Coherence relations and DRD identification: theory and analysis

Ted Sanders (Utrecht University)
Wilbert Spooren (Radboud University Nijmegen)

Radboud University

Testing patterns in the coprus analysis of crs and drd's: Statistical methods part 1 Chi2 and loglinear analysis

Statistical modeling

- We make models of the data according to our hypotheses
- The fit of the model determines how valid our hypotheses are

- Two crucial issues:
  - How wel does the model predict the data? (what is the fit of the model?)
  - Do our variables of interest contribute significantly to the prediction?

Analysis of frequencies: χ^2

- χ^2 for the analysis of relations
  - Is there a relation between two nominal variables?
    - e.g.: number of passes as a function of gender
- χ^2 as Goodness-of-fit test: do the data fit the model?
  - i.e. do the observed frequencies resemble the frequencies that were predicted by the model (the expected frequencies)?
- In χ^2 : Model = null hypothesis
  - expected frequencies reflect the null hypothesis
- Analysis of relations: big χ^2 means “yes there are relations between the two nominal variables”
- Goodness-of-fit test: big χ^2 means “no the model does not fit the data”

How to test this?

- if the H_0-model is correct, then the frequencies in the crosstable are independent of the categories

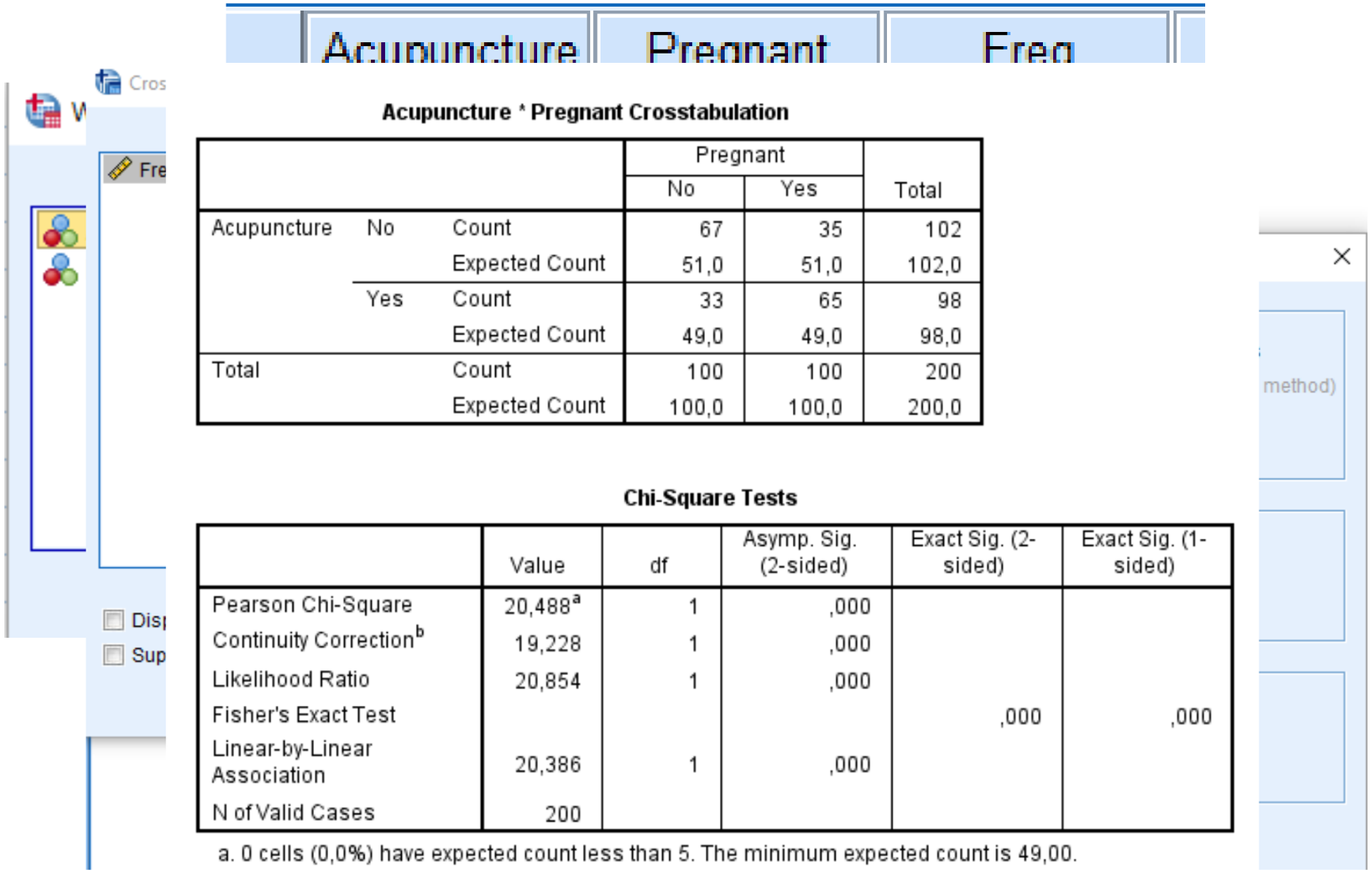
	man	woman	
pass	a	b	a+b
fail	c	d	c+d
	a+c	b+d	n

i.e. $a_{expected} = (a+b)*(a+c)/n$; $b_{expected} = (a+b)/(b+d)/n$, etc.

χ^2 in action

- Calculate for each cell in the crosstable the difference between observed frequency and expected frequency $(O - E)$
- Square the difference $(O - E)^2$
- Standardize the difference by dividing it by the expected frequency $\frac{(O - E)^2}{E}$
- Sum all the standardized squared differences $\sum \frac{(O - E)^2}{E}$
- The result is χ^2
- Calculate the probability of this χ^2, given the null hypothesis that there is no relation between the categories
- Degrees of freedom of a crosstable: (R-1)(C-1).
  - in a 2 x 2 table the df = 1; the critical value of χ^2 is 3.84.

An example



Acupuncture	Pregnant	Freq

Acupuncture * Pregnant Crosstabulation

			Pregnant		
			No	Yes	Total
Acupuncture	No	Count	67	35	102
		Expected Count	51,0	51,0	102,0
	Yes	Count	33	65	98
		Expected Count	49,0	49,0	98,0
Total		Count	100	100	200
		Expected Count	100,0	100,0	200,0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	20,488[a]	1	,000		
Continuity Correction[b]	19,228	1	,000		
Likelihood Ratio	20,854	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	20,386	1	,000		
N of Valid Cases	200				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 49,00.

b. Computed only for a 2x2 table

Assumptions behind χ^2

- You cannot use χ^2 if
  - more than 20% of the expected frequencies is less than 5;
  - one or more of the expected frequencies is less than 1;
  - there is a dependency between the data
    - every single datapoint can only contribute to one cell in the crosstable.

Effect size

- Effect size: how important is the effect?
- SPSS gives Phi, Cramer's V (and lambda)
- Cramer's V most general
  - resembles a correlation coefficient (between 0 and 1)
  - Rule of thumb: .10 = small, .30 = medium, .50= large.
- Recommendation by Andy Field (Discovering Statistics)
  - for 2 x 2-tables: odds ratio
  - for larger tables: Cramer's V

Effect size

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,320	,000
	Cramer's V	,320	,000
	Contingency Coefficient	,305	,000
N of Valid Cases		200	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Effect size

- Odds ratio: how probable is it that being pregnant and having had acupuncture go together?

Acupuncture? * Pregnant? Crosstabulation

Count

		Pregnant?		
		No	Yes	Total
Acupuncture?	No	67	35	102
	Yes	33	65	98
Total		100	100	200

Odds pregnant after acupuncture: 65/33=1.970
Odds pregnant after no acupunture: 35/67=.552
Odds ratio = 1.970/.552 = 3.57
“It is 3,57 times as likely to to be pregnant after acupuncture than after no acupuncture”

How to report χ^2?

- “There was a significant relationship between <row variable> and <column variable> (χ^2 (df) = <value>, $p < .05$): following the odds ratio <research objects> were <odds ratio> times as likely to <colum variable, level 1> after <row variable, level 1> than after <row variable, level 2>”

- “There was a significant relationship between receiving acupuncture and being pregnant (χ^2 (1) = 20.49, $p < .05$): following the odds ratio it was 3.57 times as likely that women were pregnant after having received acupuncture than after not having received acupuncture.”

Interpretation χ^2

- If there is a r x k table with a significant χ^2, is it possible to analyze in detail which cells contribute to that χ^2?
  - Yes: look at the standardized residuals
  - Analyze > Crosstabs > Cells > standardized residuals
  - Gives standardized deviation of the expected frequency; can be interpreted as a z-score (i.e. significant if |standardized residual| > 1.96)
  - to be precise, s.r. =

$$\frac{O - E}{\sqrt{E}}$$

Interpretation χ^2

genre * pattern Crosstabulation

			pattern				Total
			C	XC	CX	XCX	
genre	acdamic pr	Count	666	449	207	116	1438
		Std. Residu	-7,0	12,9	-2,8	6,2	
	newspaper	Count	1264	312	221	47	1844
		Std. Residu	4,3	-0,2	-5,7	-4,0	
	short storie	Count	1911	353	751	154	3169
		Std. Residu	-0,3	-8,2	8,3	0,8	
	leaflets	Count	1044	265	231	49	1589
		Std. Residu	2,5	-0,5	-2,9	-2,7	
Total		Count	4885	1379	1410	366	8040

Loglinear analysis

- Chi-Square: association between two nominal variables
- Loglinear analysis: more than two nominal variables
  - advantage: you can study both main effects and interactions

Loglinear analysis

- Main effects and interactions
  - Main effect: difference in frequency between the levels of one variable (#men > #women)
  - Interaction: difference in frequency between levels of one variable differs for the levels of a second variable (i.e,, there is an association between two variables)
    - Summer: #men > #women, Winter: #men = #women
  - 3-way interaction: association between two variables depends on a third variable
    - Tennis: Summer: #men > #women, Winter: #men = #women
    - Swimming: Summer: #men = #women, Winter: #men < #women

Loglinear analysis

- Model selection: Backward Elimination
  - Goal: Find the best fitting model, with the smallest number of significant parameters (i.e., with the smallest number of effects)
  - Start with the complete model (“Saturated Model”): contains all main effects and interaction effects → Perfect Fit
  - Try to eliminate effects from the mode, to begin with the highest order interaction effect
  - Stop eliminating effects if deleting the next effect significantly reduces the fit of the model

Example with SPSS

- Genre (Narrative/Expository), Relation (Additive/Causal), Explicit Marking (Present/Abstent)
- Possible interactions and main effects?
  - Genre*Relation*Marking (3-way)
  - Genre*Relation (2-way)
  - Genre*Marking (2-way)
  - Relation*Marking (2-way)
  - Genre (main effect)
  - Relation (main effect)
  - Marking (main effect)

Example in SPSS: Data

Genre	narrative				expository			
Relation	additive		causal		addative		causal	
Marking	absent	present	absent	present	absent	present	absent	present
	47	10	4	10	58	17	10	20

Example with SPSS: Data

*Untitled1 [DataSet0] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

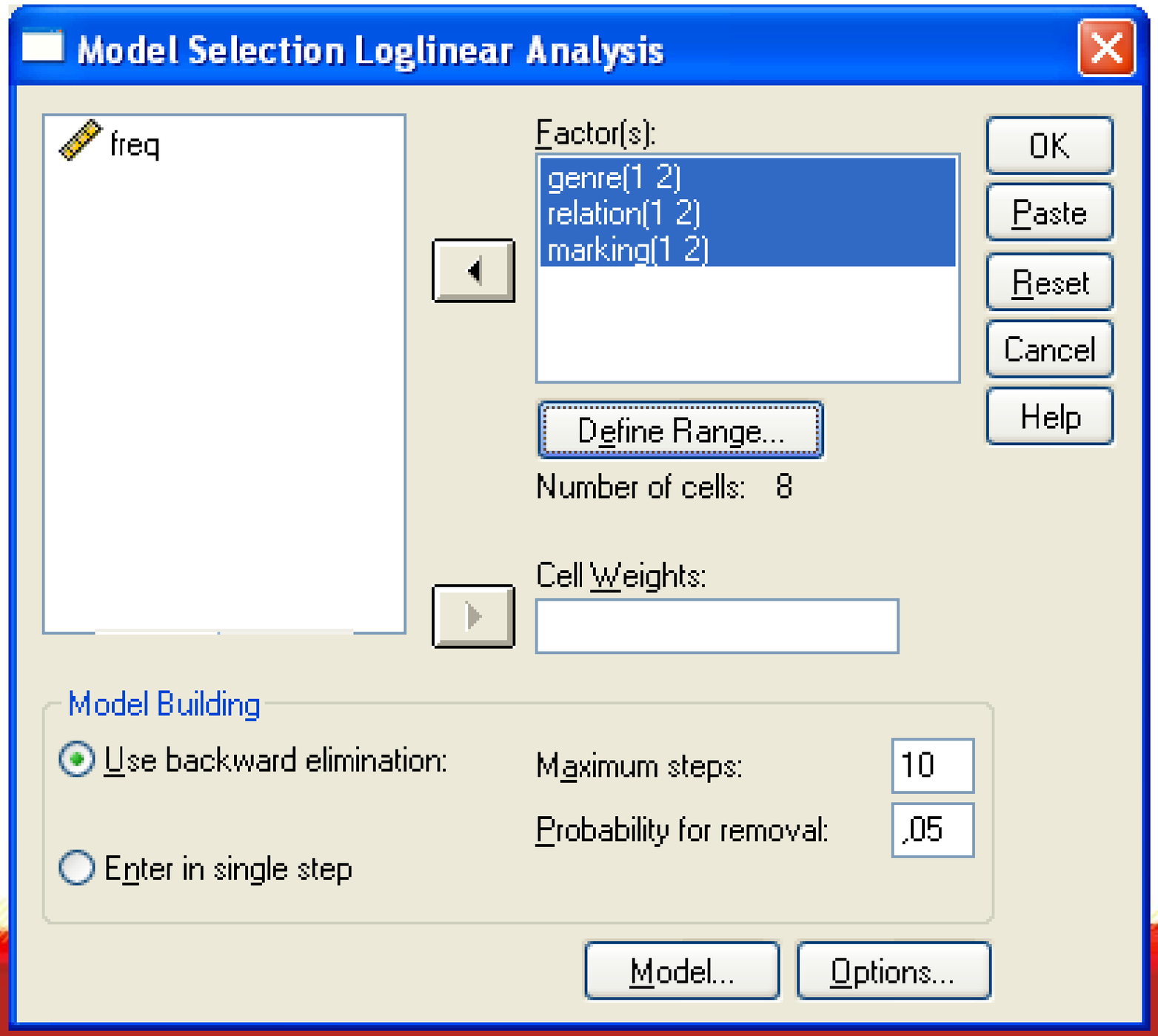
8 : freq | 20

	genre	relation	marking	freq	var	var	var	var	var	var
1	narrative	additive	absent	47						
2	narrative	additive	present	10						
3	narrative	causal	absent	4						
4	narrative	causal	present	10						
5	expository	additive	absent	58						
6	expository	additive	present	17						
7	expository	causal	absent	10						
8	expository	causal	present	20						
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										

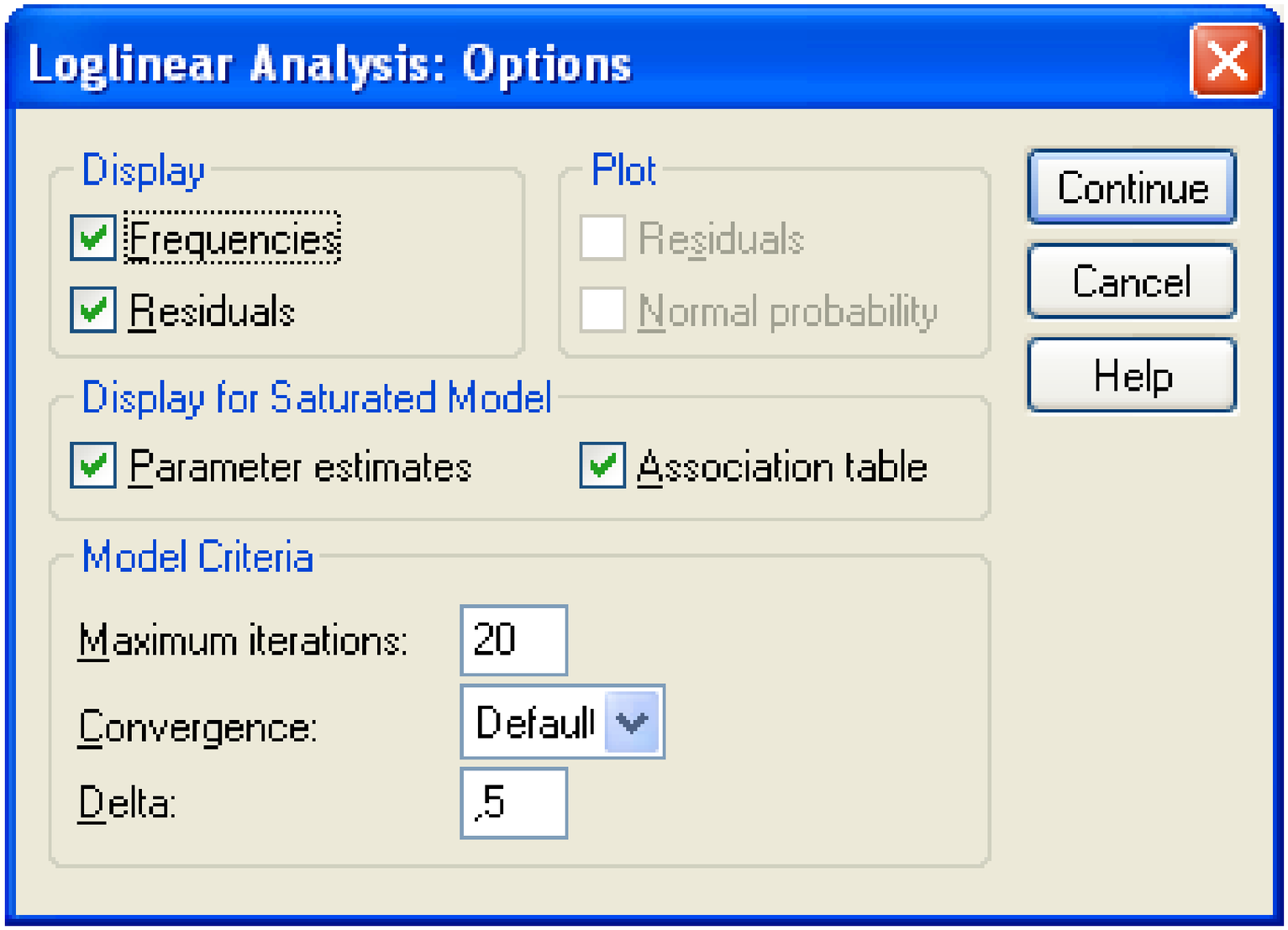
Data View | Variable View

SPSS Processor is ready | Weight On

Example with SPSS



Example with SPSS



Output of SPSS

Cell Counts and Residuals

genre	relation	marking	Observed Count[a]	Observed %	Expected Count	Expected %	Residuals	Std. Residuals
narrative	additive	absent	47,500	27,0%	47,500	27,0%	,000	,000
		present	10,500	6,0%	10,500	6,0%	,000	,000
	causal	absent	4,500	2,6%	4,500	2,6%	,000	,000
		present	10,500	6,0%	10,500	6,0%	,000	,000
expository	additive	absent	58,500	33,2%	58,500	33,2%	,000	,000
		present	17,500	9,9%	17,500	9,9%	,000	,000
	causal	absent	10,500	6,0%	10,500	6,0%	,000	,000
		present	20,500	11,6%	20,500	11,6%	,000	,000

a. For saturated models, ,500 has been added to all observed cells.

Output of SPSS

Goodness-of-Fit Tests

	Chi-Square	df	Sig.
Likelihood Ratio	,000	0	.
Pearson	,000	0	.

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects[a]	1	7	110,282	,000	123,000	,000	0
	2	4	35,310	,000	37,077	,000	2
	3	1	,431	,512	,425	,514	4
K-way Effects[b]	1	3	74,972	,000	85,923	,000	0
	2	3	34,879	,000	36,651	,000	0
	3	1	,431	,512	,425	,514	0

a. Tests that k-way and higher order effects are zero.

b. Tests that k-way effects are zero.

Output of SPSS

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
genre*relation	1	1,029	,310	2
genre*marking	1	,198	,656	2
relation*marking	1	32,098	,000	2
genre	1	6,610	,010	2
relation	1	46,046	,000	2
marking	1	22,317	,000	2

Output of SPSS

Step Summary

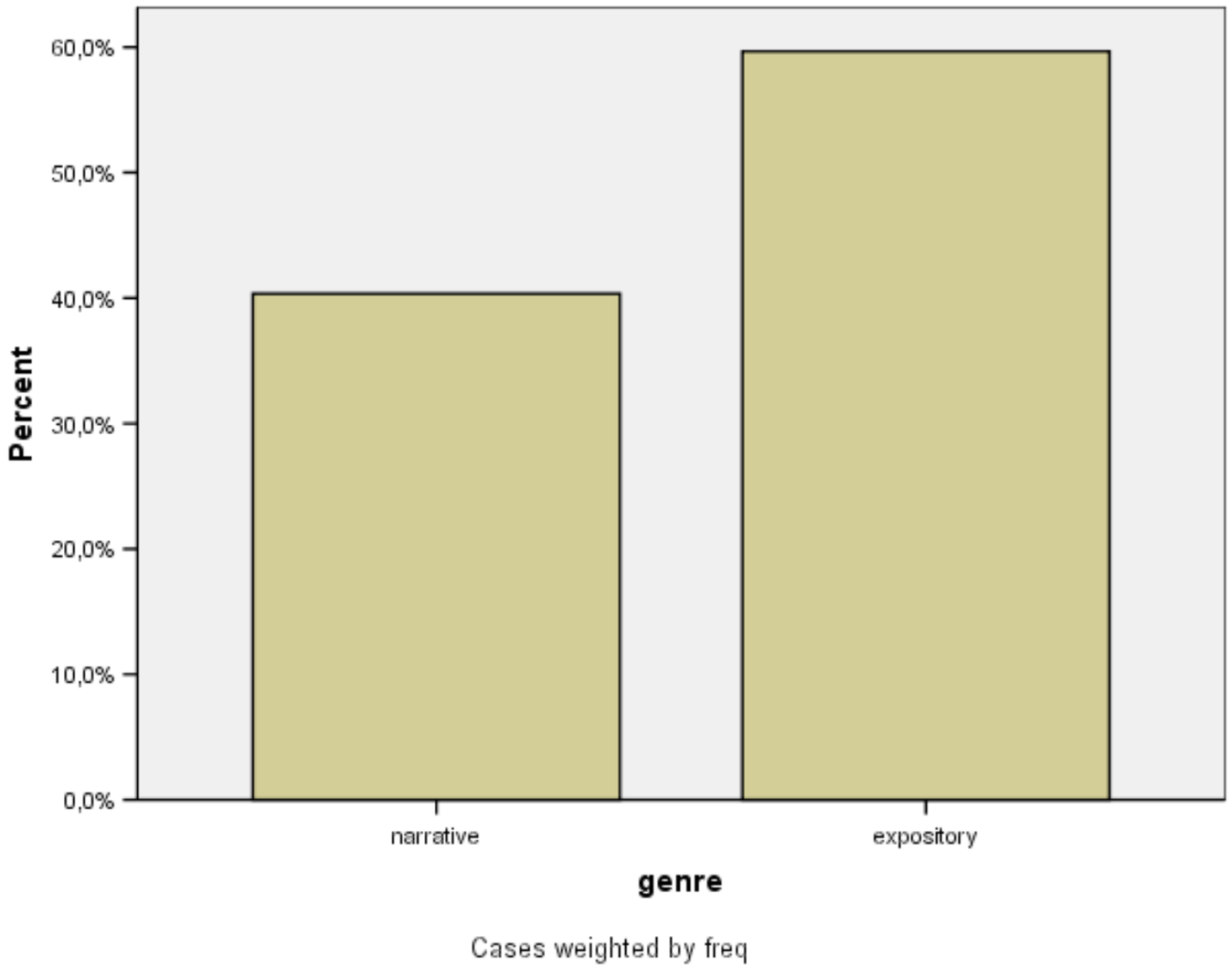
Step[b]			Effects	Chi-Square[a]	df	Sig.	Number of Iterations
0	Generating Class[c]		genre*relation*marking	,000	0	.	
	Deleted Effect	1	genre*relation*marking	,431	1	,512	4
1	Generating Class[c]		genre*relation, genre*marking, relation*marking	,431	1	,512	
	Deleted Effect	1	genre*relation	1,029	1	,310	2
		2	genre*marking	,198	1	,656	2
		3	relation*marking	32,098	1	,000	2
2	Generating Class[c]		genre*relation, relation*marking	,629	2	,730	
	Deleted Effect	1	genre*relation	1,806	1	,179	2
		2	relation*marking	32,875	1	,000	2
3	Generating Class[c]		relation*marking, genre	2,435	3	,487	
	Deleted Effect	1	relation*marking	32,875	1	,000	2
		2	genre	6,610	1	,010	2
4	Generating Class[c]		relation*marking, genre	2,435	3	,487	

a. For 'Deleted Effect', this is the change in the Chi-Square after the effect is deleted from the model.

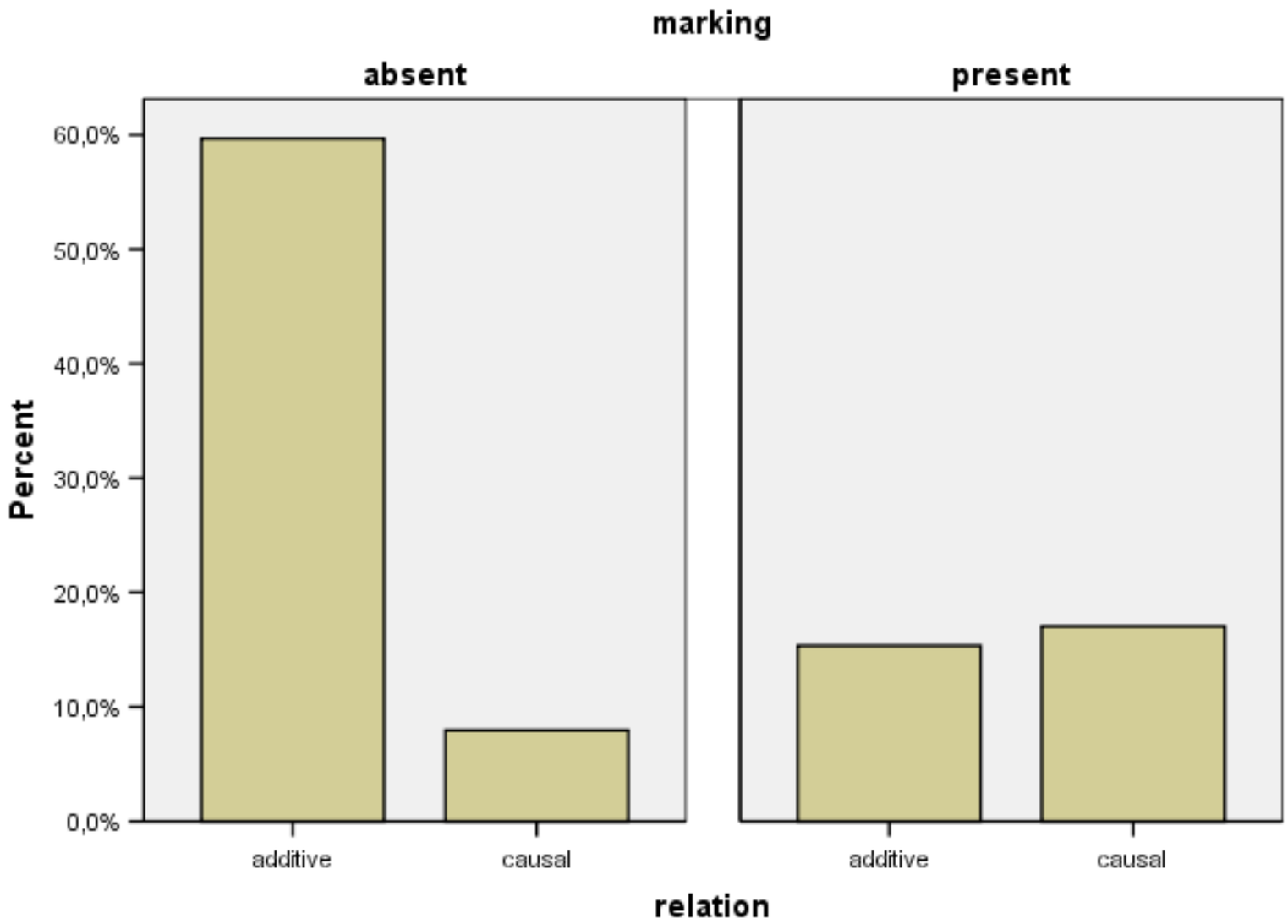
b. At each step, the effect with the largest significance level for the Likelihood Ratio Change is deleted, provided the significance level is larger than ,050.

c. Statistics are displayed for the best model at each step after step 0.

Interpretation loglinear analysis: main effect of genre



Interpretation loglinear analysis: 2-way interaction relation * marking



Cases weighted by freq

Interpretation loglinear analysis: strength of the effect

genre

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	narrative	71	40,3	40,3	40,3
	expository	105	59,7	59,7	100,0
	Total	176	100,0	100,0	

Odds ratio:
expository/narrative= 105/71 = 1.48
“It is 1.48 times likelier that a connective occurs in an expository genre than that it occurs in a narrative genre.”

Interpreting loglinear analysis: strength of the effect

relation * marking Crosstabulation

Count

		marking		
		absent	present	Total
relation	additive	105	27	132
	causal	14	30	44
Total		119	57	176

Odds ratio:
Odds additive if marking is present 27/30 = 0.90
Odds additive if marking is absent 105/14 = 7.50
Odds ratio = 7.50/0.90 = 8.33
“It is 8.33 keer times likelier that a relation is additive if a marking is absent than if a marking is present”

Interpretation loglinear analysis: strength of the effect

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for relation (additive / causal)	8,333	3,888	17,862
For cohort marking = absent	2,500	1,608	3,886
For cohort marking = present	,300	,203	,444
N of Valid Cases	176		

Analyze > Crosstabs > Statistics > Risk

Rapporting loglinear analysis

- “The 3-way loglinear analysis produced a model containing two effects: a main effect of genre ($\chi^2(1)= 6.61$, $p < .05$) and a 2-way interaction between marking and relation ($\chi^2(1)= 32.10$, $p < .01$). The goodness-of-fit of the resulting model was $\chi^2(3) = 2.44$, $p = .49$. The main effect of genre reflects the fact that there are more expository texts in the corpus (69.7 %) than narrative texts (40.3 %). The interaction between marking and relation can be interpreted in terms of an odds ratio: It is 8.33 times likelier that a relation is additive if a marking is absent than if a marking is present.”

Assumptions of loglinear analysis

- Same as χ^2:
  - Not more than 20 % of the cells in the matrix have expected frequencies < 5
  - No cell has an expected frequency < 1
  - Independence of data
- What if assumptions are violated?
  - One can consider joining categories, so that expected frequencies increase (but only if that can be motivated theoretically) (cf. Field, 2013, p. 736)