

---

## Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods\*

Sean Wallis

Survey of English Usage, University College, London, GB

---

### ABSTRACT

Many statistical methods rely on an underlying mathematical model of probability based on a simple approximation, one that is simultaneously well-known and yet frequently misunderstood. The Normal approximation to the Binomial distribution underpins a range of statistical tests and methods, including the calculation of accurate confidence intervals, performing goodness of fit and contingency tests, line- and model-fitting, and computational methods based upon these. A common mistake is in assuming that, since the probable distribution of error about the “true value” in the population is approximately Normally distributed, the same can be said for the error about an observation.

This paper is divided into two parts: fundamentals and evaluation. First, we examine the estimation of confidence intervals using three initial approaches: the “Wald” (Normal) interval, the Wilson score interval and the “exact” Clopper-Pearson Binomial interval. Whereas the first two can be calculated directly from formulae, the Binomial interval must be approximated towards by computational search, and is computationally expensive. However this interval provides the most precise significance test, and therefore will form the baseline for our later evaluations. We also consider two further refinements: employing log-likelihood in intervals (also requiring search) and the effect of adding a continuity correction.

Second, we evaluate each approach in three test paradigms. These are the single proportion interval or  $2 \times 1$  goodness of fit test, and two variations on the common  $2 \times 2$  contingency test. We evaluate the performance of each approach by a “practitioner strategy”. Since standard advice is to fall back to “exact” Binomial tests in conditions when approximations are expected to fail, we report the proportion of instances where one test obtains a significant result when the equivalent exact test does not, and vice versa, across an exhaustive set of possible values.

We demonstrate that optimal methods are based on continuity-corrected versions of the Wilson interval or Yates’ test, and that commonly-held beliefs about weaknesses of  $\chi^2$  tests

---

\*Address for correspondence: Sean Wallis, Survey of English Usage, Department of English, University College London, Gower Street WC1E 6BT, London, England. Tel: 020 7679 3120. Email: [s.wallis@ucl.ac.uk](mailto:s.wallis@ucl.ac.uk).

are misleading. Log-likelihood, often proposed as an improvement on  $\chi^2$ , performs disappointingly. Finally we note that at this level of precision we may distinguish two types of  $2 \times 2$  test according to whether the independent variable partitions data into independent populations, and we make practical recommendations for their use.

## 1. INTRODUCTION

Estimating the error in an observation is the first, crucial step in inferential statistics. It allows us to make predictions about what would happen were we to repeat our experiment multiple times, and because each observation represents a sample of the population, predict the true value in the population (Wallis forthcoming).

Consider an observation that a proportion  $p$  of a sample of size  $n$  is of a particular type. For example:

- the proportion  $p$  of coin tosses in a set of  $n$  throws that are heads,
- the proportion of light bulbs  $p$  in a production run of  $n$  bulbs that fail within a year,
- the proportion of patients  $p$  who have a second heart attack within six months after a drug trial has started ( $n$  being the number of patients in the trial),
- the proportion  $p$  of interrogative clauses  $n$  in a spoken corpus that are finite.

We have one observation of  $p$ , as the result of carrying out a single experiment. We now wish to infer about the future. We would like to know how reliable our observation of  $p$  is without further sampling. Obviously, we do not want to repeat a drug trial on cardiac patients if the drug may be adversely affecting their survival.<sup>1</sup>

## 2. COMPUTING CONFIDENCE INTERVALS

We need to estimate the “margin of error” or to use the proper term, *confidence interval*, on our observation. A confidence interval tells us that *at a given level of certainty*, 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

---

<sup>1</sup>A very important application of confidence intervals is determining *how much data is enough* to rule that a change is significant. A large decrease in survivability among patients would lead one to stop the trial early. But one early death could be accidental.

the observation will be. There are several different approaches to calculating confidence intervals, and we will begin by discussing the most common method.

**2.1 The “Wald” Interval**

The standardized “Wald” confidence interval employs the Normal approximation to the Binomial distribution sketched in Figure 1. The actual distribution, shown by the columns, is assumed to be a discrete Binomial distribution, but to obtain the interval we first approximate it to a continuous Normal curve, shown by the line. This relies on the following definitions.

$$\begin{aligned}
 \text{mean } \bar{x} &\equiv p, \\
 \text{standard deviation } s &\equiv \sqrt{p(1-p)/n}, \\
 \text{confidence interval } (e^-, e^+) &\equiv (p - z_{\alpha/2} \cdot s, p + z_{\alpha/2} \cdot s)
 \end{aligned}
 \tag{1}$$

where  $n$  represents the sample size,  $p$  the proportion of the sample in a particular class and  $z_{\alpha/2}$  is the critical value of the Normal distribution for a given error level  $\alpha$ . This means that if data is Normally distributed, and the error level  $\alpha$  is 0.05, 95% of the expected distribution is within this interval, and only 2.5% in each of the “tails” outside. This critical value is 1.95996.

The larger the value of  $n$  the more “continuous” the line, and the more confident we can be in  $p$ , so the confidence interval will shrink as  $n$  increases.

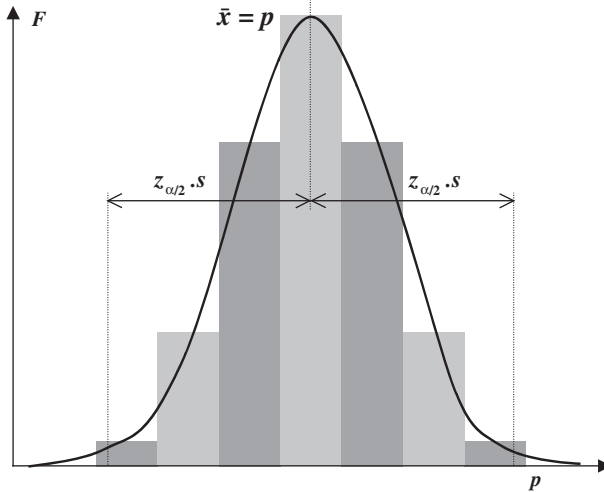


Fig. 1. The Normal approximation to the Binomial plotted within the probabilistic range  $p \in [0, 1]$ .

But what happens if  $n$  is small or  $p$  is close to zero or 1? Whereas the Normal distribution is assumed to be unconstrained (the tails go off in either direction to infinity),  $p$  cannot, for obvious reasons, exceed the range  $[0, 1]$ .

Two issues arise. First, as we shall see, where  $p$  tends to 0 or 1, the product  $p(1 - p)$  also tends to 0, leading to an underestimation of the error. Second, although  $s$  tends to zero, the interval can cross zero. However, points on the axis where  $p < 0$  (or  $p > 1$ ) are impossible to reach (Figure 2), so the approximation *fails*. Since linguists are often interested in changes in low frequency events, this is not an unimportant question!

Aarts, Close, and Wallis (2013) examined the alternation over time in British English from first person declarative uses of modal *shall* to *will* over a 30-year period by plotting over time the probability of selecting *shall* given the choice, which we can write as  $p(\text{shall} \mid \{\text{shall}, \text{will}\})$ . Their data is reproduced in Table 1. Note that the dataset has a number of attributes: data is *sparse* (this corpus is below 1 million words) and many data points are *skewed*: observed probability does not merely approach zero or 1 but reaches it.

We have added five columns to Table 1. Column A contains the Wald 95% error interval width  $z_{\alpha/2} \cdot s$ , B and C contain the lower and upper bounds  $e^-$ ,  $e^+$  respectively, obtained by subtracting and adding Column A from  $p(\text{shall})$ . Columns D and E contain the lower and upper bounds of the Wilson interval described in Section 2.2.

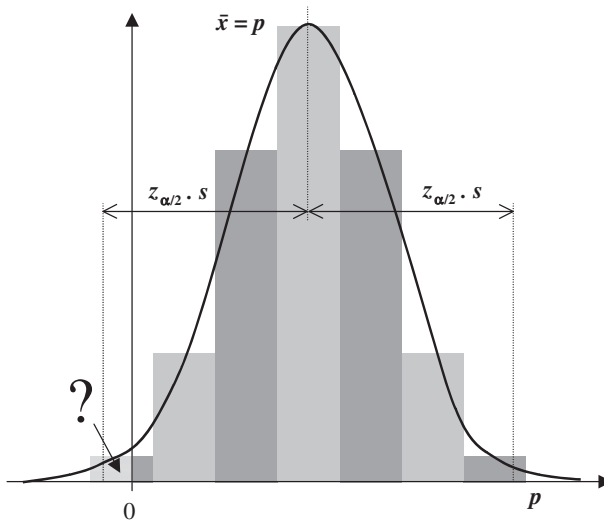


Fig. 2. As Figure 1, but  $p$  is close to zero. What happens if the curve crosses 0 or 1?

Table 1. Alternation of first person declarative modal *shall* vs. *will* over recent time, data from the spoken DCPSE corpus (after Aarts et al., 2013).

| Year | <i>shall</i> | <i>will</i> | Total $n$ | $p(\textit{shall})$ | A: $z_{\alpha/2} \cdot s$ | B: $e^-$ | C: $e^+$      | D: $w^-$ | E: $w^+$ |
|------|--------------|-------------|-----------|---------------------|---------------------------|----------|---------------|----------|----------|
| 1958 | 1            | 0           | 1         | 1.0000              | <b>0.0000</b>             | 1.0000   | 1.0000        | 0.2065   | 1.0000   |
| 1959 | 1            | 0           | 1         | 1.0000              | <b>0.0000</b>             | 1.0000   | 1.0000        | 0.2065   | 1.0000   |
| 1960 | 5            | 1           | 6         | 0.8333              | 0.2982                    | 0.5351   | <b>1.1315</b> | 0.4365   | 0.9699   |
| 1961 | 7            | 8           | 15        | 0.4667              | 0.2525                    | 0.2142   | 0.7191        | 0.2481   | 0.6988   |
| 1963 | 0            | 1           | 1         | 0.0000              | <b>0.0000</b>             | 0.0000   | 0.0000        | 0.0000   | 0.7935   |
| 1964 | 6            | 0           | 6         | 1.0000              | <b>0.0000</b>             | 1.0000   | 1.0000        | 0.6097   | 1.0000   |
| 1965 | 3            | 4           | 7         | 0.4286              | 0.3666                    | 0.0620   | 0.7952        | 0.1582   | 0.7495   |
| 1966 | 7            | 6           | 13        | 0.5385              | 0.2710                    | 0.2675   | 0.8095        | 0.2914   | 0.7679   |
| 1967 | 3            | 0           | 3         | 1.0000              | <b>0.0000</b>             | 1.0000   | 1.0000        | 0.4385   | 1.0000   |
| 1969 | 2            | 2           | 4         | 0.5000              | 0.4900                    | 0.0100   | 0.9900        | 0.1500   | 0.8500   |
| 1970 | 3            | 1           | 4         | 0.7500              | 0.4243                    | 0.3257   | <b>1.1743</b> | 0.3006   | 0.9544   |
| 1971 | 12           | 6           | 18        | 0.6667              | 0.2178                    | 0.4489   | 0.8844        | 0.4375   | 0.8372   |
| 1972 | 2            | 2           | 4         | 0.5000              | 0.4900                    | 0.0100   | 0.9900        | 0.1500   | 0.8500   |
| 1973 | 3            | 0           | 3         | 1.0000              | <b>0.0000</b>             | 1.0000   | 1.0000        | 0.4385   | 1.0000   |
| 1974 | 12           | 8           | 20        | 0.6000              | 0.2147                    | 0.3853   | 0.8147        | 0.3866   | 0.7812   |
| 1975 | 26           | 23          | 49        | 0.5306              | 0.1397                    | 0.3909   | 0.6703        | 0.3938   | 0.6630   |
| 1976 | 11           | 7           | 18        | 0.6111              | 0.2252                    | 0.3859   | 0.8363        | 0.3862   | 0.7969   |
| 1990 | 5            | 8           | 13        | 0.3846              | 0.2645                    | 0.1202   | 0.6491        | 0.1771   | 0.6448   |
| 1991 | 23           | 36          | 59        | 0.3898              | 0.1244                    | 0.2654   | 0.5143        | 0.2758   | 0.5173   |
| 1992 | 8            | 8           | 16        | 0.5000              | 0.2450                    | 0.2550   | 0.7450        | 0.2800   | 0.7200   |

Fully-skewed values, i.e. where  $p(\textit{shall}) = \text{zero or } 1$ , obtain zero-width intervals, 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 *shall* in place of *will* in first person declarative contexts! Secondly, Column C provides two examples (1960, 1970) of overshoot, where the upper bound of the interval exceeds the range  $[0, 1]$ . Again, as Figure 2 illustrates, any part of an interval outside the probabilistic range simply cannot be obtained, indicating that the interval is miscalculated. To illustrate this we plot Table 1 data in Figure 3.

Common statistical advice (the “3-sigma rule”) outlaws extreme values and requires  $p \pm 3s \in [0, 1]$  before employing the Wald interval. Some 99.7% of the Normal distribution is within three standard deviations of the mean. However, this rule has the effect that we simply give up estimating

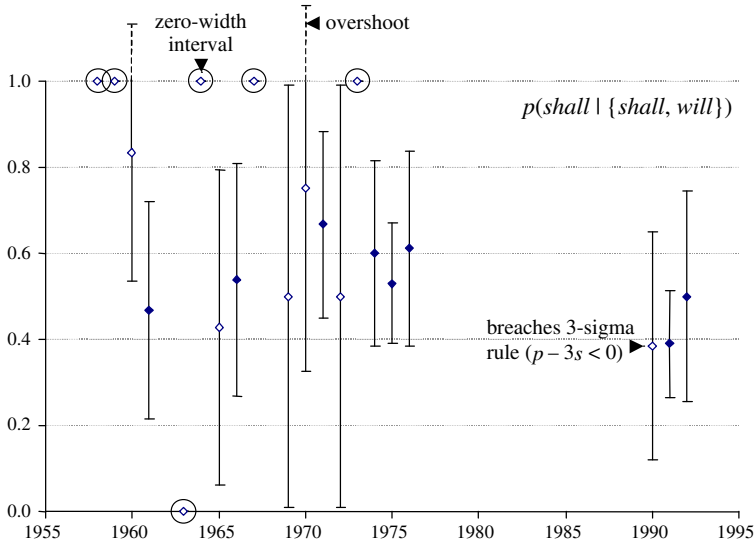


Fig. 3. Plot of  $p(\text{shall})$  over time, data from Table 1, with 95% Wald intervals, illustrating overshoot (dotted lines), zero-width intervals (circles), and 3-sigma rule failures (empty points).

the error for low or high  $p$  values or for small  $n$  – which is hardly satisfactory! Fewer than half the values of  $p(\text{shall})$  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  $\chi^2$  test (the Cochran rule) avoids employing the test where expected cell values fall below five. 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 characterizes 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.

### 2.2 Wilson’s Score Interval

The key problem with the conventional Wald definition is that *the confidence interval is incorrectly characterized*. 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 characterization is a little counter-intuitive, but it can be summarized as follows.

Imagine a true population probability, which we will call  $P$ . This is the *actual value* in the population. Observations about  $P$  will be distributed according to the Binomial. We do not 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 *search* for values of  $P$  which satisfy the formula used to characterize the Normal approximation to the Binomial about  $P$ .<sup>2</sup> Now we have the following definitions:

$$\begin{aligned}
 &\text{population mean } \mu \equiv P, \\
 &\text{population standard deviation } \sigma \equiv \sqrt{P(1 - P)/n}, \\
 &\text{population confidence interval } (E^-, E^+) \equiv (P - z_{\alpha/2} \cdot \sigma, P + z_{\alpha/2} \cdot \sigma).
 \end{aligned}
 \tag{2}$$

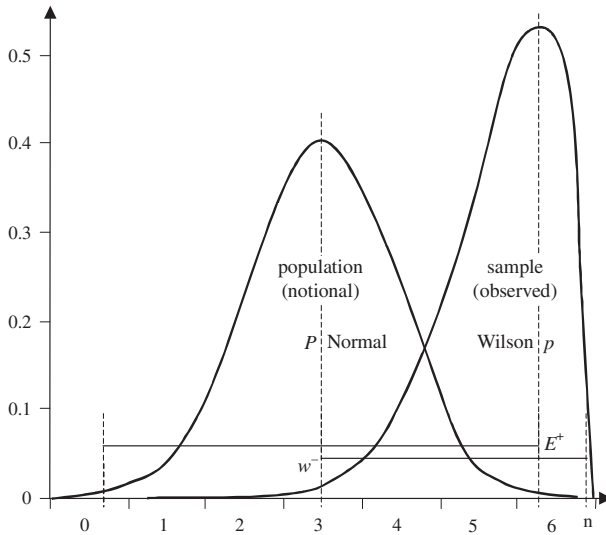


Fig. 4. The interval equality principle with Normal and Wilson intervals: the lower bound for  $p$  is  $P$ .

<sup>2</sup>In other words, we employ a computer program which estimates  $P$ , tests it, uses the resulting discrepancy between the test result and the optimum to improve the estimate, and repeat until this deviation is infinitesimal. There are a number of possible formulae for calculating the interval that can be slotted into this procedure, but we will come to this later.

The formulae are the same as (1) but the symbols have changed. The symbols  $\mu$  and  $\sigma$ , referring to the population mean and standard deviation respectively, are commonly used. This population confidence interval identifies two limit cases where  $p = P \pm z_{\alpha/2} \cdot \sigma$ .

Consider now the confidence interval around the sample observation  $p$ . We do not know  $P$  in the above, so we cannot calculate this imagined population confidence interval. It is a theoretical concept!

However the following interval equality principle must hold, where  $e^-$  and  $e^+$  are the lower and upper bounds of a sample interval for any error level  $\alpha$ :

$$\begin{aligned} e^- = P_1 &\iff E_1^+ = p \text{ where } P_1 < p, \text{ and} \\ e^+ = P_2 &\iff E_2^- = p \text{ where } P_2 > p. \end{aligned} \tag{3}$$

If the lower bound for  $p$  (labelled  $e^-$ ) is a possible population mean  $P_1$ , then the upper bound of  $P_1$  would be  $p$ , and vice-versa. Since we have formulae for the upper and lower intervals of a population confidence interval, we can attempt to find values for  $P_1$  and  $P_2$  which satisfy  $p = E_1^+ = P_1 + z_{\alpha/2} \cdot \sigma_1$  and  $p = E_2^- = P_2 - z_{\alpha/2} \cdot \sigma_2$ . 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  $2 \times 1$  goodness of fit  $\chi^2$  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 *Wilson score interval* (Wilson, 1927) and may be written as

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

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

- (1) a relocated centre estimate  $p' = \left( p + \frac{z_{\alpha/2}^2}{2n} \right) / \left( 1 + \frac{z_{\alpha/2}^2}{n} \right)$  and
- (2) a corrected standard deviation  $s' = \sqrt{\frac{p(1-p)}{n} + \frac{z_{\alpha/2}^2}{4n^2}} / \left( 1 + \frac{z_{\alpha/2}^2}{n} \right),$



such that  $w^- = p' - z_{\alpha/2} \cdot s'$  and  $w^+ = p' + z_{\alpha/2} \cdot s'$ .<sup>3</sup> We will use lower case  $w$  to refer to the Wilson interval.

The  $2 \times 1$  goodness of fit  $\chi^2$  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,  $w^- < P < w^+$ . We find that where  $P = w^-$ ,  $p = E^+$ , 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  $\chi^2$  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  $p$  (*shall*) 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 across nearly 80% of the probabilistic range. The overshooting 1960 and 1970 data points in Figure 3 fall within the probability range. 1969 and 1972, which extended over nearly the entire 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  $2z_{\alpha/2} \cdot s'$  (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 reduce with increasing  $n$  (lower) but this problem of extreme values continues to afflict Wald intervals.<sup>4</sup>

<sup>3</sup>One alternative proposal, termed the Agresti-Coull interval (Brown et al., 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.

<sup>4</sup>Newcombe (1998a) evaluates these and a number of other intervals (including the Clopper-Pearson “exact” Binomial calculation (4), and employing continuity corrections to Normal and Wilson intervals, which we discuss in the following sections). The Wilson statistic without correction performs extremely well even when compared with exact methods. He concludes that the Normal interval (1) should be abandoned in favour of the Wilson (3).

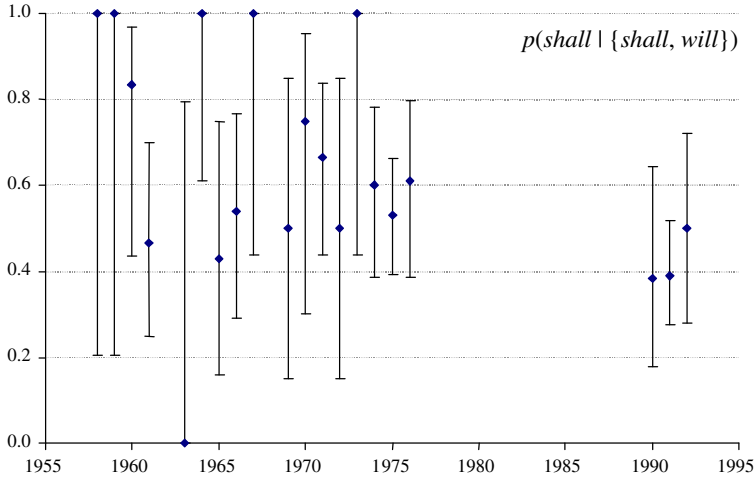


Fig. 5. Plot of  $p(\text{shall})$  over time, data from Table 1, with 95% Wilson score confidence intervals (after Aarts et al., 2013).

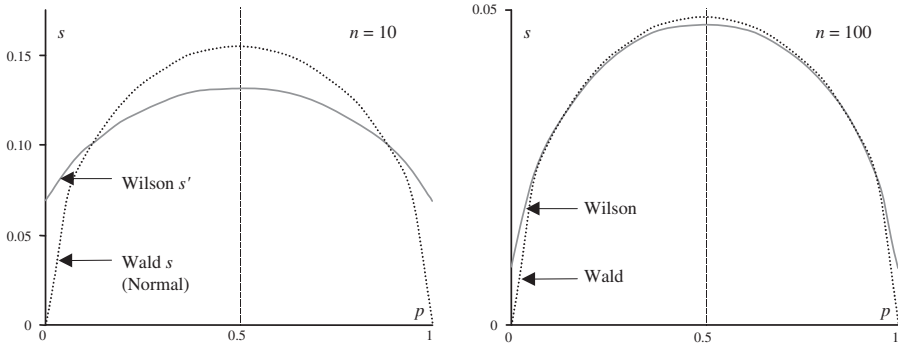


Fig. 6. Wald and Wilson standard deviations  $s, s'$  for  $p \in [0, 1]$ .

### 2.3 The “Exact” Binomial Interval

So far we have employed the Normal *approximation* to the Binomial distribution, and contrasted Wald and Wilson methods. To evaluate formulae against an ideal distribution we need a baseline. We need to calculate  $P$  values from first principles. To do this we use the Binomial formula. Recall from Figure 1 that the Binomial distribution is a discrete distribution, i.e. it can be expressed as a finite series of probability values for different values of  $x = \{0, 1, 2, 3, \dots, n\}$ .

We will consider the lower bound of  $p$ , i.e. where  $P < p$  (as in Figure 4). There are two interval boundaries on each probability, but the argument is symmetric: we could apply the same calculation substituting  $q = 1 - p$ , etc. in what follows.

Consider a coin-tossing experiment where we toss a weighted coin  $n$  times and obtain  $r$  heads (sometimes called “Bernoulli trials”). The coin has a weight  $P$ , i.e. the *true value in the population* of obtaining a head is  $P$ , and the probability of a 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, p. 115):

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

This formula consists of two components: the Binomial combinatorial  $nCr$  (i.e. how many ways one can obtain  $r$  heads out of  $n$  tosses)<sup>5</sup>, and the probability of each single pattern of  $r$  heads and  $(n - r)$  tails appearing, based on the probability of a head being  $P$ .

The total area of Binomial columns from  $x_1$  to  $x_2$  inclusive is then the *Cumulative Binomial probability*:

$$B(x_1, x_2; n, P) \equiv \sum_{r=x_1}^{x_2} B(r; n, P) = \sum_{r=x_1}^{x_2} nCr \cdot P^r(1 - P)^{(n-r)}$$

However, this formula assumes we know  $P$ . We want to find an exact upper bound for  $p = x/n$  at a given error level  $\alpha$ . The Clopper-Pearson method employs a computational search procedure to sum the upper tail from  $x$  to  $n$  to find  $P$  where the following holds:

$$B(x, n; n, P) = \alpha/2. \tag{7}$$

This obtains an exact result for any integer  $x$ . The computer modifies the value for  $P$  until the formula for the remaining “tail” area under the curve converges on the required value,  $\alpha/2$ . We then report  $P$ .<sup>6</sup>

<sup>5</sup>There is only 1 way of obtaining all heads (HHHHHH), but 6 different patterns give 1 tail and 5 heads, etc. The expression  $nCr = n! / \{r! (n - r)!\}$ , where “!” refers to the factorial.

<sup>6</sup>This method is Newcombe’s (1998a) method 5 using exact Binomial tail areas. In Figure 6 we estimate the interval for the mean  $p$  by summing  $B(0, r; n, p) < \alpha/2$ .

Note how this method is consistent with the idea of a confidence interval on an observation  $p$ : to identify a point  $P$ , sufficiently distant from  $p$  for  $p$  to be considered just significantly different from  $P$  at the level  $\alpha/2$ . As in Section 2.2, *we do not know* the true population value  $P$  but we expect that data would be Binomially distributed around it.

Figure 7 shows the result of computing the lower bound for  $p = P$  employing this Binomial formula. We also plot the Wilson formula, with and without an adjustment termed a “continuity correction”, which we will discuss in the next section. As we have noted, the Wilson formula for  $p$  is equivalent to a  $2 \times 1$  goodness of fit  $\chi^2$  based on  $P$ . The continuity-corrected formula is similarly equivalent to Yates’  $2 \times 1$   $\chi^2$ .

All three methods obtain lower confidence intervals on  $p$  which tend towards zero at  $x = 0$ , but do not converge to zero at  $x = n$ . Even with a tiny sample,  $n = 5$ , the continuity-corrected Wilson interval is very close to the “exact” population Binomial obtained using the search procedure, but it is much easier to calculate.

Recall that the argument we are using is symmetric. The dotted line at the top of Figure 7 is the upper bound for the exact population Binomial interval, which flips this around. At the extremes are highly skewed intervals, as we expected.

What happens if we use the naïve Wald interval? Figure 8 shows the effect of incorrectly characterizing the interval about  $p$ . The axes,  $n$  and  $p$ , are more-or-less swapped. The intervals tend towards zero at  $x = n$  but are very large (and become negative) for small  $x$ .<sup>7</sup>

## 2.4 Continuity Correction and Log-likelihood

We have addressed the major conceptual problem that the sample probability should not be treated as the centre of a Binomial distribution. However we have also seen that for small sample size  $n$ , the Wilson interval underestimates the error compared to the Binomial interval.

We can predict, therefore, that the corresponding uncorrected  $\chi^2$  test may find some results “significant” which would not be deemed significant if the exact Binomial test was performed. The area between the two curves in Figure 7 represents this tendency to make so-called “Type I” errors – where results are incorrectly interpreted as significant (see Section 3).

---

<sup>7</sup>The Binomial “curve” for  $p$  in Figure 8 is discrete – it consists of rationals  $r/n$  – and conservative, because the sum is *less* than  $\alpha/2$  rather than exactly equal to it.

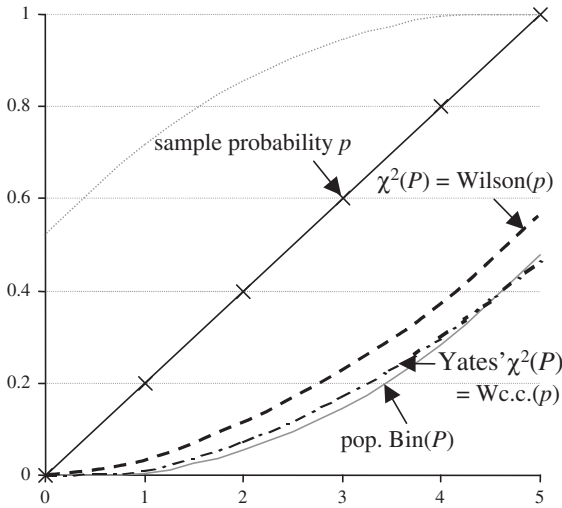


Fig. 7. Values of  $P$  where sample  $p$  is at the upper bound of  $P$ :  $n = 5, \alpha = 0.05$ .

We can now consider a couple of common alternative contingency tests against the exact Binomial population probability. In particular we have Yates'  $\chi^2$  test and the log-likelihood test (Equation (10)), both of which have been posited as improvements on  $\chi^2$ . Yates' formula for  $\chi^2$  introduces a continuity-correction term which subtracts 0.5 from each squared term:

$$Yates' \chi^2 \equiv \sum \frac{(O - E - 0.5)^2}{E}, \tag{8}$$

where  $O$  and  $E$  represent observed and expected distributions respectively. In our  $2 \times 1$  case we have  $O = \{np, n(1 - p)\}$  and  $E = \{nP, n(1 - P)\}$ . Employing a search procedure on Yates'  $\chi^2$  test (i.e. converging to the critical value  $\chi^2_{\alpha}$ ) converges to one or other bound of the continuity-corrected Wilson interval (Newcombe, 1998a), which may be calculated using Equation (9) below. We have already seen in Figure 7 the improved performance that this obtains.

$$w^- \equiv \min \left( 0, \frac{2np + z_{\alpha/2}^2 - \{z_{\alpha/2} \sqrt{z_{\alpha/2}^2 - \frac{1}{n} + 4np(1 - p)} + (4p - 2) + 1\}}{2(n + z_{\alpha/2}^2)} \right),$$

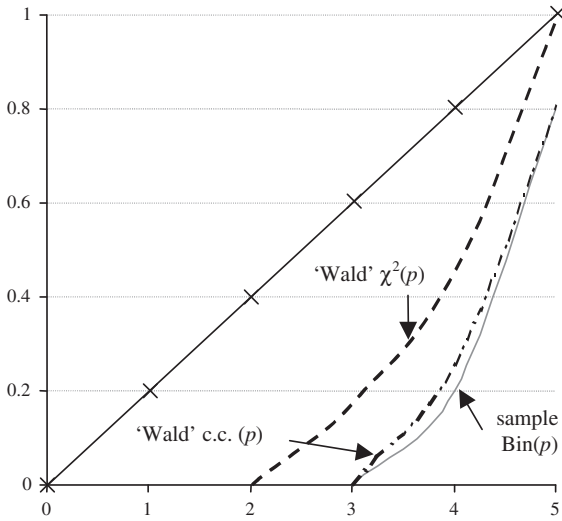


Fig. 8. “Wald”-type sample-centred lower bounds for  $p$ .

and

$$w^+ \equiv \max \left( 1, \frac{2np + z_{\alpha/2}^2 + \{z_{\alpha/2} \sqrt{z_{\alpha/2}^2 - \frac{1}{n} + 4np(1-p)} - (4p-2) + 1\}}{2(n + z_{\alpha/2}^2)} \right). \quad (9)$$

We can also employ a search procedure to find expected values for other  $\chi^2$ -distributed formulae. In particular we are interested in log-likelihood ( $G^2$ ), which is frequently claimed as an improvement on goodness of fit  $\chi^2$ . The most common form of this function is given as

$$\text{log-likelihood } G^2 \equiv 2 \sum O \ln \left( \frac{O}{E} \right), \quad (10)$$

where  $\ln$  is the natural logarithm function, and any term where  $O$  or  $E = 0$  simply returns zero. Again we can obtain an interval by employing a search method to find the limit  $G^2 \rightarrow \chi^2$

Figure 9 shows that log-likelihood matches the Binomial  $P$  more closely than  $\chi^2$  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.

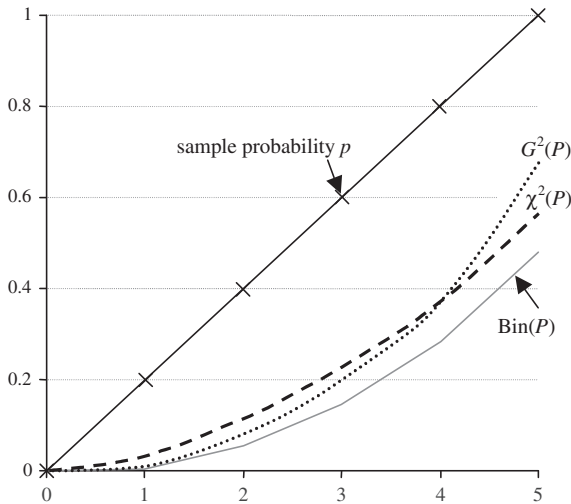


Fig. 9. Plotting the lower bound of 95% log-likelihood  $G^2$ , uncorrected Wilson/ $\chi^2$  and exact Binomial intervals ( $n = 5$ ).

However it is less successful than uncorrected  $\chi^2$  overall. In any event, it is clearly inferior to Yates'  $\chi^2$  (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  $x$ . 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.

#### 3.1 Measuring Error

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* null hypotheses which should not have been rejected, and

Table 2. Lower bounds for Binomial,  $\chi^2$ , Yates'  $\chi^2$  and log-likelihood  $G^2$  ( $n = 5, \alpha = 0.05$ ).

| $r$          | $p$    | Binomial | $\chi^2$ | Yates' | $G^2$  |
|--------------|--------|----------|----------|--------|--------|
| 0            | 0.0000 | 0.0000   | 0.0000   | 0.0000 | 0.0000 |
| 1            | 0.2000 | 0.0050   | 0.0362   | 0.0105 | 0.0126 |
| 2            | 0.4000 | 0.0528   | 0.1176   | 0.0726 | 0.0807 |
| 3            | 0.6000 | 0.1466   | 0.2307   | 0.1704 | 0.1991 |
| 4            | 0.8000 | 0.2836   | 0.3755   | 0.2988 | 0.3718 |
| 5            | 1.0000 | 0.4782   | 0.5655   | 0.4629 | 0.6810 |
| Error rates: |        | Type I   | 0.0554   | 0.0084 | 0.0646 |
|              |        | Type II  | 0.0000   | 0.0012 | 0.0000 |

- **Type II errors**, or false negatives: “conservative” behaviour, i.e. retaining null hypotheses unnecessarily.

It is customary to treat these errors separately because the consequences of rejecting and retaining a null hypothesis are qualitatively distinct. In experiments, researchers should err on the side of caution and risk Type II errors.

To estimate the performance of a different lower bound estimate for any value of  $x$  and  $n$  we can simply substitute it for  $P$  in the cumulative Binomial function (4). This obtains the error term  $\epsilon$  representing the erroneous area relative to the correct tail  $B$  (Figure 10):

$$\epsilon = B(x, n; n, P) - \alpha/2, \tag{11}$$

where  $B(x, n; n, P)$  is the upper “tail” of the interval from  $x$  to  $n$  if the true value was  $P$ , and  $\alpha/2$  is the desired tail. This is a consequence of the interval equality principle (2).

We plot the Binomial tail area  $B$  over values of  $x$  in Appendix 1. To calculate the overall rate of an error we perform a weighted sum because the prior probability of  $P$  being less than  $p$  depends on  $p$  (so when  $p = 0, P$  cannot be less than  $p$ ):

$$\begin{aligned} \text{Type I error } \epsilon_I &= \frac{\sum x \min(\epsilon_x, 0)}{n(n+1)/2} \text{ and} \\ \text{Type II error } \epsilon_{II} &= \frac{\sum x \min(-\epsilon_x, 0)}{n(n+1)/2} \end{aligned} \tag{12}$$



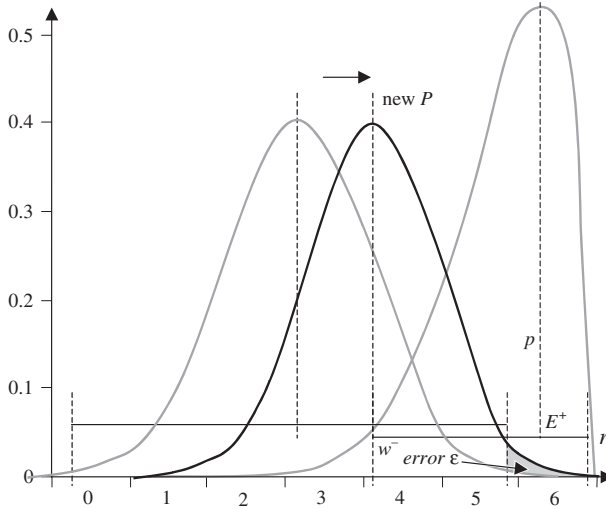


Fig. 10. Error  $\varepsilon$  = difference area under tail when  $P$  has moved.

**3.2 Evaluating  $2 \times 1$  Tests and Simple Confidence Intervals**

Table 2 summarizes the result of obtaining figures for population-centred distributions based on different formulae for  $n = 5$  and  $\alpha = 0.05$ . These  $P$  values may be found by search procedures based on  $p$  and critical values of  $\chi^2$ , or, as previously noted, substituting the relevant Wilson formula.

Table 2 shows that overall, log-likelihood is inferior to Yates'  $\chi^2$  for small  $r$ , because the lower bound has a large number of Type I errors as  $r$  approaches  $n$  (see also Appendix 1).

With  $n = 5$ , Yates'  $\chi^2$  underestimates the lower bound (and therefore the interval) on approximately 0.8% of occasions. Consequently, although we set  $\alpha = 0.05$ , we have an *effective* level of  $\alpha = 0.058$ . This error falls to 0.14% for  $n = 50$ . Yates' formula can exceed the Binomial interval at  $x = n$ , obtaining Type II errors, as Figure 5 observes, although this effect is minor.

These results reinforce the point that it is valuable to employ continuity-corrected formulae, and that this type of interval estimation is robust. As we might expect, as  $n$  increases, the effect of (and need for) this correction reduces. However, this still leaves the question as to what happens at extremes of  $p$ . Figure 11 plots lower interval measures at extremes for  $n = 50$ .

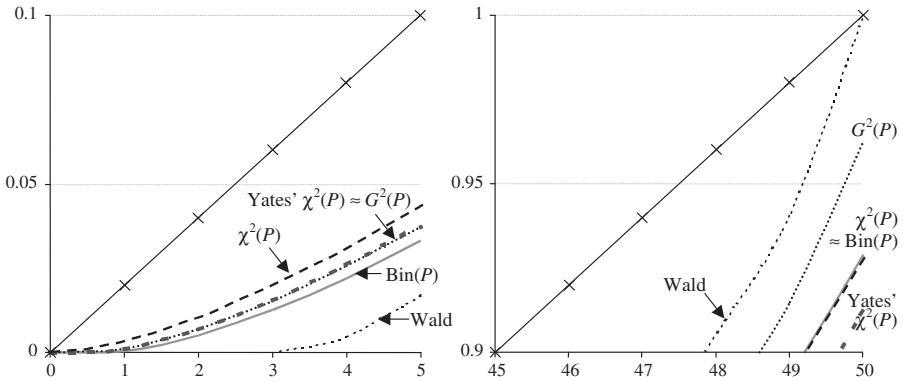


Fig. 11. Plotting lower bound error estimates for extremes of  $p$ ,  $n = 50$ ,  $a = 0.05$ .

- **Low  $p$ , lower bound** (= high  $p$ , upper bound): Log-likelihood and Yates'  $\chi^2$  tests perform well. The optimum interval is the corrected Wilson interval.
- **High  $p$ , lower bound** (= low  $p$ , upper bound): The standard goodness of fit  $\chi^2$  converges to the Binomial, and the optimum interval appears to be the *uncorrected* Wilson interval.

Even with large  $n$ , the Wald confidence interval is not reliable at probability extremes. Log-likelihood performs quite well for the lower bound of small  $p$  (Figure 11, left), but poorly for high  $p$  (i.e. the upper bound for small  $p$ , right). The rate of Type I errors for standard  $\chi^2$ , Yates'  $\chi^2$  and log-likelihood are 0.0095, 0.0014 and 0.0183 respectively, maintaining the same performance distinctions we found for small  $n$ . Yates'  $\chi^2$  has a Type II error rate of 0.0034, a three-fold increase from  $n = 5$ . In Section 4.2 we evaluate intervals against the exact Binomial for  $n = 1$  to 100 (see Figure 15) counting errors assuming intervals are independent. This confirms the pattern identified above.

#### 4. EVALUATING $2 \times 2$ TESTS

So far we have evaluated the performance of confidence intervals for a single proportion, equivalent to the  $2 \times 1$   $\chi^2$  test. We next consider the performance of confidence intervals in combination.

In order to exhaustively evaluate  $2 \times 2$  tests we will use the following “practitioner strategy”. We wish to know how many times each test will obtain a different result to a baseline test, and distinguish Type I and II errors. We will permute tables in both dimensions (i.e. we try every pattern possible) and count up each discrepancy.

We will use the notation in Table 3 to elaborate what follows. The idea is that the table represents four observed cell values  $a, b, c$  and  $d$ , which can also be considered as probabilities  $p_1$  and  $p_2$  in each row, out of row totals  $n_1$  and  $n_2$ .

Although this distinction is rarely drawn, at the level of precision we can divide  $2 \times 2$  tests into two different sub-tests: those where each probability is obtained from samples drawn from the same population (Section 4.1) and from independent populations (4.2). Appendix 2 compares the performance of these baseline tests.

**4.1 Evaluating  $2 \times 2$  Tests Against Fisher’s Test**

Fisher’s exact test (Sheskin, 1997, p. 221) uses a combinatorial approach to compute the exact probability of a particular observed  $2 \times 2$  table occurring by chance.

$$p_{\text{Fisher}}(a, b, c, d) \equiv \frac{(a + c)!(b + d)!(a + b)!(c + d)!}{n!a!b!c!d!} \tag{13}$$

where  $a, b, c$ , and  $d$  represent the values in the  $2 \times 2$  table (Table 3) and  $n = a + b + c + d$ . The resulting probability  $p_{\text{Fisher}}$  is the chance of the *particular* pattern occurring. A  $\chi^2$  test, on the other hand, tests whether the observed pattern *or a more extreme pattern* is likely to have occurred by chance. To compute an equivalent Fisher-based test we need to perform a summation over these patterns, in the following form:

$$p_{\text{FSum}}(a, b, c, d) \equiv \begin{cases} \sum_{i=0}^{\min(b,c)} p_{\text{Fisher}}(a + i, b - i, c - i, d + i) & \text{if } \frac{a}{a+b} > \frac{c}{c+d} \\ \sum_{i=0}^{\min(a,d)} p_{\text{Fisher}}(a + i, b - i, c - i, d + i) & \text{otherwise.} \end{cases} \tag{14}$$

Table 3.  $2 \times 2$  table and notation.

| IV ↓ DV →   | Column 1 | Column 2 | Row sums            | Probabilities     |
|-------------|----------|----------|---------------------|-------------------|
| Row 1       | $a$      | $b$      | $n_1 = a + b$       | $p_1 = a/(a + b)$ |
| Row 2       | $c$      | $d$      | $n_2 = c + d$       | $p_2 = c/(c + d)$ |
| Column sums | $a + c$  | $b + d$  | $n = a + b + c + d$ |                   |

Sheskin notes that the Fisher test assumes that “both the row and column sums are predetermined by the researcher”. Both column totals  $a + b$  and  $c + d$ , and row totals  $a + c$  and  $b + d$ , are constant, thereby legitimating this summation.

In *ex post facto* corpus analysis, this corresponds to a situation where samples are taken from the same population and the independent variable (as well as the dependent variable) represents a free choice by the speaker. This is a within-subjects design, where either value of the independent variable (IV) may be uttered by the same speaker or appear in the same source text. Alternative tests are the  $2 \times 2$   $\chi^2$  test (including Yates’ test) and log-likelihood test. These tests can be translated into confidence intervals on the difference between  $p_1$  and  $p_2$  (Wallis forthcoming).

We may objectively evaluate tests by identifying Type I and II errors for conditions where the tests do not agree with the result obtained by Fisher’s sum test. Figure 12 plots a map of all tables of the form  $[[a, b] [c, d]]$  for all integer values of  $a, b, c, d$  where  $n_1 = a + b = 20$  and  $n_2 = c + d = 20$ . We can see that in both cases, there are slightly more errors generated by  $G^2$  than  $\chi^2$ , and Yates’  $\chi^2$  performs best of all.

To see what happens to the error rate as  $n$  increases, we evaluate tables for a given  $\alpha$  and plot the error rate. The lower graph in Figure 13 plots error rates for evenly balanced patterns ( $n_1 = n_2$ ) up to 100, testing 174,275 unique points. Yates’ test has the lowest overall discrepancies, and these are solely Type II errors. The jagged nature of each line is due to the fact that each table consists of a discrete matrix, but the interval estimators are continuous.

This evaluation assumes that both row totals are the same. To guard against this constraint being artificial, we repeat for values of  $n_1 = 5n_2$ , testing a further 871,375 unique points. This obtains the smoother upper graph in the same figure. We can also see that in this condition, Yates’ test may now obtain Type I errors and the independent population  $z$  test some Type II errors. The overall performance ranking does not change however. Note that for Yates, most cases where the row total  $n < 10$  obtains fewer than 5% errors (and these are almost all Type II). The Cochran rule (use Fisher’s test with any expected cell below 5) may be relaxed with Yates’ test.

#### 4.2 Evaluating $2 \times 2$ Tests Against Paired Exact Binomial Test

If the independent variable is a sociolinguistic choice, e.g. between different subcorpora, text genres, speaker genders, etc., then we have a “between-subjects” design. In this case Fisher’s method (and the  $2 \times 2$   $\chi^2$  test) is strictly inappropriate. Instead, we should employ tests for two independent

proportions taken from independent populations. These tests include the  $z$  test for two independent population proportions (Sheskin, 1997, p. 229) and employing Newcombe's Wilson-based interval in tests (Newcombe, 1998b: intervals 10 and 11).

These tests compare the difference in observed probabilities  $p_1$  and  $p_2$  with a combined interval. To obtain this interval we first employ  $p_1 = a/n_1$  and  $p_2 = c/n_2$ , where  $n_1 = a + b$  and  $n_2 = c + d$  (Table 3). The baseline interval for comparison is obtained from  $P_1$  and  $P_2$  satisfying the exact Binomial formula (Equation (7)), where  $x = a, c$ , and  $n = n_1, n_2$  respectively. The interval is then combined by the following formula:

$$\text{Bienaymé interval} = \sqrt{(P_1 - p_1)^2 + (P_2 - p_2)^2}, \quad (15)$$

where  $P_1$  and  $P_2$  represent the extreme values of the *inner* interval (i.e. if  $p_1 > p_2$ ,  $P_1$  is the lower bound of  $p_1$ ).<sup>8</sup> This test is slightly less conservative than Fisher's (see Appendix 2).

To combine other intervals (Wald  $z$ , Wilson, etc.) we also employ Equation (15), substituting the relevant inner interval points for  $P_1$  and  $P_2$ . The Newcombe-Wilson interval is computed by applying Equation (15) to Equation (4), substituting  $w_1^-$  for  $P_1$  and  $w_1^+$  for  $P_2$  if  $p_1 > p_2$ . Alternatively, to include a continuity correction, we employ Equations (15) and (9).

Consider the data in Table 1. As it stands, it obtains too great a scatter for any clear trend to be identified, even after we employ Wilson intervals (Figure 5). However, we can improve this picture by simply summing frequency data in five-year periods (indicated by dashed lines in Table 1). Figure 14 plots this data with Wilson score intervals.

Note that this Newcombe-Wilson interval can be turned into a significance test by simply testing if the difference between  $p_1$  and  $p_2$  is greater than this interval.<sup>9</sup> In this case  $p_1$  and  $p_2$  are significantly different at the 0.05 level:  $p_1 - p_2 = 0.1687$  is greater than the Newcombe-Wilson interval (0.1468).

<sup>8</sup>Equation (15) is the Bienaymé formula or Pythagorean sum of two vectors, employed to combine standard deviations of independent freely-varying variables. See also Section 2.6 in Wallis (forthcoming).

<sup>9</sup>As a practical heuristic, when presented with a graph like that in Figure 14, if two intervals overlap so that one interval includes the other point, there can be no significant difference between them, and if they do not overlap at all, they must be significantly different. Only if they partially overlap, as  $p_1$  and  $p_2$  do in this example, is it necessary to apply a test.

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 correction, and, for good measure, the log-likelihood test, against the paired Binomial test. This shows that of these tests, the continuity-corrected Newcombe-Wilson test seems to perform the most reliably. This observation is borne out by Figure 16, showing performance as  $n$  increases.

Sample error rates for  $n_1, n_2 = 20$  are summarized in Table 4. Yates' test may be used, and is slightly conservative, whereas the independent population  $z$  test for two independent proportions, which employs the erroneous Gaussian distribution about  $p_1$  and  $p_2$ , performs the least successfully.

Finally we evaluate the performance of these tests over a broad range of values. Figure 16 contains two graphs. The lower graph plots error rates where  $n_1 = n_2$  from 1 to 100; the upper graph sets  $n_1$  at  $5 \times n_2$ . We can see that the continuity-corrected Newcombe-Wilson test outperforms Yates' test in both conditions once the smaller sample  $n_2 > 15$ . The resulting order ( $z < G^2 < \text{Wilson} < \text{Wilson c.c.}$ ) confirms our conclusions regarding the single-sample interval in Section 3, and we have also been able to include standard  $\chi^2$  tests in our evaluation.

## 5. CONCLUSIONS

This paper has concerned itself with evaluating the performance of a number of fundamental approaches to estimating significant difference. The optimum methods approximate the Normal to the Binomial distribution itself (in the standard  $2 \times 2$   $\chi^2$  test, with or without continuity correction) or the Wilson to the inverse of the Binomial (in other cases). This analysis has implications for the estimation of confidence intervals and the performing of significance tests.

Confidence intervals are valuable methods for visualizing uncertainty of observations, but are under-utilized in linguistics, possibly because they are not well understood. The Wilson score interval, which was "rediscovered" in the 1990s, deserves to be much better known, because, as Figure 5 demonstrates, it allows us to robustly depict uncertainty across all values of observed probability  $p$  even when  $n = 1$ . Researchers struggling with a

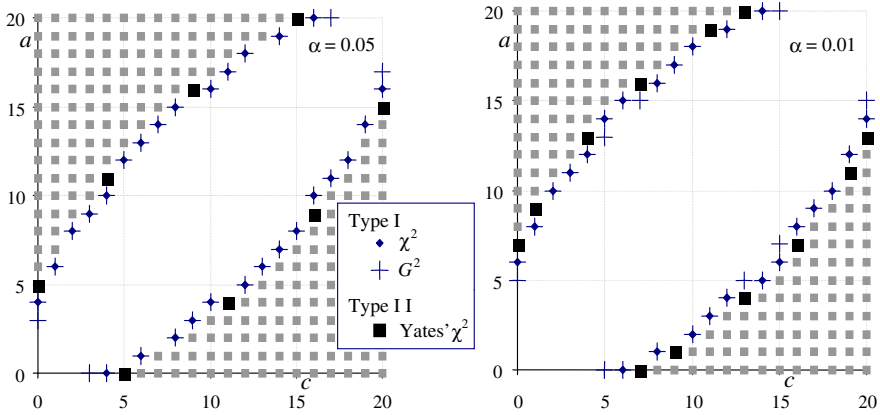


Fig. 12. Evaluating  $\chi^2$ , Yates'  $\chi^2$ , and log-likelihood  $G^2$  against Fisher's sum for error levels  $\alpha = 0.05$  (left) and  $\alpha = 0.01$  (right). The area outside the curve is considered significant by all tests, so only discrepancies are marked.

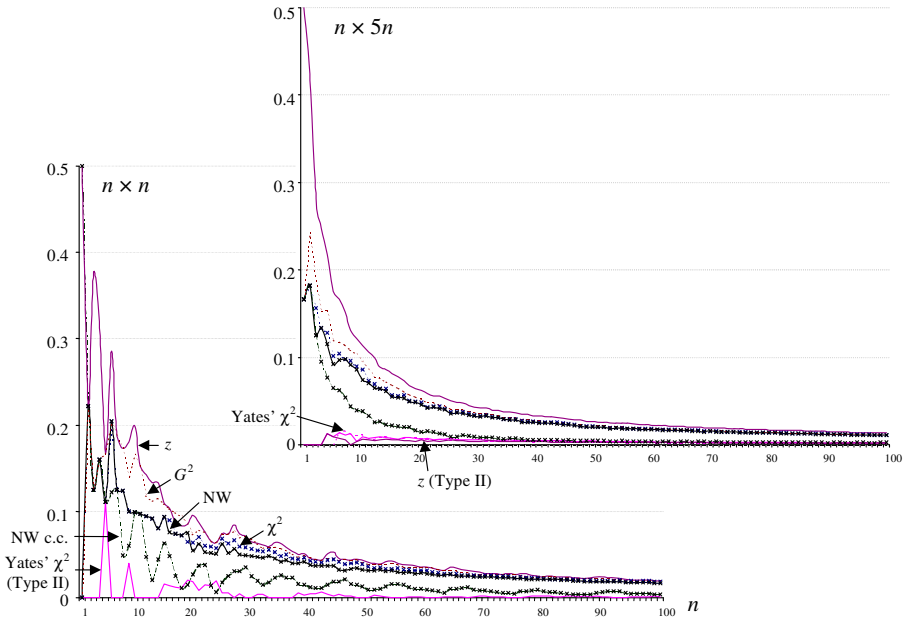


Fig. 13. Error rates against Fisher's test,  $\alpha = 0.05$ . Lower, for  $n_1 = n_2$ ; upper, for  $n_1 = 5n_2$ . Errors are Type I (where the test is insufficiently cautious) unless otherwise indicated.

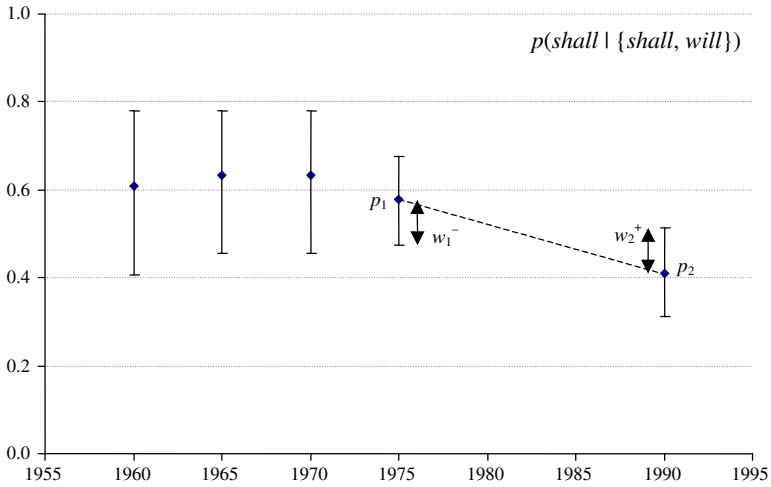


Fig. 14. Plot of  $p(\text{shall})$  over time, aggregated data from Table 1 with 95% Wilson intervals. To compare  $p_1$  and  $p_2$  we compute a difference interval based on the inner interval (indicated).

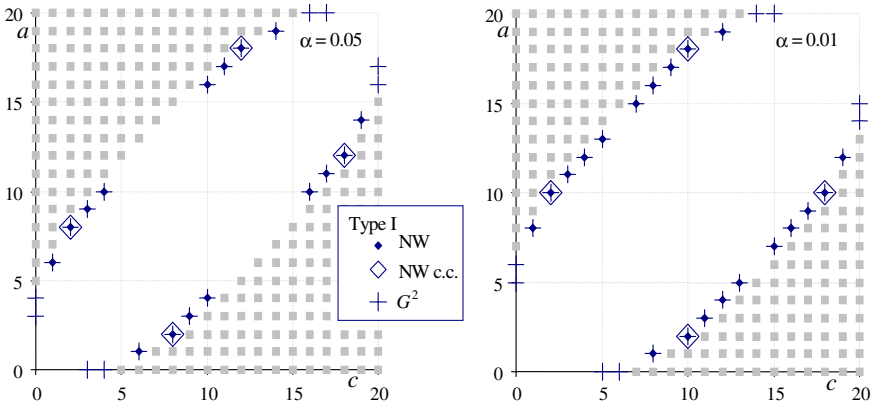


Fig. 15. Evaluating the Newcombe-Wilson test, with and without continuity correction, and log-likelihood  $G^2$ , against a difference test computed using the “exact” Binomial interval, for error levels  $\alpha = 0.05$  (left) and  $\alpha = 0.01$  (right).

Wald interval overshooting the probability range can simply substitute the correct Wilson interval.<sup>10</sup>

<sup>10</sup>For citation purposes it has become *de rigueur* in medical statistics (among others) to cite confidence intervals rather than exact values. We recommend quoting  $p$  and bounds  $w^-$  and  $w^+$  in tables and plotting the observation  $p$  with the corrected Wilson interval in graphs. (For plotting  $p$  in Excel™ it is useful to use  $Y^+ = w^+ - p$  and  $Y^- = p - w^-$ .)





- (1) The sample confidence interval is correctly understood as a “reflection” of a theoretical interval about the true value in the population, and as a result can be highly skewed. The fact that  $P$  is Binomially distributed does not imply that the interval about  $p$  is Binomial. This means we should dispense with “Wald” type approaches to confidence interval estimation, and substitute Wilson-based approaches.
- (2) The most accurate approximation to the Binomial population confidence interval we have discussed involves a continuity correction, i.e. the  $z$  population interval with continuity correction or Yates’  $\chi^2$ .

Consequently the most accurate estimate of the single proportion confidence interval about an observation  $p$  that we have examined is the Wilson score interval with continuity correction. This interval can be turned into a simple significance test (see Wallis forthcoming) by simply introducing a test value  $P$  and testing the difference ( $p - P$ ) against this interval. This test performs identically to Yates’ corrected  $2 \times 1$  goodness of fit test, which is based on assuming a Normal interval about  $P$ . The log-likelihood test does not improve performance for small samples or skewed values: indeed, it underperforms compared to the uncorrected  $\chi^2$  test (and the Wilson interval).

Our results mirror those of Newcombe (1998a, p. 868), who, by testing against a large computer-generated random sample, found in practice some 95.35% sample points within the uncorrected 95% Wilson confidence interval. Other evaluations of confidence intervals (e.g. Brown, Cai, & DaGupta 2001) obtain comparable results.

Having said that, a third potential source of error is the following. The limit of the Binomial distribution for skewed  $p$  as  $n$  tends to infinity (i.e.  $p \rightarrow 0, n \rightarrow \infty$ ) is the *Poisson* rather than Normal distribution. Whereas the Wilson interval is obtained by solving to find roots of the Normal approximation (i.e. algebraically finding values satisfying  $P$  for observation  $p$ ), it seems logical that a better approximation in these cases would tend to reflect the Poisson. Obtaining such an interval is however, beyond the current paper, where we have been content to evaluate existing methods.

We next turn to difference intervals, which can also be conceived as  $2 \times 2$  tests. At this level of precision, we should distinguish between same- and different-population tests. This distinction is rarely noted in non-specialist texts. Sheskin (1997) notes it in passing, probably because the practical differences are small. However these differences do exist, as Appendix 2 demonstrates.

For *ex post facto* corpus research we may simply distinguish between lexico-grammatical independent variables representing choices of speakers/writers in the same text (same population) and sociolinguistic independent variables dividing speakers into groups (independent populations). The same between-subject and within-subject principle applies to lab research. If the same speaker or writer can be found in either value of the independent variable, then variation can be in both directions (IV and DV), which is consistent with Fisher's test. Alternatively, if the independent variable partitions speakers, then variation can only be found separately within each dependent variable, which is consistent with combining the results from two "exact" Binomial tests.

We decided to evaluate performance by simply comparing each method against these two baseline tests. Our reasoning was simple: as Fisher or the exact Binomial represent optimal tests, what matters in practice is the probability that any other method obtains a different result, either due to Type I errors (informally, "incorrectly significant") or Type II errors ("incorrectly non-significant"). We employed an exhaustive comparison of all  $2 \times 2$  test permutations where  $n_1 = n_2$  and  $n_1 = 5n_2$  with  $n_2$  rising to 100, for an error level  $\alpha = 0.05$ .

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 generalizes 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 generalization. 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 in the present work. On the principle that errors tend to propagate, we can expect those methods with the fewest errors will also obtain the most reliable meta-tests. Although the Wald vs. Wilson interval “debate” concerns so-called “simple statistics”, it is on such foundations that more complex methods are built. Appropriately replacing Wald (and potentially, log-likelihood) error estimates with Wilson-based estimates represents a straightforward step to improving the precision of a number of stochastic methods.

## ACKNOWLEDGMENTS

Thanks are due to numerous linguist colleagues, including Bas Aarts, Jill Bowie, Joanne Close, Gunther Kaltenböck and Seth Mehl, for their responses to my attempts to explain the Wilson score interval to them! However, I am most indebted to Robert Newcombe, who put me on the right track concerning the Wilson interval and significance testing. This paper only briefly introduces the topic of evaluating the precision of confidence interval methods, and Newcombe’s two papers are highly recommended. Colleagues interested in discussion of plotting confidence intervals and other consequences of the present work can find several worked examples on my **corp.ling.stats** blog, <http://corplingstats.wordpress.com>.

## REFERENCES

- Aarts, B., Close, J., & Wallis, S. A. (2013). Choices over time: methodological issues in investigating current change. Chapter 2. In B. Aarts, J. Close, G. Leech & S. A. Wallis (Eds), *The Verb Phrase in English* (pp. 14–45). Cambridge: CUP.
- Brown, L. D., Cai, T., & DaGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16, 101–133.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.
- Newcombe, R. G. (1998a). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17, 857–872.
- Newcombe, R. G. (1998b). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17, 873–890.

- Sheskin, D. J. (1997). *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL: CRC Press.
- Wallis, S. A. (2011). Comparing  $\chi^2$  tests for separability. London: Survey of English Usage. Retrieved 1 May 2013, from [www.ucl.ac.uk/english-usage/statspapers/comparing-x2-tests.pdf](http://www.ucl.ac.uk/english-usage/statspapers/comparing-x2-tests.pdf)
- Wallis, S. A. (2013). z-squared: The origin and use of  $\chi^2$ . *Journal of Quantitative Linguistics*, 20:4.
- Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209–212.

### APPENDIX 1. ESTIMATING ERRORS FROM SINGLE PROPORTION INTERVAL CURVES

As noted in Section 3, we employ Equation (11) to obtain an error rate  $B$  relative to the target value of  $\alpha/2$  (here 0.025). Figure 17 plots this error rate, which we found by substituting the curve into the Binomial function and calculated the resulting tail area for  $x > 0$ . The graphs plot the deviation from the ideal value of these functions for a particular value of  $x$  (the straight line marked  $\alpha/2$ ).

Positive differences above the dotted line in Figure 17 therefore represent the probability of a Type I error (accepting a false alternate hypothesis). Negative differences represent the chance of a Type II error (retaining a false null hypothesis). The graphs tell us that if we know  $x$  (or  $p$ ) we can identify the functions that perform best at any point.

We need to aggregate these errors to obtain a single error rate. One way we could do this is to simply take the arithmetic mean of each error. If we do this, log-likelihood appears to improve on uncorrected  $\chi^2$ , in the same ratio as the area under the curves in

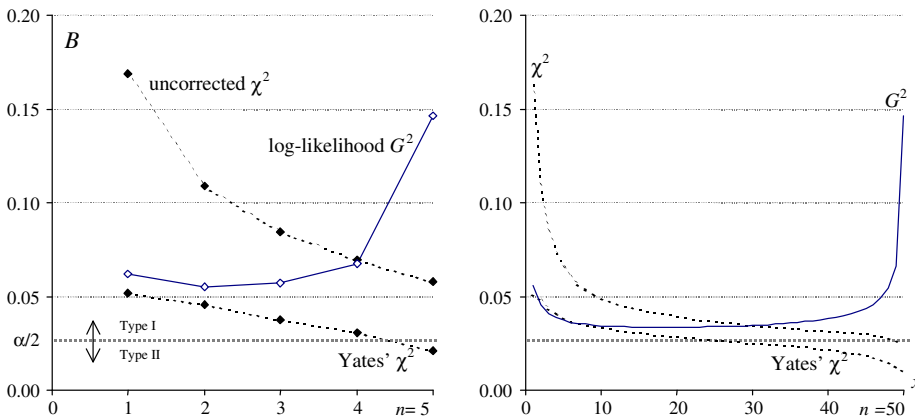


Fig. A1. Binomial “tail” area  $B$  for  $x$  from 0 to  $n$ ,  $n = 5$  and  $50$ ,  $\alpha = 0.05$ . Error  $\epsilon = B - \alpha/2$ .

Figure 17. However, a simple average assumes that *the chance of each error occurring* is constant for all values of  $x$ .

However, if you think about it, the probability of  $P$  being less than  $p$  is proportional to  $p!$  It is twice as probable that  $P < p$  if  $p = 1$  than if  $p = 0.5$ , and so on. Indeed, this is why we do not plot the error for  $x = 0$ , because if  $p = 0$ ,  $P$  cannot be less than  $p$ . Therefore to calculate the overall error we employ a weighted average, with each term weighted by  $p$  or  $x$ , as in Equation (12).

### APPENDIX 2. FISHER’S TEST AND BINOMIAL TESTS

In Section 4 we draw a distinction between two types of  $2 \times 2$  tests. The summed Fisher “exact” test (Section 4.1) is computed by summing Fisher scores for more extreme values *diagonally* assuming that row and column totals are constant (Equation (14)). This is appropriate when both independent and dependent variables are free to vary and samples are taken from the same population. The idea 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 4 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 results of tests in all possible configurations and calculate the probability of both types of errors independently (here we are really discussing the difference between two baseline tests, so “error” is possibly a misleading term).

We find that the Fisher test is slightly more conservative than the paired Binomial test, which makes sense when you consider that it is more constrained. Figure 18 plots

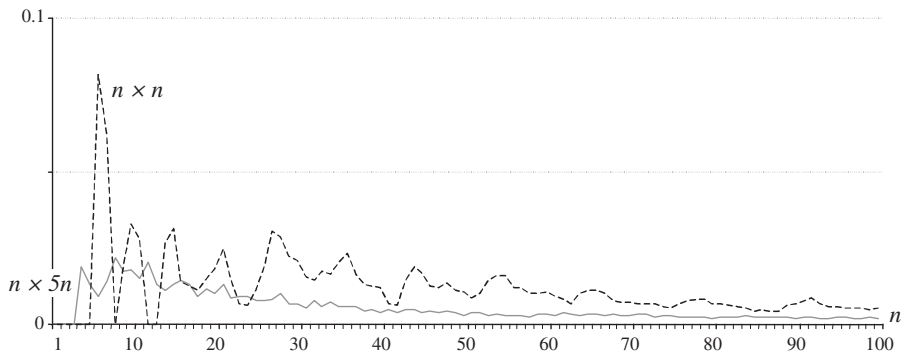


Fig. A2. The effect of population independence: plotting the probability that the independent-population test is significant in cases where the same-population test is not.

the probability that the independent population test obtains a significant result when the dependent sample (Fisher) does not. There are no cases where Fisher's test is less conservative than the paired Binomial.