CORRESPONDENCE ANALYSIS

Statistics 407, ISU

DEFINITION

Correspondence analysis is a method for exploring associations between sets of categorical variables. Mathematically it is a method for breaking down the value of the χ^2 goodness-of-fit statistic into components due to the rows and columns of the contingency table. It can also be considered as a technique for assigned order to unordered categories.

CONTINGENCY TABLE

Var 1/Var 2	Cat 1	...	Cat J	Row Total
Cat 1	n_{11}	...	n_{1J}	$n_{1.}$
⋮	⋮		⋮	⋮
Cat I	n_{I1}	...	n_{I3}	$n_{I.}$
Column Total	$n_{.1}$	...	$n_{.3}$	n

$$\chi^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

where $e_{ij} = \frac{n_{i.}n_{.j}}{n}$.

MECHANICS

The table of components, $\mathbf{C}_{I \times J} : \; c_{ij} = \frac{n_{ij}-e_{ij}}{\sqrt{e_{ij}}}$

is decomposed using singular value decomposition

$$\mathbf{C} = \mathbf{U}\Delta\mathbf{V}'$$

The columns of $\mathbf{U}$ and $\mathbf{V}$ are plotted with the corresponding category labels displayed. Categories from each variable closest to each other are considered the most associated.

EXAMPLE

The data was collected to examine the relationship between a girl's age and her relationship with her boyfriend. Each of 139 girls have been classified into one of three groups (no boyfriend, boyfriend/no sexual intercourse, boyfriend/sexual intercourse), and the second variable is the girl's age (1 = 16 or less, 2=17, 3=18, 4=19, 5=20 or older).

	1	2	3	4	5
No boyfriend	21 (17.2)	21 (18.3)	14 (13.3)	13 (17.2)	8 (11.1)
Boyfriend/ No sex	8 (7.4)	9 (7.8)	6 (5.7)	8 (7.4)	2 (4.5)
Boyfriend/Sexual relationship	2 (6.5)	3 (6.9)	4 (5.0)	10 (6.5)	10 (4.2)

EXAMPLE

C

	1	2	3	4	5
No boyfriend	0.92	0.64	0.19	-1.01	-0.93
Boyfriend/ No sex	0.24	0.42	0.13	0.24	-1.26
Boyfriend/Sexual relationship	-1.76	-1.48	-0.45	1.39	2.85

χ^2=20.6, *p*-value=0.0003

The largest values of C are the category combinations which most contribute to the significance.

EXAMPLE

C

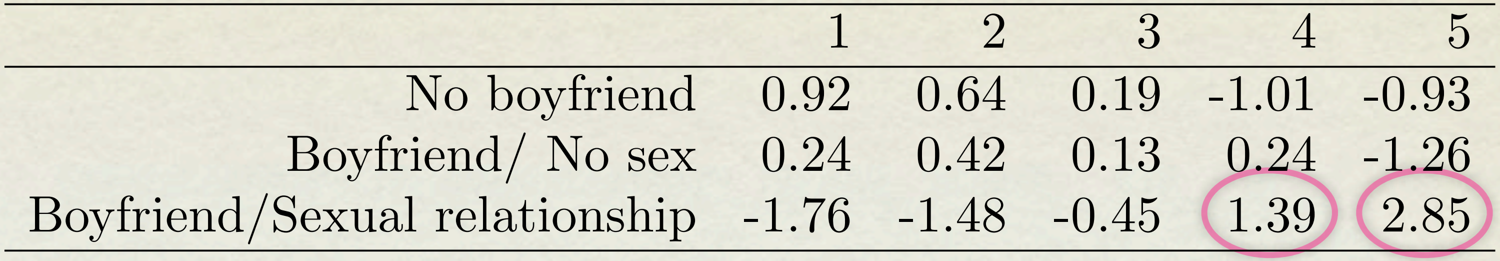
	1	2	3	4	5
No boyfriend	0.92	0.64	0.19	-1.01	-0.93
Boyfriend/ No sex	0.24	0.42	0.13	0.24	-1.26
Boyfriend/Sexual relationship	-1.76	-1.48	-0.45	1.39	2.85

χ^2=20.6, *p*-value=0.0003

The largest values of C are the category combinations which most contribute to the significance.

EXAMPLE

C



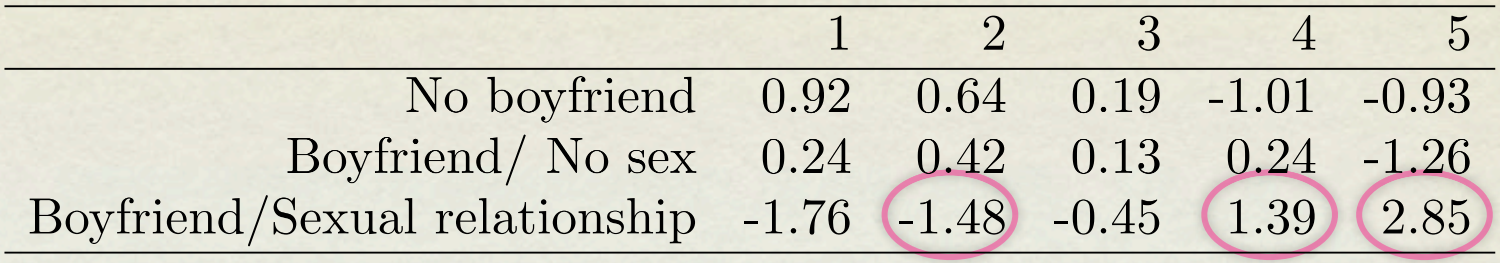
	1	2	3	4	5
No boyfriend	0.92	0.64	0.19	-1.01	-0.93
Boyfriend/ No sex	0.24	0.42	0.13	0.24	-1.26
Boyfriend/Sexual relationship	-1.76	-1.48	-0.45	1.39	2.85

χ^2=20.6, *p*-value=0.0003

The largest values of C are the category combinations which most contribute to the significance.

EXAMPLE

C



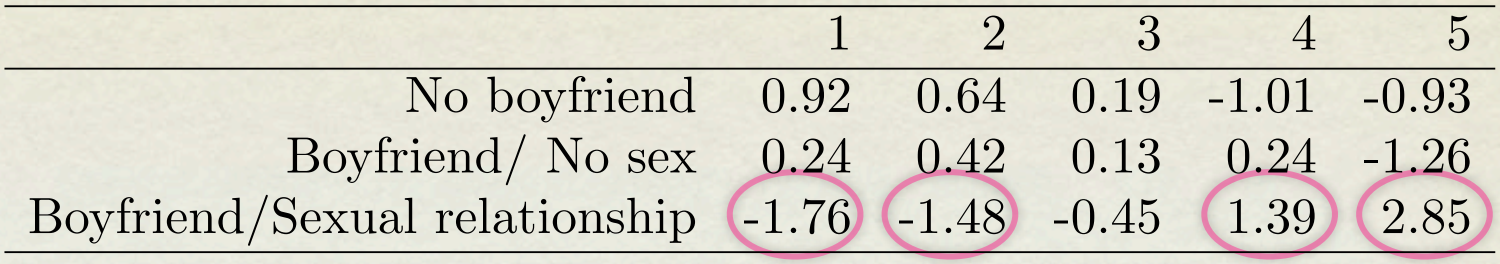
	1	2	3	4	5
No boyfriend	0.92	0.64	0.19	-1.01	-0.93
Boyfriend/ No sex	0.24	0.42	0.13	0.24	-1.26
Boyfriend/Sexual relationship	-1.76	-1.48	-0.45	1.39	2.85

χ^2=20.6, *p*-value=0.0003

The largest values of C are the category combinations which most contribute to the significance.

EXAMPLE

C

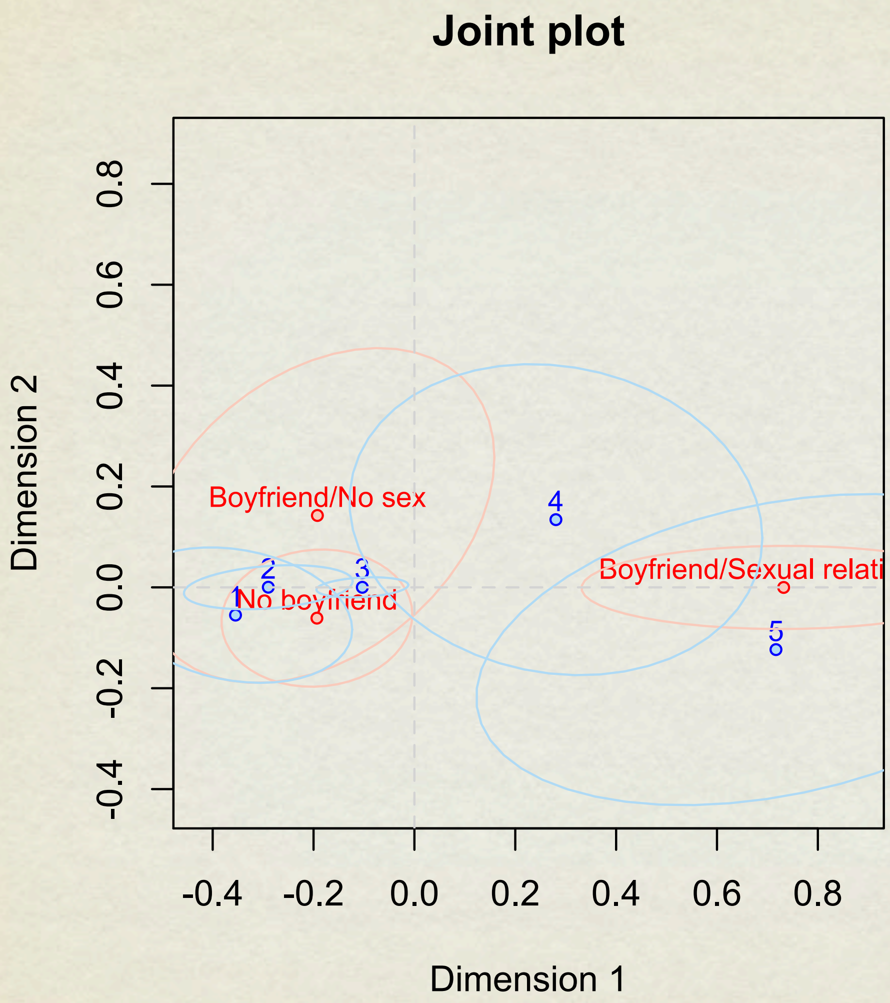


	1	2	3	4	5
No boyfriend	0.92	0.64	0.19	-1.01	-0.93
Boyfriend/ No sex	0.24	0.42	0.13	0.24	-1.26
Boyfriend/Sexual relationship	-1.76	-1.48	-0.45	1.39	2.85

χ^2=20.6, *p*-value=0.0003

The largest values of C are the category combinations which most contribute to the significance.

EXAMPLE

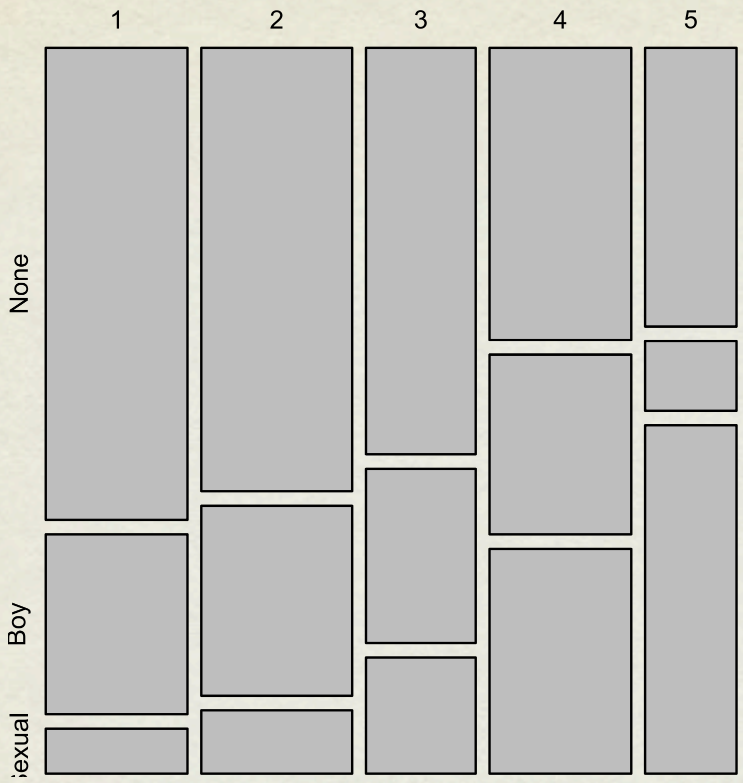


- Youngest age group is most associated platonic relationships or none!
- Older age group most associated with sexual relationships.

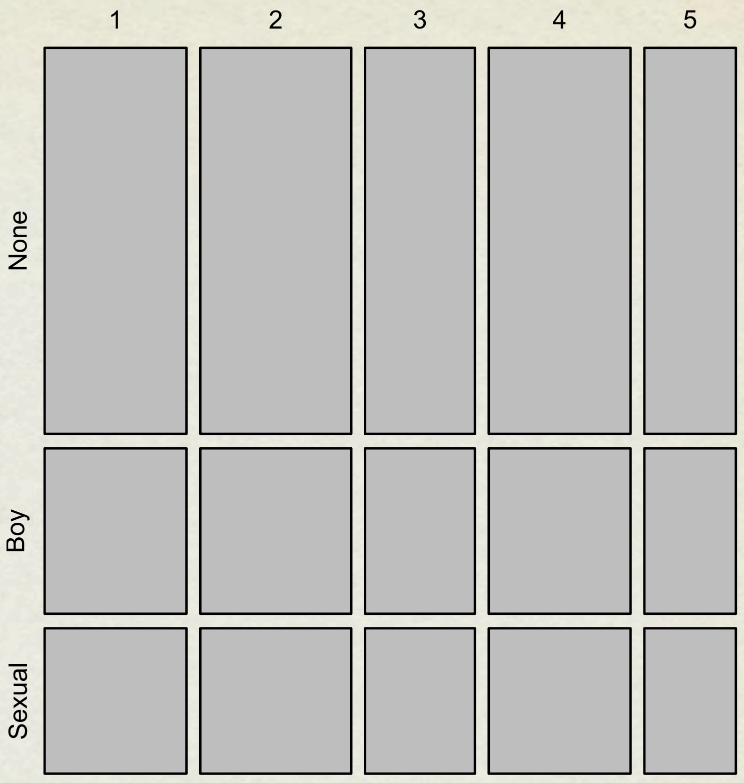
INTERPRETATION

- When the items are both large and positive then the corresponding row and column will have a large contribution to the test statistic value, and these two are said to be positively associated.
- When the items are both large but have different signs then the corresponding rows and columns are said to be negatively associated.
- When the items have both got values close to 0 then the association is close to the expected value under an assumption of independence.

ALTERNATIVE PLOT



Observed



Expected

Same basic association conclusions.