A crib sheet for statistical methods

Sean Wallis, Survey of English Usage

Readers are referred to Wallis (2013b) for more information on, and formulae for, methods summarised briefly below.[1] See also **corp.ling.stats** blog (http://corplingstats.wordpress.com) for papers.[2]

1. Confidence intervals

Confidence intervals on an observed rate *p* should be computed using the **Wilson score interval** method. 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 of confidence (e.g. 95%).

> **Note:** Confidence intervals can be applied to onomasiological change (variation in choice) and semasiological change (variation in meaning), provided that *P* is free to vary from 0 to 1 (see Wallis forthcoming). The interpretation of 'significant change' in either case is different though.

The methods for calculating intervals described below employ **the Gaussian approximation to the Binomial distribution**, i.e. we assume that the Binomial distribution predicted by combinatorial mathematics is approximately Normal.

Confidence intervals on Expected (Population) values (*P*)

The **Gaussian interval** about *P* is based on two values: the **mean** and **standard deviation** as follows:

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

The Gaussian interval about *P* can be written as $P \pm E$, where $E = z.S$, and z is the critical value of the standard Normal distribution at a given error level (e.g., 0.05). Although this is a bit of a mouthful, critical values of z are constant, so for any given level you can just substitute the constant for z. [$z(0.05) = 1.95996$ to six decimal places.] In summary:

$$\textit{Gaussian interval} \equiv P \pm z\sqrt{P(1-P)/N}\ .$$

Confidence intervals on Observed (Sample) values (*p*)

We cannot use the formula above for confidence intervals about **observations**. (Many people do this and it is, quite simply, wrong!) Most obviously, if *p* gets close to zero, the error *e* can be greater than *p*, so the lower bound of the interval can fall below zero, which is clearly impossible! The problem is most apparent on smaller samples (larger intervals) and skewed values of *p* (close to 0 or 1).

The Gaussian (Normal) distribution is a reasonable approximation for the expected distribution of an as-yet-unknown population probability *P*. It is incorrect for an interval around an observation *p* (Wallis 2013a). However, the latter case is precisely where the Gaussian interval is used most often!

To plot accurate intervals around observed *p* we need to use **Wilson's score interval**:

$$\textit{Wilson's score interval } (w^-, w^+) \equiv \left(p + \frac{z^2}{2N} \pm z\sqrt{\frac{p(1-p)}{N} + \frac{z^2}{4N^2}} \right) \Bigg/ \left(1 + \frac{z^2}{N} \right).$$

The score interval is **asymmetric** (except where *p*=0.5) and tends towards the middle of the distribution. It cannot exceed the probability range [0, 1] and it should always be used instead of the

[1] A spreadsheet is also at www.ucl.ac.uk/english-usage/staff/sean/resources/2x2chisq.xls

[2] See also http://corplingstats.wordpress.com/2012/04/01/crib-sheet

Gaussian, particularly with skewed data and small samples (a common condition in corpus linguistics). A continuity-corrected version of Wilson's interval should be used where the sample size N is small.

To employ intervals on 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 decreases the interval width.

2. Contingency correlation tests

Contingency correlation tests, including **log-likelihood**, **chi-square** (**χ^2**), and its variations, are premised on the **population** **z** **test** (Wallis 2013b).

The 2 × 1 **goodness of fit** χ^2 test (Figure 1) is a reformulation of a single sample z test based on a expected baseline frequency.[3]

We might use this to check

- whether the ratio of a term correlates with a baseline.
- to compare two competing frequencies (proportions of values of the same variable) for significant difference.

Similarly, the **2 × 2 χ^2 test of homogeneity (independence)** is identical to a **two-sample** **z** **test where samples are drawn from the same population** (Figure 2, see also Wallis 2013b).

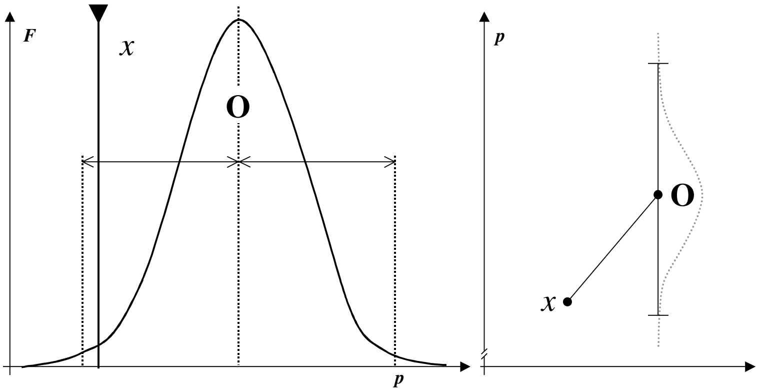

Figure 1: The single-sample population z test (= 'goodness of fit' χ^2 test).

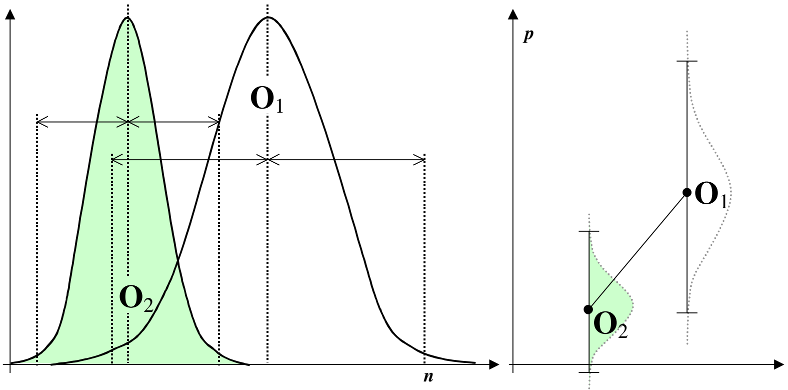

Figure 2: The 2 × 2 χ^2 test assumes uncertainty in both observations **O**$_1$ and **O**$_2$.

In this test the following χ^2 formula is applied to all cells o_{ij}, where $e_{ij} = n_r \times n_c / n$, where n_r and n_c are the row and column totals respectively, and n is the grand total.

$$\chi^2 = \sum_{\substack{i=1..r \\ j=1..c}} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

The z tests work by creating a new confidence interval out of the inner intervals at each point. For χ^2 the equivalent combined interval is based on the overall probability. So in Figure 2, O$_1$ and O$_2$ represent observed distributions about two points, and the new combined interval is related to the standard deviation (a measure of spread) of each distribution.

The optimum method of calculation is to employ **Yates' χ^2 test**. This can also be used for evaluating larger tables with more than two columns or rows. The main problem with larger $r \times c$ tables is interpretation: with more than 1 degree of freedom, a significant result merely tell you that the variables interact. One approach is discussed in Wallis (2013b): to restructure tables and refocus the

[3] The goodness of fit χ^2 test uses an interval on the expected probability P, so the Gaussian (z) interval is acceptable.

experimental design on key areas of variation. See also Nelson *et al.* (2002). A possibly more effective approach is to plot probabilities with confidence intervals and then identify areas of variation visually.

Although we have plotted confidence intervals on a dependent variable x, in fact, chi-square tests of independence pay no attention to which axis represents the dependent variable and which the independent variable. Swap them and you get the same result. These tests also assume that both variables are free to vary from zero to 1, and that data is drawn from the same population. This would be reasonable if both variable was lexical or grammatical, e.g. if we wished to know if one grammatical decision had an impact on another.

When samples are taken from **different** populations an alternative method is recommended. See Wallis (2013a). Robert Newcombe employs Wilson's score interval to create a more precise 2 × 2 test for samples taken from different populations. His method compares the difference in p values ('simple swing', below) with a new combined confidence interval based on Wilson's interval.

This **Newcombe-Wilson test** (preferably with a continuity-correction) should be used instead of Yates' χ^2 test when the independent variable classifies entire texts (e.g., by date, speaker gender etc.).

3. Effect size

To compare different results we can focus on these difference measures alone.

Wallis (2013b) notes that **simple swing**, $d = p_2 - p_1$, and **percentage swing**, $d^{\%} = d/p_1$, are commonly used for comparing p values, and explains how these may be plotted with confidence intervals. See also Aarts *et al.* (2013) for an illustration.

Faced with a different problem, Bowie *et al.* (2013) used a goodness of fit χ^2 score to compare the present perfect against two different baselines (present- and past- marked VPs). Although normally we are interested in large χ^2 values to demonstrate significant difference, in this case the smaller the χ^2 score, the closer the correlation can be said to be (hence 'goodness of fit').

More advanced methods use adaptations of χ^2. These include **Cramér's ϕ** and a modified goodness of fit ϕ_p (Wallis 2012), both of 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 (measuring change on both variables A and B).

$$\textit{Cramér's } \phi \equiv \sqrt{\chi^2 / (k-1)N} \,,$$

where $k = \min(r, c)$. Standardised **root mean square error ϕ_p** is designed for *goodness of fit* applications (estimating the degree of variation of a single subset a against a fixed baseline A). ϕ_p can measure variation over multiple points (such as text categories), whereas simple swing (difference) d can obviously only be based on two p values.

$$\textit{r.m.s. error } \phi_p \equiv \sqrt{\tfrac{1}{2}\sum (O_i - E_i)^2 \Big/ N} \,,$$

where O_i and E_i observed (term) and expected (baseline) frequencies for category i. Both ϕ measures are standardised to the probabilistic range [0, 1].

4. Separability tests and explaining results

Finally, Wallis (2013b) also points out that it is possible to compare a pair of 2 × 2 contingency tests for statistical separability, that is, **to test if the results are significantly different from each other.**

The idea is an extension of the derivation of the z test for the difference between two proportions (2 × 2 contingency test), by evaluating *the difference between two differences.* Wallis (2011) extends the paradigm to compare outcomes from any pair of identically-structured χ^2 tests (with equations for 2 × 1 goodness of fit, $r \times 1$ goodness of fit and $r \times c$ contingency test for independence).

> Suppose that you carry out the same experiment twice, but vary the conditions slightly. On the second attempt you appear to get a stronger effect than on the first. A separability test determines whether the difference between these two test outcomes is significant, i.e. that *one result is significantly greater than the other*.

Note that just because two results are *individually* significant (i.e. a change is significantly different from zero) does not mean that they are significantly different *from each other*. Likewise, just because one result reports a numerically greater size of effect, χ^2 score or error level than another does not mean that results are 'stronger'.

This is also why I advise against quoting χ^2 scores (or p error values) in papers. This practice, whilst common, is misleading.[4] **A better approach is to pick an error level, say, 0.05, and then stick to it.**

When reporting results you should use appropriate tests to draw out distinctions, and cite confidence intervals around probability values. For example you might say 'the number of finite VPs rose from between 34 and 40% to between 45 and 50%' when discussing change.

It is also perfectly legitimate to say 'around 37 percent' instead of 'between 34 and 40%', and leave the detail of intervals to tables and graphs. See, for example, Aarts *et al*. (2013).

References

ACLW: Aarts, B., J. Close, G. Leech and S.A. Wallis (eds.) (2013). *The Verb Phrase in English: Investigating recent language change with corpora*. Cambridge: CUP.
Preview at www.ucl.ac.uk/english-usage/projects/verb-phrase/book.

Aarts, B., J. Close and S.A. Wallis. 2013. Choices over time: methodological issues in investigating current change. ACLW 2. http://www.ucl.ac.uk/english-usage/projects/verb-phrase/book/aartsclosewallis.pdf

Bowie, J., S.A. Wallis and B. Aarts, 2013. The perfect in spoken English. ACLW 13. www.ucl.ac.uk/english-usage/projects/verb-phrase/book/bowiewallisaarts.pdf

Nelson, G., S.A. Wallis and B. Aarts. 2002. *Exploring Natural Language*. Amsterdam: John Benjamins.

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

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

Wallis, S.A. 2012. *Goodness of fit measures for discrete categorical data*. Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/gofmeasures.pdf

Wallis, S.A. 2013a. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *JQL* **20**:3, 178-208. www.ucl.ac.uk/english-usage/statspapers/binomialpoisson.pdf

Wallis, S.A. 2013b. z-squared: the origin and application of χ^2. *JQL* **20**:4, 350:378. www.ucl.ac.uk/english-usage/statspapers/z-squared.pdf

Wallis, S.A. forthcoming. *That vexed problem of choice*. London: Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/vexedchoice.pdf

[4] A χ^2 score is based on two things: the size of effect and the quantity of data, so a high score could simply mean that you have a lot of data. An error level (0.05, 0.01, 0.001, or series of asterisks, '*', '**', '***') is based on this score, so the logic of claiming greater strength with a smaller error is equally erroneous! In statistics, sadly, the commonality of a practice is no guide to its mathematical legitimacy.