frequencies under the assumption of independence as follows:

$$f_{jk}^* = f_{j.} f_{.k}$$

and

$$n_{jk}^* = \frac{n_{j.} n_{.k}}{n} = f_{jk}^* n.$$

Example, Education and Salary

	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: Observed frequencies

	B_1	B_2	B_3	
A_1	70	100	30	200
A_2	210	300	90	600
A_3	70	100	30	200
	350	500	150	1000

Table: Theoretical frequencies under independence

Example, Education and Salary

	B_1	B_2	B_3	
A_1	0.15	0.04	0.01	0.20
A_2	0.19	0.35	0.06	0.60
A_3	0.01	0.11	0.08	0.20
	0.35	0.50	0.15	1

Table: Observed relative frequencies

	B_1	B_2	B_3	
A_1	0.07	0.10	0.03	0.20
A_2	0.21	0.30	0.09	0.60
A_3	0.07	0.10	0.03	0.20
	0.35	0.50	0.15	1

Table: Theoretical relative frequencies under independence

Attraction Repulsion Matrix

Attraction Repulsion Matrix

The attraction repulsion indices

$$d_{jk} = \frac{n_{jk}}{n^*_{jk}} = \frac{f_{jk}}{f^*_{jk}} = \frac{f_{jk}}{f_{j.}f_{.k}}$$

can be used to measure dependencies between categorical variables. The attraction repulsion matrix D is a matrix whose elements are the attraction repulsion indices. The element ij of the matrix D is d_{jk}.

Attraction Repulsion Matrix

Note that

$$d_{jk} > 1 \;\Leftrightarrow\; f_{jk} > f_{j.} f_{.k} \;\Leftrightarrow$$
$$f_{j|k} > f_{j.} \text{ and } f_{k|j} > f_{k.}$$

and

$$d_{jk} < 1 \;\Leftrightarrow\; f_{jk} < f_{j.} f_{.k} \;\Leftrightarrow$$
$$f_{j|k} < f_{j.} \text{ and } f_{k|j} < f_{k.}$$

If $d_{jk} > 1$, then the modalities (categories) A_j and B_k are said to be attracted to each other. If $d_{jk} < 1$, then the modalities A_j and B_k are said to repulse each other.

Salary Example

	B_1	B_2	B_3
A_1	2.14	0.40	0.33
A_2	0.90	1.16	0.67
A_3	0.14	1.10	2.67

Table: Attraction repulsion indices

- High salary is more frequent for people with University education.
- High salary is less frequent for people with a Primary School education.
- Low salary is less frequent for people with University education.
- ...

Chi-square Test Statistic

Independence Testing

If the variables (for example salary level and education level) were independent of each other, it would not make sense to assess dependencies between the categories. One can start the analysis by independence testing to see whether there is statistically significant dependency between the variables.

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

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

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

Decomposition of the Chi-square Statistic

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

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

Thus, the matrix Z gives shifted and scaled relative frequencies of the variables. The variables are shifted and scaled such that the elements

$$Z_{jk} = \frac{f_{jk} - f_{j.}f_{.k}}{\sqrt{f_{j.}f_{.k}}} = \frac{f_{jk} - f^*_{jk}}{\sqrt{f^*_{jk}}} = \frac{n_{jk} - n^*_{jk}}{\sqrt{n^*_{jk}}}$$

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

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

Next Week

Next week we will continue discussion about correspondence analysis.

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.