ICAME 34: *Modelling patterns of linguistic interaction*

Sean Wallis, Survey of English Usage, UCL

Readers are referred to Wallis (to appear 2013) for more information on the methods summarised briefly below. See also my **corp.ling.stats** blog (http://corplingstats.wordpress.com) for discussion, worked examples, links to papers, and spreadsheets for carrying out calculations. A pre-publication version of a paper describing the work under discussion is published online as Wallis (2012).

1. Analysing sequential probabilities

In this type of experiment, we wish to study the probability of making a decision at different points in the process of constructing a larger structure, such as an NP. At each point the speaker has a simple choice: **to add** a term or **not to add** a term. The null hypothesis is that the probability of adding the x-th term (for simplicity, $p(x)$) is constant.

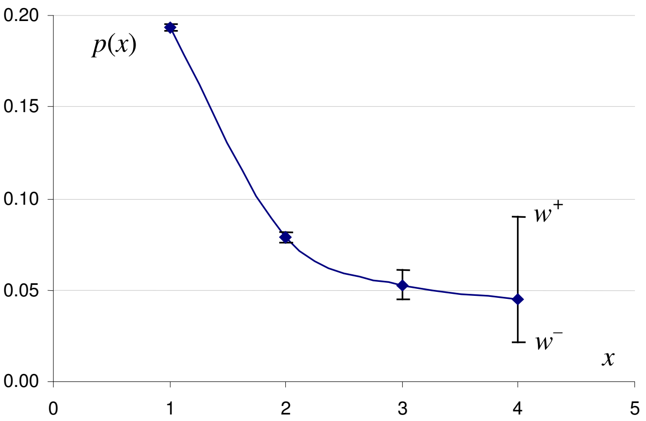

Figure 1: Probability distribution of adding attributive adjective phrases to an NP, all ICE-GB data, with Wilson confidence intervals (after Wallis 2012).

To determine the probability of adding the first term $p(1)$ to a construction of size 0, we simply divide the frequency of all cases of a construction *with at least one term* by the frequency of all cases. To put it another way, the probability of adding the first adjective to an NP is the proportion of cases of NPs with at least one adjective. By extension, for all x:

$$p(x) = f / n = F(x) / F(x\text{-}1).$$

The following data is taken from ICE-GB. We can lay the data out like this and plot the probability curve as in Figure 1.

***x* adjective phrases**	**0**	**1**	**2**	**3**	**4**
‘at least *x*’ *F*	193,123	37,306	2,944	155	7
probability *p*		0.1932	0.0789	0.0526	0.0452
upper bound w^+		0.1949	0.0817	0.0613	0.0903
lower bound w^-		0.1914	0.0762	0.0451	0.0220
Significance			s-	s-	ns

Each interval represents the likely range of potential values that $p(x)$ may take given the amount of data supporting the observation, at a particular error level α.

2. Confidence intervals on single observations

Confidence intervals on the true rate p are commonly computed using either Gaussian (Normal) or (less frequently) Wilson score interval (Poisson) methods. The first, most commonly used, method is wrong, however!

A confidence interval on an observation p represents the range that the true population value, P (which we cannot observe directly) may take at a given level (e.g. 95%).[1] To plot error bars around observed p supported by n observations ($p = f / n$, where $f = F(x)$ and $n = F(x\text{-}1)$), we should use Wilson’s score interval:

$$\textit{Wilson score interval}\ (w^-, w^+) \equiv \left(p + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n} + \frac{z^2_{\alpha/2}}{4n^2}}\right) \Bigg/ \left(1 + \frac{z^2_{\alpha/2}}{n}\right),$$

where $z_{\alpha/2}$ is ‘the two-tailed critical value of the standard Normal distribution at a given error level’. Although this is a bit of a mouthful, critical values of z are constant, and α is usually set at $^1/_{20}$ (0.05). $z(0.05) = 1.95996$ (to six decimal places).

The score interval is not symmetric but tends towards the middle of the distribution. It cannot exceed the probability range (0, 1) and is **strongly recommended** over the Gaussian, particularly for skewed data and small samples (common in corpus linguistics). In our case, both p and n can become very small, and the Wilson interval is used.

3. Testing falls/rises for significant difference (‘goodness of fit’)

Contingency correlation tests, including log-likelihood, χ^2, and its variations, are premised on the population z test (Wallis to appear 2013). The 2 × 1 goodness of fit χ^2 test is a reformulation of a single sample z test based on an expected baseline frequency. If we plot Wilson intervals, there is a short-cut involving testing against those intervals, which has exactly the same result.

[1] On **corp.ling.stats** I generally use capital letters to refer to population measures and lower case to refer to sample ones. In this paper, however, I follow an additional convention and use F to refer to cumulative (‘at least’) frequency. In the literature you may also see the symbols μ for a population mean and σ for a population standard deviation.

In other cases we might use this type of test to check whether a term (e.g. modal *shall*) correlates with a baseline (e.g. tensed VPs, or the set {*will*, *shall*}). In our experiment, we wish to compare $p(x)$ with the prior, $p(x\text{-}1)$ because our null hypothesis was that $p(x)$ was constant (flat) over all x.

The easiest way of testing for a significant difference is simply to examine the Wilson score intervals for $p(x)$ and check whether the previous probability, $p(x\text{-}1)$, falls outside the interval. If so, the difference is significant at the given error level α. Thus the significance test for $p(2)$ (two attributive adjective phrases) being significantly different from $p(1)$ (one AJP) is carried out by checking whether $p(1)$ is outside the range for $p(2)$, and so on:

***x* adjective phrases**	**0**	**1**	**2**	**3**	**4**
‘at least *x*’ *F*	193,123	37,306	2,944	155	7
probability *p*		**0.1932**	0.0789	0.0526	0.0452
upper bound w^+		0.1949	**0.0817**	0.0613	0.0903
lower bound w^-		0.1914	**0.0762**	0.0451	0.0220
significance			↑ s-	s-	ns

4. Testing parallel runs for significant difference (‘separability’)

Finally, Wallis (to appear 2013) also points out that it is possible to compare a pair of contingency tests for a ‘meta-test’ of *statistical separability*, i.e. to test if results are significantly different from each other. This can be used for comparing results from different corpora, or exploring permutations of an experimental design.

Crucially, whereas contingency tests assume that observations under comparison derive from the same data set, in this case the observed probabilities derive from **different** data sets – either from different subcorpora (spoken vs. written) or as the result of different processes (embedded vs. serial postmodification).

The method is discussed in detail in Wallis (2011), which gives formulae for comparing different experimental tables (2 × 1, 2 × 2, r × 1, r × c etc). In our case we performed 2 × 1 tests using the Wilson score interval (w^-, w^+) for each observation.

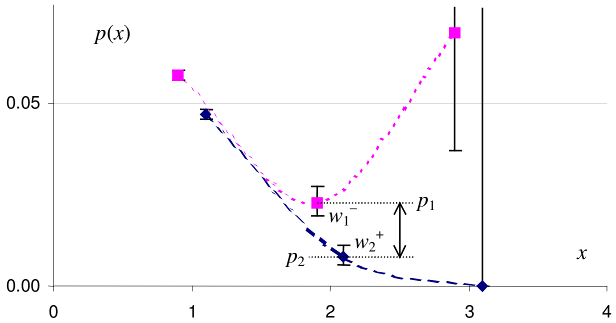

Figure 2: Comparing results for separability, sequential postmodifying clauses in ICE-GB (spoken vs. written).

Let us use the notation p_1 and p_2 for the two different observations, and w_1^-, w_1^+ etc., for their interval range. We can test the difference $D = p_1 - p_2$ against a new interval range ($-w_D^-$, w_D^+) derived using the formula:

$$w_D^- = \sqrt{(p_1 - w_1^-)^2 + (p_2 - w_2^+)^2} \ , w_D^+ = \sqrt{(p_1 - w_1^+)^2 + (p_2 - w_2^-)^2} \ .$$

All we need do now is compare D with this interval. If it is outside the range, the difference is significant:

$$-w_D^- < p_1 - p_2 < w_D^+.$$

Each term in the new interval is the square root of the sum of the squares of each Wilson interval width (e.g., $p_1 - w_1^-$) on the *near side* of the difference between $p_1 - p_2$. In other words, if p_1 is greater than p_2 (as in Figure 2), the near side width is based on the lower bound of p_1 and the upper bound of p_2. See Wallis (2011).

A quick way of reading graphs like this is as follows:

- If two intervals do not overlap (cf. Figure 2) the difference must be **significant**,
- If one interval contains the other’s central point p, the difference must be **non-significant**,
- Otherwise, if the interval ranges overlap, the difference *may* be significant and must be tested.

5. References

Nelson, G., Wallis, S. & Aarts, B. 2002. *Exploring natural language*. Benjamins.

Pickering, M. & Ferreira, V. 2008. Structural priming. *Psychological Bulletin* 134, 427–459.

Wallis, S.A. 2011. *Comparing χ^2 tests for separability*. Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/comparing-x2-tests.pdf

Wallis, S.A. 2012. *Capturing patterns of linguistic interaction in a parsed corpus: an insight into the empirical evaluation of grammar?* Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/analysing-grammatical-interaction.pdf

Wallis, S.A. to appear 2013. *z*-squared: the origin and use of χ^2. *Journal of Quantitative Linguistics* 20:4. www.ucl.ac.uk/english-usage/statspapers/z-squared.pdf

See also **corp.ling.stats** (http://corplingstats.wordpress.com).