

Introducing Configural Frequency Analysis

Daniel Wiechmann – Summer 2008

What is configural frequency analysis (cfa) ?

- CFA is a multivariate statistical method that
 - identifies **INDIVIDUALS (CONFIGURATIONS)** that are, statistically speaking, **SPECIAL**
 - may be applied in either a confirmatory or **EXPLORATORY** fashion
 - Exploratory approaches typically test simultaneously all possible local null hypotheses
 - Confirmatory CFA tests each cell wrt whether F_{obs} differs sig. from F_{exp}
 - can look at descriptors either simultaneously or hierarchically (cfa versus **HCFA**)

On the analysis of contingency tables

- Two main approaches
 - (1) Relationships among variables
 - includes coefficients of association that can be interpreted in a manner analogous to correlations
 - Log-linear modeling
 - Chi-square decomposition techniques
 - yields results that target variables and their interaction relationships

On the analysis of contingency tables

- Two main approaches
 - (2) Patterns of characteristics
 - Focuses on **groups of subjects**
 - Groups contain subjects who differ from all other subjects in that they display a unique pattern of characteristics (configuration)
 - CFA allows one to **identify such groups**
 - As corpus-linguists, we are usually not so much interested in comparing (groups of) subjects
 - **patterns of characteristics (configurations)** can also be identified at the level of **complex linguistic units** (say complex sentences)

Analysing Contingency Tables

		DEFINITENESS		
		TRUE	FALSE	
ANIMACY	TRUE	50 (the man)	20 (a man)	70
	FALSE	10 (the chair)	40 (a chair)	50
		60	60	120

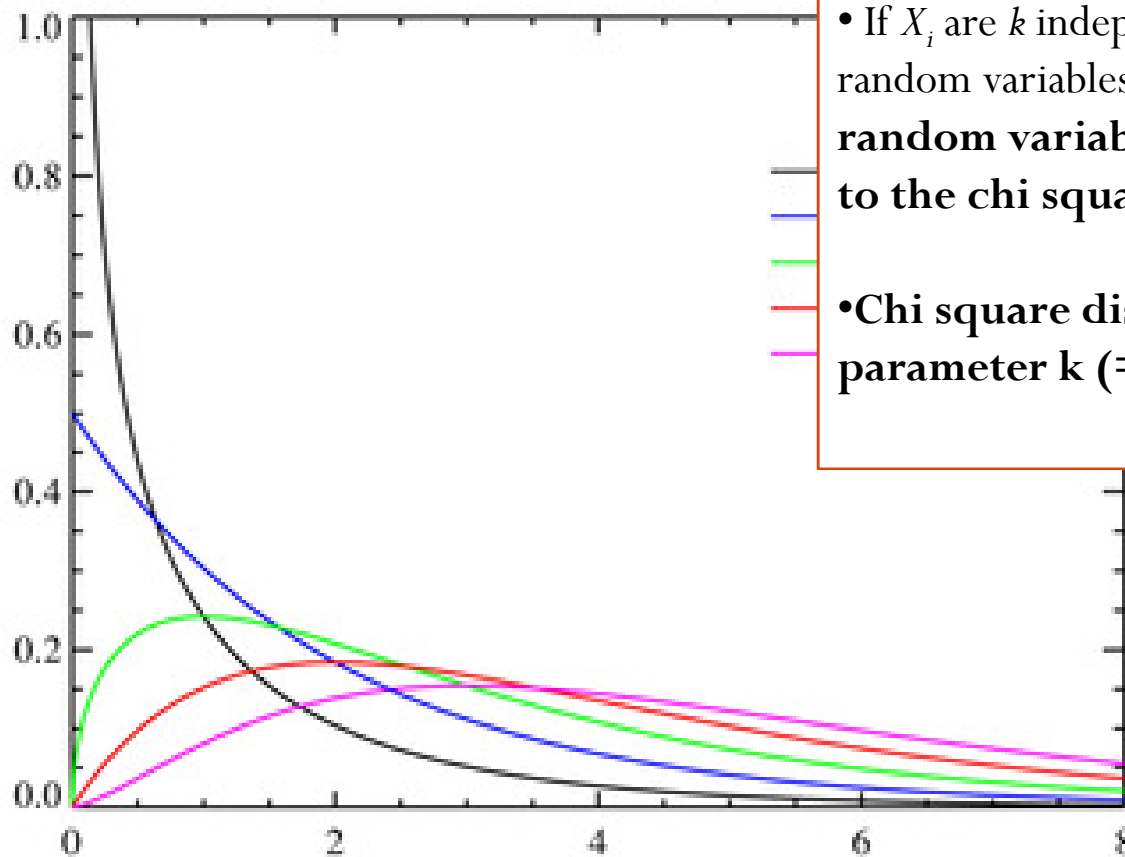
Analysing Contingency Tables

		DEFINITENESS		
		TRUE	FALSE	
ANIMACY	TRUE	50 35	20 35	70
	FALSE	10 25		
		60	60	120

Frequency distribution of certain events in a sample is consistent with a particular theoretical distribution, the **CHI SQUARE DISTRIBUTION**

Chi square distribution

probability density function



- If X_i are k independent, normally distributed random variables (mean =0, variance =1), then **the random variable Q is distributed according to the chi square distribution**

- **Chi square distribution has just 1 parameter k (=degrees of freedom)**

$$Q = \sum_{i=1}^k X_i^2$$

$$Q \sim \chi_k^2$$

Chi square distribution

$$Q = \sum_{i=1}^k X_i^2$$

$$Q \sim \chi_k^2.$$

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Analysing Contingency Tables

		DEFINITENESS		
		TRUE	FALSE	
ANIMACY	TRUE	50 (the man)	20 (a man)	70
	FALSE	10 (the chair)	40 (a chair)	50
		60	60	120

X-squared = 28.8343, df = 1, p < 7.8 E-08,

Analysing Contingency Tables

		DEFINITENESS		
		TRUE	FALSE	
ANIMACY	TRUE	5 (the man)	2 (a man)	7
	FALSE	1 (the chair)	4 (a chair)	5
		6	6	12

X-squared = 1.3741, df = 1, p < 0.24

Analysing Contingency Tables

Global versus local measures of independence

```
R Console
> a<-matrix(c(500,100,200,400), nrow=2, ncol=2)
> a
      [,1] [,2]
[1,]  500  200
[2,]  100  400
> b<-chisq.test(a)
> b

      Pearson's Chi-squared test with Yates' continuity correction

data:  a
X-squared = 306.5177, df = 1, p-value < 2.2e-16

> round(b$residuals^2, 2) # Chi-square Components
      [,1] [,2]
[1,]  64.29 64.29
[2,]  90.00 90.00
> |
```

X-squared = 306.5177, df = 1, p < 2.2 E-16

Analysing contingency tables

p-values & effect sizes

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Hence, χ^2 gets larger as N gets larger
and p gets smaller as χ^2 gets larger

Corresponding
effect size measure
Cramer's ϕ

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Analysing Contingency Tables

Corresponding
effect size measure

Cramer's Φ

$$\Phi = \sqrt{\frac{\chi^2}{N}}$$

Analysing Contingency Tables

Corresponding
effect size measure

Cramer's Φ

$$\Phi = \sqrt{\frac{\chi^2}{N}}$$

alternative
effect size measure

ODDS RATIO

$$R(A : B) = \frac{R(A)}{R(B)} = \frac{\frac{P(A)}{1-P(A)}}{\frac{P(B)}{1-P(B)}} = \frac{P(A) \cdot (1 - P(B))}{P(B) \cdot (1 - P(A))}$$

Analysing Contingency Tables

Corresponding
effect size measure

Cramer's ϕ

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

alternative
effect size measure

ODDS RATIO

$$\text{odds-ratio} = \log \frac{O_{11}O_{22}}{O_{12}O_{21}}$$

Asymptotic and exact hypothesis test

```
R R Console
> plot(a)
> chisq.test(a)

      Pearson's Chi-squared test with Yates' continuity correction

data:  a
X-squared = 306.5177, df = 1, p-value < 2.2e-16

> fisher.test(a)

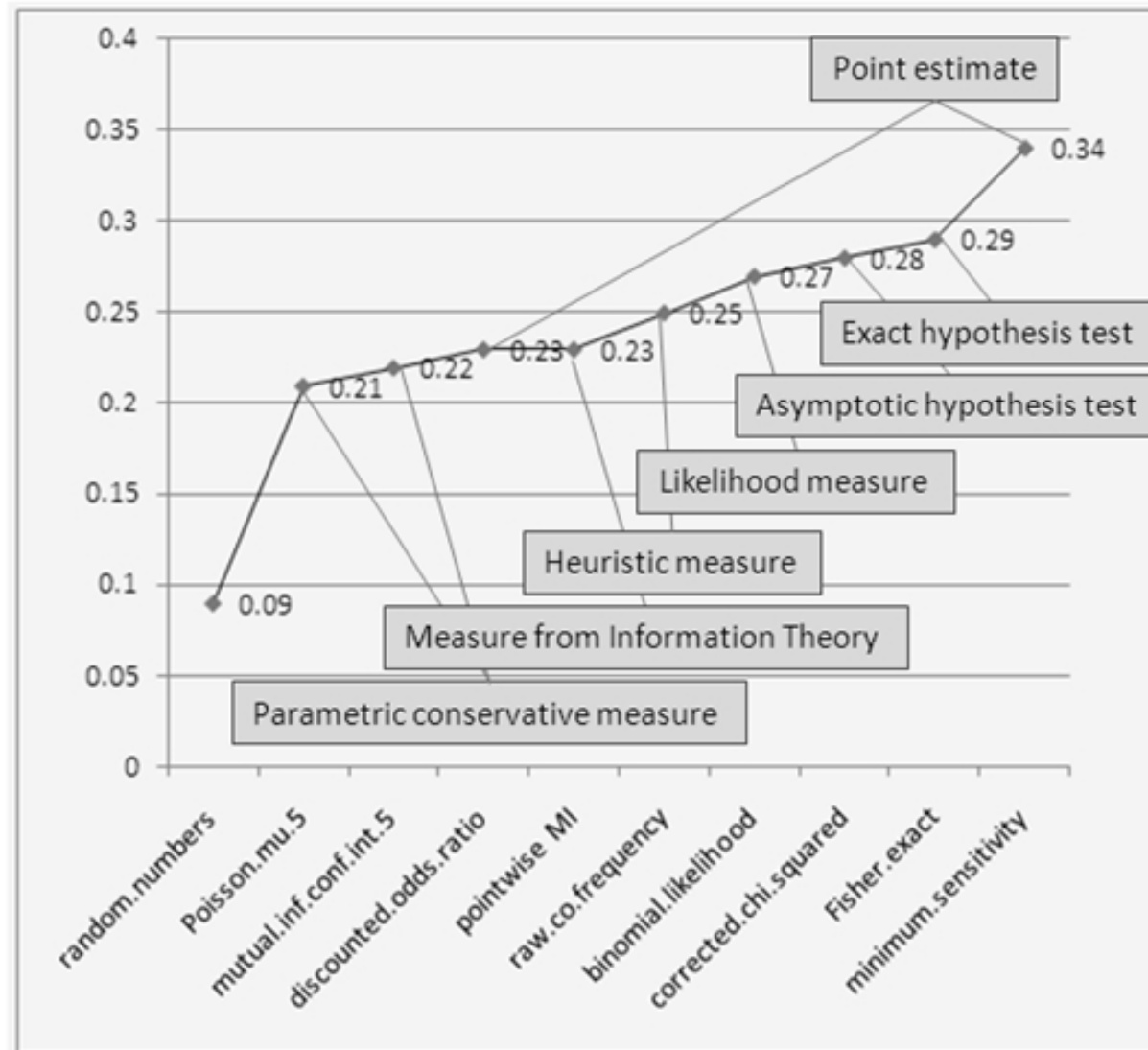
      Fisher's Exact Test for Count Data

data:  a
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 7.540593 13.287310
sample estimates:
odds ratio
 9.974676

> |
```

Comparing association measures

Wiechmann (to appear)



Contingency tables

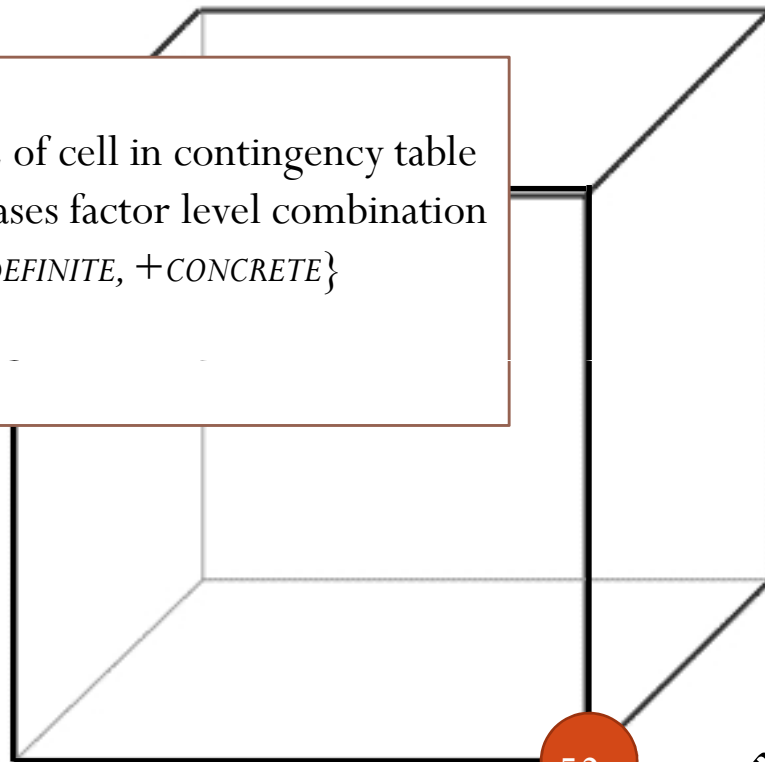
		DEFINITENESS	
		TRUE	FALSE
ANIMACY	TRUE		
	TRUE	50	20
	FALSE	10	40
	FALSE		

Adding dimensions:

Contingency cubes (d=3) and beyond

VALUE of a corner = VALUE of cell in contingency table representing the number of cases factor level combination of, say, $\{+ANIMATE, -DEFINITE, +CONCRETE\}$

ANIMATE



CONCRETENESS

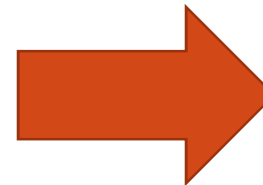
DEFINITENESS

52

Adding dimensions: Contingency cubes and beyond

animacy.head	definiteness.head	concreteness.head
inanimate.head	definite.head	abstract.head
inanimate.head	definite.head	abstract.head
inanimate.head	indefinite.head	abstract.head
inanimate.head	indefinite.head	abstract.head
inanimate.head	definite.head	abstract.head
inanimate.head	indefinite.head	abstract.head
inanimate.head	indefinite.head	abstract.head
inanimate.head	indefinite.head	abstract.head
inanimate.head	definite.head	abstract.head
inanimate.head	definite.head	abstract.head
...

data frame



Configuration	F_obs	F_exp
TTT	50	25
TTF	0	25
TFT	0	25
TFF	50	25
FTT	0	25
FTF	50	25
FFT	50	25
FFF	0	25

tabular format

Testing for types and antitypes

animacy	0				1			
definiteness	0		1		0		1	
concreteness	0	1	0	1	0	1	0	1
counts n_{ijk}	1	10	11	40	35	17	8	19

GOAL of Configurational Frequency Analysis:

Evaluate frequency distribution
 not at a *global* level (=whole table)
 but at a *local* level (=level of individual cells)

Evaluates particular states of a given state space

Types and Antitypes

- CFA compares the observed to expected frequencies in a cross tabulation
- Goal of this comparison is to determine whether the difference between the obs and exp frequency for a given configuration is larger than some critical value and is statistically significant
- $F_{\text{obs}} > F_{\text{exp}}$ is TYPE
- $F_{\text{obs}} < F_{\text{exp}}$ is ANTITYPE

CFA output

Configuration	F_obs	F_exp	$z (= (o-e)/e^{1/2})$	p(z)	T/A
TTT	38	37.4	0.049	0.9609	
TTF	52	33.01	3.304	0.001	T
TFT	23	48.14	3.623	0.0003	A
TFF	48	42.15	0.901	0.3676	
FTT	39	47.07	1.176	0.2396	
FTF	30	41.22	1.747	0.0806	
FFT	93	60.09	4.245	0.00002	T
FFF	39	52.62	1.878	0.0604	

How to proceed: exploratory CFA

1. Choose descriptor variables (features)
2. Fix scale of measurement
3. Determine sample size
 - Rule of thumb: $N = 5 * 2^d$
4. Collect and prepare data (-> contingency table)
5. Calculate test statistics
 - Global & local chi square
6. Determine α (Holm adjustment)
7. Interpret results

[Link to example](#)

Some things I am NOT going to talk about
but which we need to think about...

Some problems with induction...

- What exactly is **probability**?
 - **Frequentist** interpretation
 - well-defined random experiments
 - **Bayesian** interpretation
 - degrees of belief; prior probabilities
 - no random process needs to be involved
- **Conditional probability fallacy**
 - **$P(\text{data} \mid \text{hypothesis}) \neq P(\text{hypothesis} \mid \text{data})$**
 - Bayes theorem

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- Does refuting a NIL hypothesis help us to build theories



Reverend Thomas Bayes
(1702–1761)

Criteria for the description of groups

- Subjects can be grouped so that the groups are either disjunct or overlapping
 - **Disjunctive classification subjects belong to exactly one group**
 - Overlapping classification they may belong to more than one group
- Groups of subjects can be analyzed with respect to either (1) *similarity* or (2) *spatial distance* among group members
 - (1): use coefficients of similarity
 - (2): distance measures

Basic forms of configural frequency analysis

- There are hierarchical and non-hierarchical versions of CFA
 - hCFA systematically excludes variables from the analysis that contribute little to the contribution of *types* and *antitypes*
 - non-hCFA uses all variables simultaneously

CFA and cluster analysis

- In contrast to (hierarchical) cluster analytical method (e.g. Ward), CFA does not lead to solely descriptive statements
- Results of CFA are both *descriptive* and *inferential*
 - Descriptive: they result in labels for a configuration
 - Inferential: they assign a probability to the frequency of a given configuration relative to some expected frequency

Using CFA

- CFA can be used exploratively or **confirmatively**
 - Confirmatory CFA restricts testing to an a priori specified number of configurations