Corpus Linguistics in the South V: *Statistics for variationists*

Sean Wallis, Survey of English Usage, UCL

Readers are referred to Wallis (to appear, a, b) for more information on 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.

1. 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!

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 ($p = f / n$), 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 it is **strongly recommended** over the Gaussian, particularly for skewed data and small samples (common in corpus linguistics).[2] It outperforms log-likelihood as well.

If we employ intervals to estimate proportions from large samples in finite populations we may obtain a more precise interval by first dividing N by $\nu = \sqrt{1 - N/N_p}$, where N_p is the size of the population (Singleton *et al*. 1988). Since $\nu < 1$, this boosts the effective size of N and shrinks the interval.

2. Contingency tests

Contingency correlation tests, including log-likelihood, χ^2, and its variations, are premised on the population z test (Wallis to appear, a). The 2 × 1 goodness of fit χ^2 test is a reformulation of a single sample z test based on an expected baseline frequency. We might use this to check whether a term (e.g. modal *shall*) correlates with a baseline (e.g. tensed VPs, or the set {*will*, *shall*}).

$$\textit{Population mean } x \equiv P = F/N, \textit{ standard deviation } S \equiv \sqrt{P(1-P)/N}\ .$$

The Gaussian interval about P is then $(P - E, P + E)$, where $E = z_{\alpha/2}.S$. To convert the Gaussian interval into a 2 × 1 goodness of fit χ^2 we simply test if $P - E < p < P + E$. In our case, $p = p(b \mid a)$, $P = p(b)$, $F = F(b)$ and $N = F(a)$. That is, the support for the interval about P within an observed subset a depends on the data in the observation.

Important: The Gaussian interval **should not** be used for computing intervals on observations of the sample, p. This is a common mistake that has perplexed many researchers working with skewed data: not least because it is easy to get intervals which cover an impossible negative probability range!

---

[1] We will use capital letters to refer to population measures and lower case to refer to sample ones. You may also see the symbols μ for a population mean and σ for a population standard deviation.

[2] Wallis (to appear, b) compares the performance of this method, and a continuity-corrected version of it, with an exact approach based on the Binomial distribution. If one were to be cautious, with small samples one should use the continuity-corrected formula, but overall, continuity correction trades coverage accuracy for caution.

The 2 × 2 χ^2 test is identical to a two-sample *z* test (see Wallis, to appear, a). Wallis (to appear, b) evaluates a range of methods for computing 2 × 2 tests, concluding:

1) where samples are drawn from the **same population** (if the same speakers appear in both samples), Yates' continuity-corrected χ^2 test should be used; and
2) where samples are drawn from **distinct populations** (typically, where samples are subdivided by speakers), Newcombe's (1998) method should be used, applying a similar 'continuity correction'.[3]

3. Confidence intervals on differences

The *z* test for two samples drawn from the same population is another way of calculating the 2 × 2 χ^2 test. This compares the difference, $p_1 - p_2$, with the following interval centred on zero. In comparing columns in a table, $p_1 = p(b \mid a)$, $p_2 = p(b \mid \neg a)$, $n_1 = F(a)$ and $n_2 = F(\neg a)$:

$$\textit{Gaussian difference interval: } \pm z_{\alpha/2} \sqrt{P(1-P)\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)}\,.$$

Newcombe employs the Wilson score interval to create a new interval for a 2×2 contingency test when comparing across sub-populations of data (e.g. corpus data plotted over time).

$$\textit{Newcombe-Wilson interval: } \left(-\sqrt{(p_1 - w_1^-)^2 + (w_2^+ - p_2)^2}\,,\sqrt{(w_1^+ - p_1)^2 + (p_2 - w_2^-)^2}\,\right),$$

where w_1^+ represents the upper bound of the Wilson score interval for p_1 etc. This formula combines the upper width of the Wilson interval at p_1 and the lower at p_2 to obtain an upper bound for $p_1 - p_2$.

4. Measures of association

To compare different results we can focus on these difference measures alone. Wallis (to appear, a) notes that simple swing, $d = p_2 - p_1$, and percentage swing, $d^{\%} = d/p_1$, are commonly used for comparing *p* values. These may be plotted with Newcombe-Wilson confidence intervals (note *d* is simply the negative of the difference used above).

More advanced methods include Cramér's ϕ, which can be extended to assess the size of an effect of an independent variable across multiple dependent values. Cramér's ϕ is a *measure of association* based on a χ^2 test of homogeneity that evaluates the degree to which the value of one variable predicts the value of another. For any $r \times c$ table, we take *k* is the minimum of *r* and *c* and *N* is the grand total.

$$\text{Cramér's } \phi \equiv \sqrt{\frac{\chi^2}{N \times (k-1)}}\,.$$

However it is not easy to apply this measure to 'goodness of fit' conditions. Unpublished research papers on measures of association can be found on **corp.ling.stats**.[4]

5. Statistical separability

Finally, Wallis (to appear, a) 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.

This question is discussed in detail in Wallis (2011), which gives formulae for comparing different experimental tables (2 × 1, 2 × 2, $r \times 1$, $r \times c$ etc).

---

[3] See Wallis (to appear, a) for more information.
[4] A paper on goodness of fit measures of association is at http://corplingstats.wordpress.com/2012/03/31/gof-measures

6. Illustrative data for some worked examples[5]

Suppose we have two variables *A*, *B*, where

> $A = \{a, \neg a\}$ is the **independent variable** (IV),
> $B = \{b, \neg b\}$ is the **dependent variable** (DV).

A typical 2 × 2 experiment may ask whether *A* influences *B*, i.e. is the value of *B* dependent to some extent on the value of *A*?

For example, *B* could be the modal choice {*shall*, *will*}, *A* could represent different sociolinguistic conditions (speech vs. writing) or different grammatical conditions (interrogative vs. declarative).

Table 1 is an example 2 × 2 **contingency table** for *A* and *B*, which can be read as saying that for condition *a*, the number of cases that are *b* is 20 out of 30.

F	*a*	¬*a*	Σ
b	20	5	25
¬*b*	10	10	20
Σ	30	15	45

Table 1. An example 2 × 2 contingency table (frequency data *F*).

p	*a*	¬*a*	Σ
b	$^2/_3$	$^1/_3$	$^5/_9$
¬*b*	$^1/_3$	$^2/_3$	$^4/_9$

Table 2. Dividing by column totals rewrites Table 1 in terms of probability *p*.

This can be rewritten in terms of probabilities (Table 2), so $p(b \mid a) = {}^2/_3$, etc. The final column, Σ, summarises the overall probabilities $p(b)$, $p(\neg b)$.

We will use an error level of 0.05, so $z_{\alpha/2} \approx 1.96$, $\chi^2_{crit} \approx 3.841$ (for one degree of freedom = $z_{\alpha/2}{}^2$).

7. 2 × 1 goodness of fit χ^2 test / single-sample *z* test

We apply this method to compare the distribution for a single column (e.g. *a*) across the DV *B* with the overall distribution Σ (or, to put it another way, to compare $p(b \mid a)$ with $p(b)$):

(1) $\chi^2 = \Sigma\sqrt{(o-e)^2}/e = 0.667+0.833 = 1.5$,
$\chi^2 < \chi^2_{crit}$, so **non-significant**.

(2) *Gaussian interval* about $P = p(b) \pm z_{\alpha/2}.S = 0.556 \pm 0.178 = (0.378, 0.733)$,
$p(b \mid a) = 0.667$, $0.378 < p(b \mid a) < 0.733$, so **non-significant**.

The *z* test (2) allows us to pick a different value of *P* other than $p(b)$. The method only uses one row (*b*), whereas χ^2 employs both rows. Here is a third method using Wilson's interval about $p(b \mid a)$:

(3) *Wilson's score interval* $p(b \mid a) = (0.488, 0.808)$,
$P = p(b) = 0.556$, $0.488 < p(b) < 0.808$, so **non-significant**.

All three methods are alternative calculations for the same test, and obtain the same result.

8. 2 × 2 χ^2 test / *z* test for two independent proportions

Compare IV *A* across DV *B*, with overall expected distributions in proportion to final column Σ.

(1) $\chi^2 = 0.667+0.833+1.333+1.667 = 4.5$,
$\chi^2 > \chi^2_{crit}$, so **significant**.

(2) $P = p(b) = 0.556$, *standard deviation* $S' = 0.157$,
Gaussian difference interval = (-0.308, +0.308),
Difference $d = p_1 - p_2 = {}^2/_3 - {}^1/_3 = 0.333 > 0.308$, so **significant**.

(3) *Newcombe-Wilson interval* = (-0.230, +0.307): $p_1 - p_2 = 0.333 > 0.307$, so **significant**.

[5] A useful spreadsheet for replicating these calculations is at www.ucl.ac.uk/english-usage/statspapers/2x2chisq.xls

The **first two methods** always obtain the same result. However method (3) assumes *a* and ¬*a* are drawn from different populations (a sociolinguistic IV, such as speech vs. writing), whereas (1) and (2) assume they are drawn from the same population (e.g. a grammatical IV), where the same speaker may perform utterances in either category. See section 3 above.

With a continuity correction applied (see Wallis to appear, b), these results are non-significant.

9. References and suggested reading

Good range of 'state of art' papers in current change

Aarts, B., J. Close, G. Leech and S.A. Wallis (eds.) 2013. *The Verb Phrase in English*. Cambridge: CUP.

Introductions to statistical methods

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

Wallis, S.A. to appear, b. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*. www.ucl.ac.uk/english-usage/statspapers/binomialpoisson.pdf

General statistics references

Gries, S. Th. 2009. *Statistics for Linguistics with R*. Berlin/New York: Mouton de Gruyter.

Singleton, R. Jr., B.C. Straits, M.M. Straits and R.J. McAllister, 1988. *Approaches to social research*. New York, Oxford: OUP.

Sheskin, D.J. 1997. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, Fl: CRC Press.

Mathematical papers *(in many cases a more difficult read, but useful for citation)*

Newcombe, R.G. 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* **17**: 873-890.

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

Wilson, E.B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**: 209-212.

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