Hypothesis Testing I: χ^2 for collocations

Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul

OCTOBER 4, 2016

Distributional Independence

- If x and y are independent, $P(x,y) = P(x)P(y)$.
- Can we test of two distributions are independent?
- This also is a χ^2 test

Example: Collocations

- Selectional preferences: “strong tea”, not “powerful tea”
- Phrases: “intents and purposes”, “helter skelter”
- Some words just go together more than others
- I.e., they’re not independent

Can't use frequency

Most frequent bigrams are just the most frequent words. (Independent distribution.)

80871	of the
58841	in the
26430	to the
21842	on the
21839	for the
18568	and the
16121	that the
15630	at the
15494	to be
13899	in a
13689	of a
13361	by the

Contingency tables

	$w_1 = \text{new}$	$w_1 \neq \text{new}$
$w_2 = \text{companies}$	8 (new companies)	4667 (e.g., old companies)
$w_2 \neq \text{companies}$	15820 (e.g., new machines)	14287181 (e.g., old machines)

Contingency tables: degrees of freedom

- Given row and column totals, one cell can fill in the rest (as you did in first quiz)
- In general, for a contingency table with r rows and c columns, $(r-1)(c-1)$ degrees of freedom

Observed

	$w_1 =$ **new**	$w_1 \neq$ **new**
$w_2 =$ **companies**	8	4667
$w_2 \neq$ **companies**	15820	14287181

Observed

	$w_1 =$ **new**	$w_1 \neq$ **new**	
$w_2 =$ **companies**	8	4667	*4675*
$w_2 \neq$ **companies**	15820	14287181	*14303001*
	15828	*14291848*	*14307676*

Expected

	$w_1 =$ **new**	$w_1 \neq$ **new**
$w_2 =$ **companies**	$\frac{15828}{14307676} \frac{4675}{14307676} \cdot 14307676 = 5.17$	1669.83
$w_2 \neq$ **companies**	15822.83	14287178.17

Observed

	$w_1 = \textbf{new}$	$w_1 \neq \textbf{new}$
$w_2 = \textbf{companies}$	8	4667
$w_2 \neq \textbf{companies}$	15820	14287181

Expected

	$w_1 = \textbf{new}$	$w_1 \neq \textbf{new}$
$w_2 = \textbf{companies}$	5.17	1669.83
$w_2 \neq \textbf{companies}$	15822.83	14287178.17

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \quad (1)$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \quad (2)$$

Observed

	$w_1 =$ **new**	$w_1 \neq$ **new**
$w_2 =$ **companies**	8	4667
$w_2 \neq$ **companies**	15820	14287181

Expected

	$w_1 =$ **new**	$w_1 \neq$ **new**
$w_2 =$ **companies**	5.17	1669.83
$w_2 \neq$ **companies**	15822.83	14287178.17

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{1}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{2}$$

Observed

	$w_1 = $ **new**	$w_1 \neq$ **new**
$w_2 = $ **companies**	8	4667
$w_2 \neq$ **companies**	15820	14287181

Expected

	$w_1 = $ **new**	$w_1 \neq$ **new**
$w_2 = $ **companies**	5.17	1669.83
$w_2 \neq$ **companies**	15822.83	14287178.17

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{1}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{2}$$

Observed

	$w_1 =$ **new**	$w_1 \neq$ **new**
$w_2 =$ **companies**	8	4667
$w_2 \neq$ **companies**	15820	14287181

Expected

	$w_1 =$ **new**	$w_1 \neq$ **new**
$w_2 =$ **companies**	5.17	1669.83
$w_2 \neq$ **companies**	15822.83	14287178.17

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{1}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{2}$$

Observed

	$w_1 =$ **new**	$w_1 \neq$ **new**
$w_2 =$ **companies**	8	4667
$w_2 \neq$ **companies**	15820	14287181

Expected

	$w_1 =$ **new**	$w_1 \neq$ **new**
$w_2 =$ **companies**	5.17	1669.83
$w_2 \neq$ **companies**	15822.83	14287178.17

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{1}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{2}$$

Observed

	$w_1 = \textbf{new}$	$w_1 \neq \textbf{new}$
$w_2 = \textbf{companies}$	8	4667
$w_2 \neq \textbf{companies}$	15820	14287181

Expected

	$w_1 = \textbf{new}$	$w_1 \neq \textbf{new}$
$w_2 = \textbf{companies}$	5.17	1669.83
$w_2 \neq \textbf{companies}$	15822.83	14287178.17

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{1}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} = 1.55 \tag{2}$$

Can we reject the null?