Odds Ratios

applied to Negative Polarity Items

Methodology and Statistics
Laura Handojo

Outline

- Negative Polarity Items
- Research Question
- Data
- Odds Ratios
- Statistical Analysis

Negative Polarity Items

- Negative Polarity Items (NPIs)
  - can only occur in negative contexts
    - *He hasn't seen any students*
    - * *He has seen any students*
  - single words or word groups
    - *ever* vs. *lift a finger*
  - various parts of speech
    - advers, verb phrases, noun phrases etc.

Negative Contexts

- called licensers, include:
  - negation
  - neg raising verbs (e.g. *think*)
  - N-words (e.g. *never*, *nobody*)
  - negative verbs (e.g. *doubt*)
  - negative conjunctions (e.g. *without*)
  - conditionals
  - universal quantifiers
  - superlatives
  - comparatives
  - questions
  - downward entailing expressions (e.g. *hardly*)
  - other (e.g. *only*)

NPI Classification

- the distribution of every NPI can be different
- classification by means of their licenser (Zwarts 1997)

NPI	Negation		
	classical/ antimorphic	regular/ antiadditive	minimal/ downward entailing
weak	+	+	+
strong	+	+	-
superstrong	+	-	-

Research Question

- NPI classification by means of their licenser possible?
- use of an association measure: **Odds Ratio**
  - association strength between an NPI and all negative contexts
    - determines the 'negative polarity' of an item
  - association strength between an NPI and the three classes of negation
    - shows if there is statistical evidence for Zwarts' (1997) theory

Research Hypothesis

- all NPIs:
  - occur more often than expected in negative contexts
- weak NPIs:
  - occur more often than expected at least in DE contexts and possibly also in AA and AM contexts
- strong NPIs:
  - occur more often than expected at least in AA contexts and possibly also in AM contexts
  - occur less often than expected in DE contexts
- superstrong NPIs:
  - occur more often than expected in AM contexts
  - occur less often than expected in AA and DE contexts

Data

- data set by Lichte & Soehn (2007)
  - 5.8 million sentences from the TüPP-D/Z corpus
  - lemmatized, annotated for clause structure
  - annotated for negative contexts:
    - PTKNEG } antimorphic contexts
    - AM }
    - AA → anti-additive contexts
    - DE } downward entailing contexts
    - DEINT }
  - → not all possible negative contexts are identified

Data

- select three NPIs:
  - one that is supposed to be **weak**: *alle Tassen im Schrank haben* (have all cups in the cupboard) – to have a screw loose
  - one that is supposed to be **strong**: *sonderlich* – particularly
  - one that is supposed to be **superstrong**: *jedermanns Sache* (everyone‘s thing) – everyone‘s cup of tea

Odds Ratios

- association measure for categorical data
- uses a 2 x 2 contingency table
- present the odds of an outcome in the presence of some other variable

Odds Ratio

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

	NPI	~ NPI	total
negative context	n_{11}	n_{12}	n_{1+}
~ negative context	n_{21}	n_{22}	n_{2+}
total	n_{+1}	n_{+2}	n

number of clauses

Odds Ratio

- the odds ratio is a nonnegative number
- $\hat{\theta} = 1$ → the variables are independent
- $\hat{\theta} > 1$ → the odds in row 1 are higher
  - the bigger the number, the stronger the association
- $\hat{\theta} < 1$ → the odds in row 2 are higher
  - the smaller the number, the stronger the association

Log Odds Ratio

- the sampling distribution of odds ratio is skewed for small to moderate sample sizes
- use of **Log Odds Ratio**
  - the natural logarithm of $\hat{\theta}$: $\log(\hat{\theta})$
  - with log odds ratio, independence of the variables corresponds to $\log(\hat{\theta}) = 0$

Log Odds Ratio

- with log odds ratio, we can calculate the standard error and confidence intervals
- $SE(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$
- confidence intervals: $\log \hat{\theta} \pm z_{a/2} \times SE(\log \hat{\theta})$
  - $z_{a/2}$ defines the confidence limits
  - for a 95% confidence interval, $z_{a/2} = 1.96$
  - confidence intervals for odds ratio can be calculated by exponentiating those of log odds ratio

Tassen im Schrank

	Tassen im Schrank	~ Tassen im Schrank	total
negative contexts	26	1,423,766	1,423,792
~ negative contexts	2	8,076,905	8,076,907
total	28	9,500,671	9,500,699

- $\hat{\theta} = 73.75$; $\log \hat{\theta} = 4.3$
- 95% confidence interval for $\hat{\theta}$: (17.5, 310.7)
- 95% confidence interval for $\log \hat{\theta}$: (2.9, 5.7)
- the odds for *Tassen im Schrank* to occur in a negative context are 74 times higher than in a non-negative context
- strongly associated with negative polarity

Tassen im Schrank

	frequencies (n_{11})	odds ratio + confidence interval		log odds ratio + confidence interval	
anti-morphic	10	**7.09**	3.3, 15.4	1.96	1.2, 2.7
anti-additive	0	**0**			
downward entailing	16	**25.92**	12.3, 54.8	3.26	2.5, 4

- *Tassen im Schrank* is 26 times more likely in a 'weak' context and 7 times more likely in a 'superstrong' context than in other contexts
- but: for a classification, n should not be the number of *all* clauses, but that of all *negative* clauses, right?

Tassen im Schrank

	frequencies (n_{11})	odds ratio + confidence interval		log odds ratio + confidence interval	
anti-morphic	10	**7.09** **0.59**	3.3, 15.4 0.3, 1.3	1.96 – 0.53	1.2, 2.7 – 1.3, 0.25
anti-additive	0	**0**			
downward entailing	16	**25.92** **2.75**	12.3, 54.8 1.3, 5.8	3.26 1.01	2.5, 4 0.3, 1.8

- *Tassen im Schrank* is 3 times more likely to occur in a 'weak' context than in 'strong' and 'superstrong' ones
- can be classified as a **weak** NPI?

n = number of all clauses
n = number of negative clauses

Sonderlich

	sonderlich	~ sonderlich	total
negative contexts	879	1,422,913	1,423,792
~ negative contexts	102	8,076,805	8,076,907
total	981	9,499,718	9,500,699

- $\hat{\theta} = 48.92$; $\log \hat{\theta} = 3.89$
- 95% confidence interval for log $\hat{\theta}$: (3.7, 4.1)
- 95% confidence interval for $\hat{\theta}$: (39.9, 60.1)
- the odds for *sonderlich* to occur in a negative context are 49 times higher than in a non-negative context
- strongly associated with negative polarity

Sonderlich

	frequencies (n_{11})	odds ratio + confidence interval		log odds ratio + confidence interval	
anti-morphic	781	**49.9** 4.15	42.7, 58.3 3.6, 4.8	3.91 1.42	3.8, 4.1 1.3, 1.6
anti-additive	94	**3.64** 0.46	2.9, 4.5 0.4, 0.6	1.29 – 0.78	1.1, 1.5 – 0.99, – 0.6
downward entailing	4	**0.08** 0.01	0.03, 0.2 0.004, 0.03	– 2.53 – 4.61	– 3.5, – 1.5 – 5.6, – 3.6

- *sonderlich* is 50 times more likely in a 'superstrong' and 4 times more likely in a 'strong' context than in other contexts
- it is 4 times more likely in a 'superstrong' context than in a 'strong' or 'weak' one
- **can** or **cannot** be classified as a **strong** NPI?

n = number of all clauses
n = number of negative clauses

Jedermanns Sache

	jedermanns Sache	~ jedermanns Sache	total
negative contexts	66	1,423,726	1,423,792
~ negative contexts	0	8,076,907	8,076,907
total	66	9,500,633	9,500,699

- $\hat{\theta} = 374.42$; $\log \hat{\theta} = 5.93$
- the odds for *jedermanns Sache* to occur in a negative context are 374 times higher than in a non-negative context
- strongly associated with negative polarity

Jedermanns Sache

	frequencies (n_{11})	odds ratio + confidence interval		log odds ratio + confidence interval	
anti-morphic	64	**408.5** 34	100, 1668.9 8.3, 138.9	6.01 3.53	4.6, 7.4 2.1, 4.9
anti-additive	0	**0**			
downward entailing	2	**0.61** 0.06	0.1, 2.5 0.01, 0.2	– 0.49 – 2.81	– 1.9, 0.9 – 4.2, – 1.4

- *jedermanns Sache* is 409 times more likely to occur in 'superstrong' contexts than in others
- it is 34 times more likely to occur in a 'superstrong' context than in a 'strong' or 'weak' one
- can be classified as a **superstrong** NPI?

n = number of all clauses
n = number of negative clauses

Discussion

- for n, is the number of all clauses or that of all negative clauses relevant (does not always result in the same classification)
- can the method really prove that Zwarts‘ theory is appropriate

References

- Alan Agresti (1996). *An Introduction to Categorical Data Analysis*. Wiley: New York.
- Timm Lichte and Jan-Philipp Soehn (2007): "The retrieval and classification of negative polarity items using statistical profiles." In: Sam Featherston and Wolfgang Sternefeld (eds.). *Roots: Linguistics in Search of its Evidential Base*. Berlin: Mouton de Gruyter. pp. 249 – 266.

n = number of clauses