Corpus Linguistics: Analysing word frequencies
INTRODUCTION
A corpus of British English and American English
- Books from 19th century British and American writers downloaded from the Gutenberg Project

- Number of individual words: 30,723

- Number of occurrences in AE corpus: 11,709,009

- Number of occurrences in BE corpus: 13,795,791

- Total size of corpus: 25,504,800
Is the word “colour” used more often in American or British English?

<table>
<thead>
<tr>
<th>Occ. of “colour”</th>
<th>Total number of words</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>255</td>
<td>11 709 009</td>
</tr>
<tr>
<td>BE</td>
<td>1772</td>
<td>13 795 791</td>
</tr>
</tbody>
</table>
Is it significant?
\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]
CHI SQUARE

\[ \chi^2 = 906.71 \]

\[ p < .001 \]
Is the word “the” used more often in American or British English?

<table>
<thead>
<tr>
<th></th>
<th>Occ. of “the”</th>
<th>Total number of words</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>848 729</td>
<td>11 709 009</td>
<td>72 485 pwm</td>
</tr>
<tr>
<td>BE</td>
<td>914 669</td>
<td>13 795 791</td>
<td>66 300 pmw</td>
</tr>
</tbody>
</table>
WORD FREQUENCIES

CHI SQUARE

\[ \chi^2 = 3503.73 \]

\[ p < .001 \]
But...

Is $\chi^2$ really appropriate?
Normal distribution: 95% of the values lie within two standard deviations
Word frequencies: zeta distribution

ZIPF’S LAW

A small number of words have a very high frequency, and a large number of words have a very low frequency.
NULL HYPOTHESIS

The null hypothesis is the idea that there is no relationship between two measured phenomena.
IN OTHER WORDS...

The null hypothesis is the hypothesis that chance alone can explain what we’re observing.
LANGUAGE IS NOT RANDOM
“Words are not selected at random. There is no a priori reason to expect them to behave as if they had been, and indeed they do not.”

Adam Kilgarriff, 1996
Chi-square: How bad is it?
USING CHI-SQUARE ON THE WHOLE CORPUS:

❖ Only keep words with over 5 occurrences
❖ Only keep words that occur in both AE and BE corpus
USING CHI-SQUARE ON THE WHOLE CORPUS:

- Only keep words with over 5 occurrences
- Only keep words that occur in both AE and BE corpus

- Number of significant results: \textbf{15197}
- Number of non-significant results: \textbf{15526}
- \textbf{49.4\%} of tests turn out significant (p < .05)
ALTERNATIVE #1:
Cramer’s V
Cramer’s Phi & Cramer’s V

\[ \Phi = \sqrt{\frac{\chi^2}{N}} \]

\[ V = \sqrt{\frac{\Phi}{k-1}} \]
USING CRAMER’S PHI ON THE WHOLE CORPUS:

❖ Same data as for Chi-square test
❖ Maximum value of Phi coefficient is determined by the distribution of the two variables
USING CRAMER’S PHI ON THE WHOLE CORPUS:

- Same data as for Chi-square test
- Maximum value of Phi coefficient is determined by the distribution of the two variables

\[ \Phi > 0.5 : \]
- Significant results: 1198
- Non-significant results: 29525
- 4.06% of tests turn out significant
USING CRAMER’S PHI ON THE WHOLE CORPUS:

❖ Same data as for Chi-square test
❖ Maximum value of Phi coefficient is determined by the distribution of the two variables

Φ > 0.5 :
❖ Significant results: 1198
❖ Non-significant results: 29525
❖ 4.06% of tests turn out significant

Φ > 0.6 :
❖ Significant results: 698
❖ Non-significant results: 30025
❖ 2.32% of tests turn out significant
USING CRAMER'S PHI ON THE WHOLE CORPUS:

- Phi coefficient for “colour”: 0.447
- Phi coefficient for “the”: 0.001
ALTERNATIVE #2:
Wilcoxon-Mann-Whitney ranking test
Wilcoxon-Mann-Whitney ranking test
WMW

- Uses frequency to rank items and determine the value of the statistic (U)
- Divide the data in equal sized samples
- For each observation, retain frequency and origin of the sample (AE or BE)

<table>
<thead>
<tr>
<th>“raining”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 4 4 5 7</td>
</tr>
<tr>
<td>AE AE BE AE BE BE</td>
</tr>
<tr>
<td>1 2 3 4 5 6</td>
</tr>
</tbody>
</table>
The significance of the U statistic can be checked using normal distribution tables.

AE and BE were divided in 10 equal sized chunks.
Tests made on all words with a frequency over 30 (n = 15756)
**WORD FREQUENCIES**

**WMW**

\[ p < 0.05: \]

- Significant results: \(2357\)
- Non-significant results: \(13399\)
- \(17.59\%\) of tests turn out significant
WORD FREQUENCIES

WMW

p < 0.05:
- Significant results: 2357
- Non-significant results: 13399
- 17.59% of tests turn out significant

p < 0.01:
- Significant results: 889
- Non-significant results: 14867
- 5.98% of tests turn out significant
Let’s see some results...
CONCLUSION
Choosing the most appropriate test
<table>
<thead>
<tr>
<th><strong>CRAMER’s V</strong></th>
<th><strong>WILCOXON-MANN-WHITNEY</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>No local copy of corpus needed</td>
<td>Local copy of corpus needed</td>
</tr>
<tr>
<td>No programming skills required</td>
<td>Some programming skills required</td>
</tr>
<tr>
<td>Interpretation can be difficult</td>
<td>Interpretation is easy</td>
</tr>
</tbody>
</table>
THANKS!

Any questions?