Theory

Consider testing for whether a phrase like “new york” is a collocation. The occurrence counts are $C(\text{new}) = 794$, $C(\text{york}) = 149$, $C(\text{new, york}) = 124$, and $N = 477813$. In a two-by-two table, the data are

<table>
<thead>
<tr>
<th></th>
<th>york</th>
<th>¬york</th>
</tr>
</thead>
<tbody>
<tr>
<td>new</td>
<td>124</td>
<td>670</td>
</tr>
<tr>
<td>¬new</td>
<td>25</td>
<td>476994</td>
</tr>
</tbody>
</table>

In class, we developed a test of the null hypothesis of $H_0$ (no collocation) versus $H_1$ (collocation) where the hypotheses are

$H_0 : \text{Pr}(\text{york} | \text{new}) = \text{Pr}(\text{york} | ¬\text{new})$,  
$H_1 : \text{Pr}(\text{york} | \text{new}) > \text{Pr}(\text{york} | ¬\text{new})$.

To perform the test, we conditioned on the row sums in the two-by-two table, so that we could treat $C(\text{new, york})$ and $C(¬\text{new, york})$ like independent binomial random variables. We then used a likelihood ratio test.

In your homework assignment, you will consider one of the following two alternative tests. Choose either Option 1 or Option 2 on one of the subsequent pages.

Application

Download the anc-masc.json corpus from the course webpage. Use the test you develop in Option 1 or Option 2 to test for collocations in the corpus. Print out the chi squared statistics and $p$-values for the top 30 collocations. You can use segment.Rmd as a starting point.
Option 1

Perform a test conditional on the second word, not the first word. Specifically, define

\[ p_1 = \Pr(\text{first word is “new” | second word is “york”}) \]
\[ p_2 = \Pr(\text{first word is “new” | second word is not “york”}) \]

Suppose you have seen \( n_1 \) occurrences of “york”, and \( n_2 \) occurrences of “¬york”. Let

\[ X_1 = \#(\text{occurrences of “new” followed by “york”}), \]
\[ X_2 = \#(\text{occurrences of “new” followed by “¬york”}). \]

1. Argue that \( X_1 \) and \( X_2 \) can be approximated as independent binomial random variables.

2. Find expressions for the observed values \( n_1, n_2, x_1, \) and \( x_2 \) in terms of \( C(\text{new}), C(\text{york}), C(\text{new, york}), \) and \( N. \)

3. Give an expression for the log-likelihood function

\[ l(p_1, p_2) = \log P(X_1 = x_1, X_2 = x_2 | n_1, n_2, p_1, p_2). \]

4. Write down the appropriate null and alternative hypothesis for testing for a collocation, in terms of \( p_1 \) and \( p_2. \)

5. Derive an expression for \( \hat{l}_0 = \sup_{p_1} l(p_1, p_2). \)

6. Derive an expression for \( \hat{l}_1 = \sup_{p_1} l(p_1, p_2). \)

7. Under the null hypothesis, what is the distribution of the likelihood ratio statistic \( \chi^2 = -2(\hat{l}_0 - \hat{l}_1)? \)
Option 2

Perform a test conditional on the total. Let \( Y_1, \ldots, Y_N \) be the consecutive bigrams in the corpus. For \( 1 \leq k \leq N \), define

\[
\begin{align*}
    p_{11} &= \Pr(Y_k = (\text{new, york})) \\
    p_{12} &= \Pr(Y_k = (\text{new, } \neg \text{york})) \\
    p_{21} &= \Pr(Y_k = (\neg \text{new, york})) \\
    p_{22} &= \Pr(Y_k = (\neg \text{new, } \neg \text{york}))
\end{align*}
\]

Note that \( p_{11} + p_{12} + p_{21} + p_{22} = 1 \). Also, define

\[
\begin{align*}
    X_{11} &= C(\text{new, york}) \\
    X_{12} &= C(\text{new, } \neg \text{york}) \\
    X_{21} &= C(\neg \text{new, york}) \\
    X_{22} &= C(\neg \text{new, } \neg \text{york})
\end{align*}
\]

Note that \( X_{11} + X_{12} + X_{21} + X_{22} = N \).

1. Assume that \( Y_1, \ldots, Y_N \) are independent. Do you think this is reasonable? Why or why not?

2. Under the independence assumption, argue that \( X = (X_{11}, X_{12}, X_{21}, X_{22}) \) is a multinomial random variable.

3. Write the log-likelihood function

\[
l(p) = \log \Pr(X = x \mid N, p),
\]

where \( p = (p_{11}, p_{12}, p_{21}, p_{22}) \), and \( x = (x_{11}, x_{12}, x_{21}, x_{22}) \).

4. In terms of \( p \), write the null and alternative hypotheses, corresponding to “new york is a collocation” and “new york is not a collocation,” respectively.

5. Derive an expression for \( \hat{l}_0 = \sup_{H_0} l(p) \).

6. Derive an expression for \( \hat{l}_1 = \sup_{H_1} l(p) \).

7. Under the null hypothesis, what is the distribution of the likelihood ratio statistic \( \chi^2 = -2(\hat{l}_0 - \hat{l}_1) \)?