Common statistics used in corpus linguistics

Laurence Anthony

September 27, 2023

This paper describes some of the common statistics used in corpus linguistics research. The paper focuses on statistics that are used in the *AntConc* corpus analysis toolkit (Anthony 2023), but the statistics are known to be used in a wide number of other online and offline corpus analysis tools, as well in many research papers. The equations shown here are largely based on the work of Andrew Hardie of Lancaster University as presented in an unpublished, internal working paper of 2014. The equations from the original paper are reproduced here with his kind permission, but all responsibility for the accuracy of the work presented here lies with the current author. The notation used in the Hardie paper of 2014 and here is taken from the work of Evert (2004: 36-37).

Foundations - Equivalence of Keyness and Collocation Measures

Keyness measures are used to identify and rank the degree to which words in a target corpus appear unusually frequently compared with their occurrence in a reference corpus. To calculate this property, contingency tables of observed and expected values are used, where "O" represents the observed value, "E" represents the expected value, C represents the column total, R represents the row total, and N represents the total number of words in target and reference corpora combined.

	freq. (target word)	**freq. (all other words)**	**Totals**
Target Corpus	O_{11}	O_{12}	**R_1**
Reference Corpus	O_{21}	O_{22}	**R_2**
Totals	**C_1**	**C_2**	**N**

Table 1. Contingency table for observed values as used in a keyness measure.

	freq. (target word)	**freq. (all other words)**	**Totals**
Target Corpus	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$	**R_1**
Reference Corpus	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$	**R_2**
Totals	**C_1**	**C_2**	**N**

Table 2. Contingency table for expected values as used in a keyness measure.

Collocation measures are analogous to keyness measures in that they are used to identify and rank the degree to which words in a corpus appear unusually frequently in a context span surrounding a target word (i.e., collocate with the target word) compared with their occurrence in the corpus as a

whole. If we conceive the context span around a target word as analogous to a 'target corpus' (i.e., the words that are the target of our analysis) and the words outside of this span as analogous to a reference corpus (i.e., the words that we are using as a reference to determine if a word collocates with a target word), then it is clear that keyness measures and collocation measures are mathematically equivalent and the same contingency tables can be used.

	freq. (candidate collocate)	**freq. (all other words)**	**Totals**
Target 'Corpus' (set of context spans around a target word)	O_{11}	O_{12}	**R_1**
Reference 'Corpus' (all words not in the context spans of a target word)	O_{21}	O_{22}	**R_2**
Totals	**C_1**	**C_2**	**N**

Table 3. Contingency table for observed values as used in a collocation measure.

	freq. (candidate collocate)	**freq. (all other words)**	**Totals**
Target 'Corpus' (set of context spans around a target word)	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$	**R_1**
Reference 'Corpus' (all words not in the context spans of a target word)	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$	**R_2**
Totals	**C_1**	**C_2**	**N**

Table 4. Contingency table for expected values as used in a collocation measure.

Using the values in the cells of the contingency tables as presented in Tables 1-4, various keyness measures and collocation measures can be calculated. Traditionally, some of these measures have been more commonly used to determine keyness and others more commonly used to determine collocation strength, but they are presented here without any such distinction.

Dice coefficient
(Dice 1945)

$$Dice\ coefficient = \frac{2O_{11}}{R_1 + C_1}$$

LogDice
(Pavel Rychlý 2008)

$$LogDice = 14 + \log_2\left(\frac{2O_{11}}{R_1 + C_1}\right)$$

Log Ratio
Hardie (2014)

$$Log\ Ratio = \log_2\left(\frac{R_2 O_{11}}{R_1 O_{21}}\right)$$

Hardie (2014) recommends that for cases when O_{11} or O_{21} =is zero, add 0.5 to the value.

Mutual Information (MI)

$$MI = \log_2\left(\frac{O_{11}}{E_{11}}\right)$$

Mutual Information^2 (MI^2)

$$MI^2 = \log_2\left(\frac{(O_{11})^2}{E_{11}}\right)$$

Mutual Information^3 (MI^3)

$$MI^3 = \log_2\left(\frac{(O_{11})^3}{E_{11}}\right)$$

Minimum sensitivity coefficient
(Pedersen and Bruce: 1996:12)

$$MS = min\left\{\frac{O_{11}}{R_1}, \frac{O_{11}}{C_1}\right\}$$

Ratio of observed:expected (Evert's mu)
Evert (2004: 54)

$$Mu\ value = \frac{O_{11}}{E_{11}}$$

Ratio of relative frequencies (RRF)

$RRF = \frac{O_{11}/R_1}{O_{21}/R_2}$	$RRF = \frac{R_2 O_{11}}{R_1 O_{21}}$
(standard form)	(simplified form)

Difference of relative frequencies (DRF)

$$DRF = O_{11}/R_1 - O_{21}/R_2$$

T-score

$$T - score = \frac{O_{11} - E_{11}}{\sqrt{O_{12}}}$$

Z-Score

$$z = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$

Hofland and Johansson's Difference Coefficient
(Hofland and Johansson, 1982: 14, 471-544)

$$H\&J's\ Difference\ Coefficient = \frac{O_{11} - O_{21}}{O_{11} + O_{21}}$$

$Difference\ Coefficient\ (Relative) = \frac{O_{11}/R_1 - O_{21}/R_2}{O_{11}/R_1 + O_{21}/R_2}$	$Difference\ Coefficient\ (Relative) = \frac{R_2 O_{11} - R_1 O_{21}}{R_2 O_{11} + R_1 O_{21}}$
(standard form)	(simplified form)

Chi-squared (χ^2)

Standard form

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Expanded form

$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}}$$

Chi-squared (χ^2) with Yates Correction

Standard form

$$\chi^2\ (Yates)\ = \sum_i \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Expanded form

$$\chi^2\ (Yates) = \frac{(|O_{11} - E_{11}| - 0.5)^2}{E_{11}} + \frac{(|O_{12} - E_{12}| - 0.5)^2}{E_{12}} + \frac{(|O_{21} - E_{21}| - 0.5)^2}{E_{21}} + \frac{(|O_{22} - E_{22}| - 0.5)^2}{E_{22}}$$

Log Likelihood (G^2)

Standard form

$$Log\ Likelihood = 2\ \sum_{ij} O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

Expanded form

$$Log\ Likelihood = 2\left(O_{11} \ln\left(\frac{O_{11}}{E_{11}}\right) + O_{21} \ln\left(\frac{O_{21}}{E_{21}}\right) + O_{12} \ln\left(\frac{O_{12}}{E_{12}}\right) + O_{22} \ln\left(\frac{O_{22}}{E_{22}}\right)\right)$$

Expanded form (2 -term)

$$Log\ Likelihood = 2\left(O_{11} \ln\left(\frac{O_{11}}{E_{11}}\right) + O_{21} \ln\left(\frac{O_{21}}{E_{21}}\right)\right)$$

The exact same formula for Log Likelihood (G^2) are also used in the Text Dispersion statistic with frequency values replaced with range values.

The following statistics have been described in the literature, but they do not appear in AntConc.

Yule's Q
(Yule, 1944)

$$Association\ Coefficient: Yule's\ Q = \frac{O_{11}O_{22} - O_{12}O_{21}}{O_{11}O_{22} + O_{12}O_{21}}$$

%DIFF
(Gabrielatos and Marchi (2012)

$\%DIFF = \frac{O_{11}/R_1 - O_{21}/R_2}{O_{21}/R_2} \times 100$	$\%DIFF = \frac{R_2 O_{11}/R_1 - O_{21}}{O_{21}} \times 100$
(standard form)	(simplified form)

Odds-ratio

$OR = \frac{O_{11}/O_{12}}{O_{21}/O_{22}}$	$OR = \frac{O_{11}O_{22}}{O_{12}O_{21}}$
(standard form)	(simplified form)

References:

Evert, S. 2004. *The Statistics of Word Cooccurrences*: *Word Pairs and Collocations*. Unpublished Ph.D. thesis. University of Stuttgart. (Published 2005; available online at http://elib.unistuttgart.de/opus/volltexte/2005/2371/.)

Gabrielatos, C. and Marchi, A. 2012. "Keyness: Appropriate metrics and practical issues". Presentation at CADS 2012 conference. Available online at http://repository.edgehill.ac.uk/4196/1/Gabrielatos%26Marchi-Keyness-CADS2012.pdf

Hofland, K. and Johansson, S. 1982. *Word frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Pedersen, T. and Bruce, R. 1996. *What to infer from a description*. Technical Report 96-CSE-04, Southern Methodist University, Dallas, TX.

Rychlý, P. (2008, December). A Lexicographer-Friendly Association Score. In *RASLAN* (pp. 6-9).

Yule, G.U. 1944. *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press. (Reprinted 1968 by Archon Press.)

Hardie, A. (2014). Unpublished internal working paper.