Homework 2

STAT-GB.4310: Statistics for Social Data
Instructor: Patrick O. Perry

Due February 16, 2016

Theory

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

	york	¬york
new	124	670
¬new	25	476994

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

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

To perform the test, we conditioned on the row sums in the two-by-two table, so that we could treat $C(\text{new}, \text{york})$ and $C(\neg\text{new}, \text{york})$ 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

$$p_1 = \Pr(\text{first word is "new"} \mid \text{second word is "york"})$$
$$p_2 = \Pr(\text{first word is "new"} \mid \text{second word is not "york"})$$

Suppose you have seen n_1 occurences of "york", and n_2 occurrences of "$\neg$york". Let

$$X_1 = \#\{\text{occurrences of "york" preceded by "new"}\},$$
$$X_2 = \#\{\text{occurrences of "york" preceded by "}\neg\text{new"}\}.$$

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 $C(\text{new})$, $C(\text{york})$, $C(\text{new}, \text{york})$, and N.
3. Give an expression for the log-likelihood function
$$l(p_1, p_2) = \log P(X_1 = x_1, X_2 = x_2 \mid n_1, n_2, p_1, p_2).$$
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 $\hat{\ell}_0 = \sup_{H_0} l(p_1, p_2)$.
6. Derive an expression for $\hat{\ell}_1 = \sup_{H_1} l(p_1, p_2)$.
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 $1 \leqslant k \leqslant N$, define

$$\begin{aligned} p_{11} &= \Pr\{Y_k = (\text{new}, \text{york})\} \\ p_{12} &= \Pr\{Y_k = (\text{new}, \neg\text{york})\} \\ p_{21} &= \Pr\{Y_k = (\neg\text{new}, \text{york})\} \\ p_{22} &= \Pr\{Y_k = (\neg\text{new}, \neg\text{york})\} \end{aligned}$$

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

$$\begin{aligned} X_{11} &= C(\text{new}, \text{york}) \\ X_{12} &= C(\text{new}, \neg\text{york}) \\ X_{21} &= C(\neg\text{new}, \text{york}) \\ X_{22} &= C(\neg\text{new}, \neg\text{york}) \end{aligned}$$

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-likelilhood 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 $\hat{\ell}_0 = \sup_{H_0} l(p)$.

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

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