observation. A confidence interval tells us that *t given eve of cert inty*, if our scientific model is correct, the true value in the population will likely be in the range identified. The larger the confidence interval, the less certain the observation will be. There are several different approaches to calculating confidence intervals, and we will begin by discussing the most common method.

. e d n er

The standardised ‘Wald’ confidence interval employs

Fully-skewed values, i.e. where *p*(*sh*) = zero or 1, obtain **zero- d n er** , highlighted in bold in Column A. However an interval of zero width represents complete certainty. We cannot say on the basis of a single observation that it is certain that all similarly-sampled speakers in 1958 used *sh* in place of *i* in first person declarative contexts! Secondly, Column C provides two for low or high p values or for small n – which is hardly satisfactory! Fewer than half the values of p(*sh*) in Table 1 satisfy this rule (the empty points in Figure 3). Needless to say, when it comes to line-fitting or other less explicit uses of this estimate, such limits tend to be forgotten.

A similar heuristic for the χ^2 test (the Cochran rule) avoids employing the test where expected cell values fall below 5. This has proved so unsatisfactory that a series of statisticians have proposed competing alternatives to the chi-square test such as the log-likelihood test, in a series of attempts to cope with low frequencies and skewed datasets. In this paper we distinguish two mathematical problems with the Wald interval – that it incorrectly characterises the interval about p and that it fails to correct for continuity – and then evaluate competing test methods by a combination of plotting limits and exhaustive computation.

. on core n erv

The key problem with the conventional Wald definition is that *the confidence interv is incorrect y ch r cterised*. Note how we assumed that the interval about p was Binomial and could be approximated by the Normal distribution. This is the wrong way to think about the problem, but it is such a common error that it needs to be addressed.

The correct characterisation is a little counter-intuitive, but it can be summarised as follows.

Imagine a true population probability, which we will call P. This is the *ct v e* in the population. Observations about P will be distributed according to the Binomial. We don’t know precisely what P is, but we can try to observe it indirectly, by sampling the population.

Given an observation p, there are, potentially, two values of P which would place p at the outermost limits of a confidence interval about P. See Figure 4. What we can do, therefore, is *se rch* for values of P which satisfy the formula used to characterise the Normal approximation to the Binomial about P.[2] Now we have the following definitions:

pop tion e n $\mu \equiv P$,
pop tion st nd rd dev i tion $\sigma \equiv \sqrt{P(1-P)/n}$, (2)
pop tion confidence interv $(E^-, E^+) \equiv (P - z_{\alpha/2}.\sigma, P + z_{\alpha/2}.\sigma)$.

The formulae are the same as (1) but the symbols have changed. The symbols μ and σ

$P_1 + z_{\alpha/2}.\sigma_1$ and $p = \overline{E_2} = P_2 - z_{\alpha/2}.\sigma_2$. With a computer we can perform a search process to converge on the correct values.

The formula for the population confidence interval above is a Normal z interval about the population probability P. This interval can be used to carry out the z test for the population probability. This test is equivalent to the 2 × 1 goodness of fit χ^2 test, which is a test where the population probability is simply the expected probability $P = E/n$.

Fortunately, rather than performing a computational search process, it turns out that there is a simple method for directly calculating the sample interval about p. This interval is called the **on core n er** (Wilson, 1927) and may be written as

$$i\ son\ score\ inter \quad (\ ^-, \quad) \equiv \left. p + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n} + \frac{z^2_{\alpha/2}}{4n^2}} \right/ 1 + \frac{z^2_{\alpha/2}}{n} . \qquad (4)$$

The score interval can be broken down into two parts on either side of the plus/minus ('±') sign:

1) a relocated centre estimate $p \ = \ \left. p + \frac{z^2_{\alpha/2}}{2n} \right/ 1 + \frac{z^2_{\alpha/2}}{n}$ and

2) a corrected *st nd rd de i tion* $s \ = \ \left. \sqrt{\frac{p(1-p)}{n} + \frac{z^2_{\alpha/2}}{4n^2}} \right/ 1 + \frac{z^2_{\alpha/2}}{n}$,

such that $^- = p \ - z_{\alpha/2}.s$ and $\ = p \ + z_{\alpha/2}.s$.[3] We will use lower case to refer to the Wilson interval.

The 2 × 1 goodness of fit χ^2 test checks for the sample probability falling within Gaussian intervals on the population distribution, i.e. $E^- < p < E$. This obtains the same result as testing the population probability within the sample confidence intervals, $^- < P <$. We find that where $P = \ ^-$, $p = E$, which is sketched in Figure 4. As the diagram indicates, whereas the Normal distribution is symmetric, the Wilson interval is asymmetric (unless $p = 0.5$).

Employing the Wilson interval on a sample probability does not itself improve on this ^{2} test. It obtains exactly the same result by approaching the problem from p rather than P. The improvement is in estimating the confidence interval around p!

If we return to Table 1 we can now plot confidence intervals on first person p(*sh*) over time, using the upper and lower Wilson score interval bounds in Columns D and E. Figure 5 depicts the same data. Previously zero-width intervals have a large width – as one would expect, they represent highly uncertain observations rather than certain ones – in some instances, extending nearly 80% of the probabilistic range. The overshooting 1960 and 1970 datapoints in Figure 3 fall within the probability range. 1969 and 1972, which extended over nearly the entire

[3] One alternative proposal, termed the Agresti-Coull interval (Brown *et* 2001) employs the adjusted Wilson centre p and then substitutes it for p into the Wald standard deviation s (see Equation 1). We do not consider this interval here, whose merits primarily concern ease of presentation. Its performance is inferior to the Wilson interval.

range, have shrunk.

How do these intervals compare overall? As we have seen, the Wilson interval is asymmetric. In Equation 4, the centre-point, p , is pushed towards the centre of the probability range. In addition, the total width of the interval is $2z_{\alpha/2}.s$ (i.e. proportional to s). We compare s and s by plotting across p for different values of sample size n in Figure 6. Note that the Wilson deviation s never reaches zero for low or high p, whereas the Gaussian deviation always converges to zero at extremes (hence the zero-width interval behaviour). The differences between curves reduces with increasing n

tail is (1 – P). The coin may be biased, so P need not be 0.5!

The population Binomial distribution of r heads out of n tosses of a coin with weight P is defined in terms of a series of discrete probabilities for r, where the height of each column is defined by the following expression (Sheskin, 1997: 115):

$$\textit{Binomial probability } B(r; n, P) \equiv nCr.P^r(1-P)^{(n-r)}. \qquad (5)$$

This formula consists of two components: the Binomial combinatorial nCr (i.e. how many ways one can obtain r heads out of n [illegible]

Figure 9 shows that log-likelihood matches the Binomial P more closely than χ^2 for $r \leq 3$, $n = 5$ and $\alpha = 0.05$, which may explain why some researchers such as Dunning (1993) have (incorrectly) claimed its superiority. However it is less successful than uncorrected χ^2 overall. In any event, it is clearly inferior to Yates' χ^2 (cf. Figure 7 and Table 2).

3. *Evaluating confidence intervals*

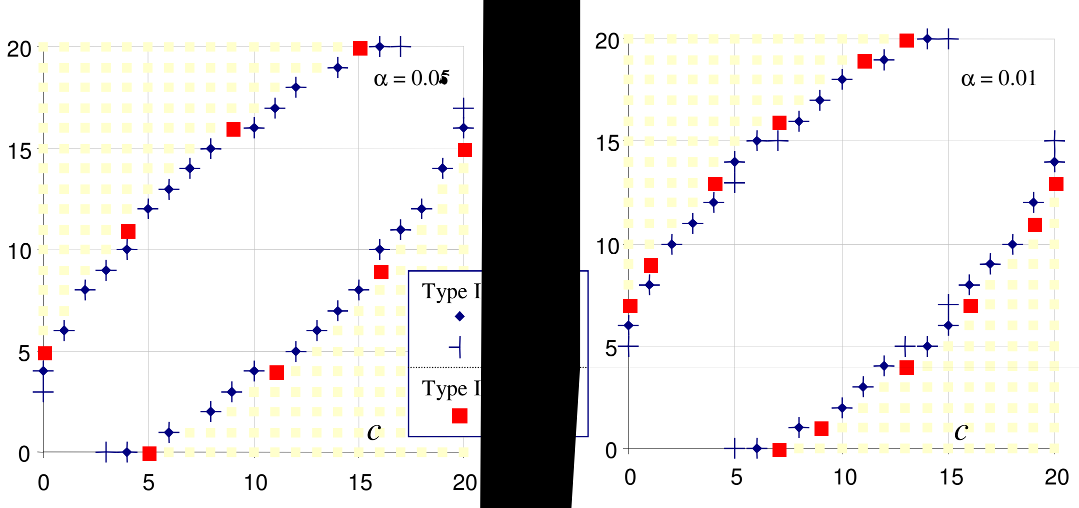
Thus far we have simply compared the behaviour of the interval lower bound over values of . This tells us that different methods obtain different results, but does not really inform us about the scale of these discrepancies and their effect on empirical research. To address this question we need to consider other methods of evaluation.

3. Measuring errors

Statistical procedures should be evaluated in terms of the rate of two distinct types of error:

- **Type I errors**, or false positives: this is so-called 'anti-conservative' behaviour, i.e. *rejecting*

likelihood are 0.0095, 0.0014 and 0.0183 respective



In *e post f cto* corpus analysis, this corresp o a situation where samples are taken from the **e pop on** and the independent variab well as the dependent variable) represents a free choice by the speaker. This is a within-sub design, where either value of the **ndependen r e** (IV) may be uttered by the same s r or appear in the same source text. Alternative tests are the 2 × 2 χ^2 test (including Yates' tes l log-likelihood test. These tests can be translated into confidence intervals on the difference be p_1 and p_2 (Wallis forthcoming).

We may objectively evaluate tests by identif Type I and II errors for conditions where the tests do not agree with the result obtained by Fish um test. Figure 12 plots a map of all tables of the form [[, *b*] [*c*, *d*]] for all integer values of , *d* where n_1 = + *b* = 20 and $n_2 = c + d = 20$. We can see that in both cases, there are slightly e errors generated by G^2 than χ^2, and Yates' χ^2

Given this common derivation, we would anticipate that this second pairwise comparison will obtain comparable results to the evaluation of intervals for the single proportion discussed in section 3. Figure 15 plots the result of comparing Newcombe-Wilson tests, with and without continuity

$5 \times n_2$

We found that the optimum tests were Yates' test (when data is drawn from the same population) and the Newcombe-Wilson test with continuity correction (for data drawn from independent populations). Yates' test can also be used in the latter condition, and is advisable if the smaller sample size (row total) is 15 or below.

It is worth noting that the corresponding z test suggested by Sheskin (1997) performs poorly because it generalises from the Wald interval. Log-likelihood also performs poorly in all cases, despite its adherents (e.g. Dunning 1993) whose observations appear premised on only the lower part of the interval range. Our results are consistent with Newcombe (1998b) who uses a different evaluation method and identifies that the tested Newcombe-Wilson inner ('mesial') interval is reliable.

Finally, the Bienaymé formula (15) may also be employed to make another useful generalisation. In Wallis (2011) we derive a set of "meta-tests" that allow us to evaluate whether the results of two structurally identical experiments performed on different data sets are significantly different from one another. This allows researchers to compare results obtained with different data sets or corpora, compare results under different experimental conditions, etc. Meta-testing has also been used to pool results which may be individually insignificant but are legitimate to consolidate.

Our approach is superior to comparing effect size numerically or making the common logical error of inferring that, e.g., because one result is significant and another not, the first result is 'significantly greater' than the second. (Indeed, two individually non-significant test results may be significantly different because observed variation is in opposite directions.)

The resulting meta-test is based on comparing the optimum sub-tests we evaluate Wower

is that if any utterance by any speaker could be accounted for in any cell in the table, then the summation should be performed in both directions at the same time.

An alternative test using the same configuration is more appropriate when samples are taken from different populations, and the independent variable is not free to vary. In this case we sum ‘exact’ Binomial (Clopper-Pearson) intervals (section 4.2) in one direction only: within each sample (finding P for Equation 7), and then combine intervals assuming that variation is independent (Equation 15).

We may compare the performance of the two tests by the same method as in section 6 of the paper: identify table configurations where one test obtains a significant result and the other does not. For $n_1 = n_2$ up to 100 and $n_1 = 5n_2$ we compare the ahmepawo tp ell 1y mlsee b tmecone