Overview

This workshop discusses different statistical procedures available for analysis of sociolinguistic data in large language corpora. I will demonstrate that the traditional approach of using aggregated data with the log-likelihood statistic is in principle unreliable. Instead, the workshop will offer suggestions for alternative methodologies and statistical procedures, which take into account within group differences and therefore produce more meaningful results. As part of the workshop, a new research tool **BNC64 Search & Compare** will be introduced. **BNC64 Search & Compare** can carry out detailed analyses based on a socially-balanced spoken corpus BNC64 (1.5 million words). BNC64 represents the speech of 64 speakers - 32 men and 32 women - extracted from the British National Corpus (BNC). BNC64 Search and Compare is a web-based environment that creates simple visualisations, calculates statistics and produces concordances. The website was created to allow for easy visualisations of complex corpus data and easy testing of a number of different sociolinguistic hypotheses. The workshop will be structured around a series of practical exercises guiding the participants through different types of analysis of corpus data and statistical procedures. The following areas will be covered:

- Sociolinguistic data in language corpora
- Descriptive and inferential statistics
- Individual and social variation
- The null-hypothesis testing paradigm and the "new" statistics

Workshop convenor: Vaclav Brezina (v.brezina@lancaster.ac.uk), ESRC Centre for Corpus Approaches to Social Science, Lancaster University

Exercises

***Task 1:* Find the best model for the data.** If you were to choose ONE number to describe each of the following datasets, what would this number be (It doesn't have to be a number from the dataset)?

Table 1. Three data sets

Data set	Model
Data set 1: 10, 10, 10, 10, 10, 10, 10, 10, 10, 10	
Data set 2: 10, 11, 8, 9, 10, 11, 12, 9, 10, 10	
Data set 3: 10, 11, 8, 9, 100, 11, 12, 9, 10, 10	

***Task 2:* Use the three data sets from Task 1 and calculate the mean & SD, trimmed mean & Winsorized SD.** Use the online statistics tool to help you with the calculations. http://corpora.lancs.ac.uk/bnc64/workshop.php

Compare your models from Task 1 with the results obtained in Task 2.

Table 2. Measures of central tendency

Dataset	Mean	SD	20% trimmed mean	Winsorized SD
Dataset 1				
Dataset 2				
Dataset 3				

***Task 3:* Compare the use of linguistic variable X in a contingency table (traditional approach)**

A) Look at Table 3. It provides data about the frequency of variable X in the Female and the Male subcorpus: *What can you say about the use of variable X in the speech of male and female speakers?*

Table 3. Contingency table based on aggregate data

	Female subcorpus	**Male subcorpus**
Occurrences of variable X	140	100
Corpus size (tokens)	5,000	5,000

Based on the contingency table (Table 3) we can conclude that_____

_____

B) Calculate the **log-likelihood score** and the **p-value** using the *UCREL LL calculator* (http://ucrel.lancs.ac.uk/llwizard.html). Log likelihood score is: _____

The p-value is:_____

Do the results of the statistical analysis confirm your initial observation?

_____

***Task 4:* Looking inside corpora…**

Look at Table 4. It provides data about the distribution of variable X among Female and Male speakers in the corpus. What can you say about the use of variable X in the speech of male and female speakers?

Table 4. Distribution of linguistic variable X in the speech of individual speakers

Individual speakers	Freq. of ling. variable X	Sample size
F1	12	1,000
F2	10	1,000
F3	100	1,000
F4	10	1,000
F5	8	1,000
M1	22	1,000
M2	20	1,000
M3	20	1,000
M4	20	1,000
M5	18	1,000

Based on the distribution table (Table 4) we can conclude that_____

_____

Task 5: Appropriate generalising: Measures of central tendency (mean, 20% trimmed mean), Robust mean difference & Robust Cohen's *d.*

a) Use the data from Table 4. Calculate the *mean* for the female and for the male group.
b) Calculate *20% trimmed mean* by removing the lowest and the highest 20% of the values from the data set and by calculating the arithmetic average of the remaining values.
c) Calculate the difference between the female and male group using *Robust mean difference* and *Robust Cohen's d.*

Table 5. Measures of central tendency

	Female group	Male group	Difference between the female and male group
Mean			
20 % trimmed mean			
Robust mean difference, incl. 95% CI			
Robust Cohen's *d*			

Task 6: *Do men swear more than women?* Use the **BNC64 Search & Compare** to test different socio-linguistic hypotheses about swearing. Pay attention to the different statistical measures and their interpretation.

Table 6. Swearing and gender: BNC64

Swear word	Statistically significant result?	Meaningful difference?
1.		
2.		
3.		
4.		

Task 7: Find typically male and typically female linguistic features in BNC 64. Use the **BNC64 Search & Compare** to test different socio-linguistic hypotheses.