

# Binomial distributions, probability and Wilson's confidence interval

Sean Wallis, Survey of English Usage, University College London  
December 20 2009 (revised 2012)

This short paper concerns an important problem in a core approach to a number of statistical methods and tests. These include calculating confidence intervals (or 'error bars'),  $\chi^2$  tests (Sheskin, 1997), log-likelihood tests, and certain types of line fitting. The idea that all of these methods have in common is that with sufficient data, we can estimate a confidence interval using the Normal  $z$  distribution, and thereby construct simpler tests, or apply mathematical simplifications, than would otherwise be possible. (This is such a common assumption it is usually referred to as "The Central Limit Theorem".)

However this statement leaves open the question as to how much data is 'sufficient', and whether this question is affected by the degree to which the distribution is off-centre.

## 1. Conventional estimates of error

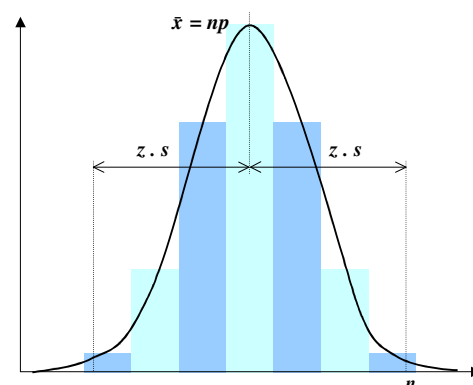
Estimating the error in an observation is a crucial step in inferential statistics. It allows us to make predictions about what would happen were we to repeat our experiment any number of times, and because each observation represents a sample of the population, predict the true value in the population (Wallis 2010).

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

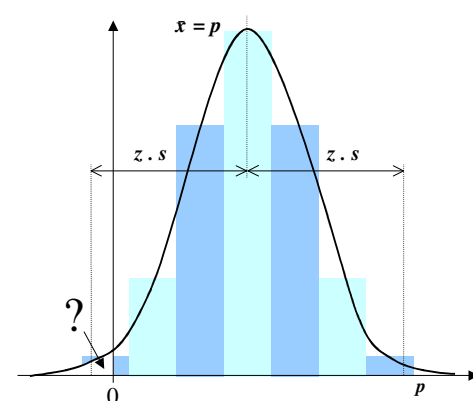
- the proportion  $p$  of coin tosses in a set of  $n$  throws that are heads,
- the proportion of light bulbs  $p$  in a production run of  $n$  bulbs that fail within a year,
- the proportion of patients  $p$  who have a second heart attack within six months after a drug trial has started ( $n$  being 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 don't want to repeat a drug trial on cardiac patients if the drug may be adversely affecting their survival.<sup>1</sup>

To do this 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. This is immensely important.



**Figure 1.** Binomial approximation to the Normal distribution plotted over a linear range  $n \in (-\infty, \infty)$ .



**Figure 2.** As above but on a finite probabilistic range  $p \in [0, 1]$ . The assumption that the distribution is approximately Normal may be misleading. In particular what happens if the curve crosses 0 or 1?

<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.

By far the most common ('Wald') approach is to employ the Binomial approximation to the Normal (Gaussian) shown in Figure 1, with the following definitions.

$$\begin{aligned} \text{mean } x &\equiv np, \\ \text{standard deviation } s &\equiv \sqrt{np(1-p)}, \\ \text{confidence interval} &\equiv (np - z.s, np + z.s), \end{aligned}$$

where  $n$  represents the sample size,  $p$  the proportion of the sample in a particular class and  $z$  is the critical value of the Normal distribution for a given error level. If the error level  $\alpha$  is 0.05, 95% of the expected distribution is within the interval, and only 2.5% in each of the 'tails' outside.

The actual distribution, shown by the columns, is assumed to be a discrete Binomial distribution, but to obtain this interval we approximate it to a continuous Normal curve, shown by the line. The larger the value of  $n$  the more 'continuous' the line, and the more confident we can be in  $p$ . With lots of data the size of the confidence interval relative to the range will decrease. But what happens if  $n$  is small?

In our case it is probably more useful to standardize margins of error by converting the range  $[0, n]$  to a probabilistic range  $[0, 1]$ . To do this we simply divide the formulae above by  $n$ .

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

This exposes an obvious problem however. 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  $p$  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 attain (Figure 2), so the approximation **fails**. (This is known as a 'floor' or 'ceiling' effect.)

Some authors employ the limit  $p \pm 3s \in [0, 1]$  before using the Binomial approximation to the Normal. This means that we simply give up estimating the error for low or high  $p$  values or for small  $n$ , a situation that is not exactly satisfactory!

A similar heuristic for the  $\chi^2$  test 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 in a series of attempts to cope with low frequencies and skewed datasets. In this paper we demonstrate two distinct problems with this 'Wald' method for single sample intervals.

## 2. Confidence intervals on population probability

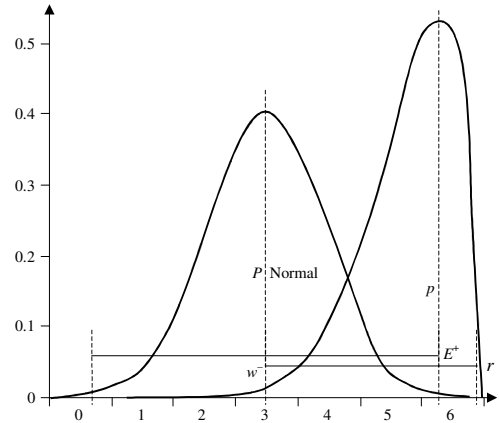
The problem with the conventional definition given above is that *the confidence interval is incorrectly characterised*. Note that we assumed in the above that the interval about  $p$  was Binomial and therefore this could be approximated to the Normal distribution. This is the wrong way to think about the problem.

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 *actual value* 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  $P$ . The idea is illustrated by Figure 3.

What we need to do therefore is **search** for values of  $P$  which satisfy the formula<sup>2</sup> used to characterise the Binomial distribution about  $P$ . Now we have the following definitions:

$$\begin{aligned} \text{pop. mean } \mu &\equiv P, \\ \text{pop. standard deviation } \sigma &\equiv \sqrt{P(1-P)/n}, \\ \text{pop. confidence interval } (E^-, E^+) &\equiv (P - z.\sigma, P + z.\sigma). \end{aligned}$$



**Figure 3.** The interval equality principle with Normal and Wilson intervals: the lower bound for  $p$  is  $P$ .

The symbols  $\mu$  and  $\sigma$ , referring to the **population** mean and standard deviations respectively, are commonly used. This population confidence interval identifies two limit cases where  $p = P \pm z.\sigma$ .

Consider now the confidence interval around the sample observation  $p$ . We don't know  $P$  in the above and we can't 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  $\alpha$ :

$$\begin{aligned} e^- = P_1 &\leftrightarrow E_1^+ = p \text{ where } P_1 < p, \text{ and} \\ e^+ = P_2 &\leftrightarrow E_2^- = p \text{ where } P_2 > p. \end{aligned} \quad (2)$$

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  $p = E_1^+ = P_1 + z.\sigma_1$  and  $p = E_2^- = P_2 - z.\sigma_2$ . With a computer we can perform a search process which iteratively focuses on the correct value.

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  $P = E/n$ .

It turns out that there is a simpler method for directly calculating the sample interval about  $p$  than this computational search process. This interval is called the *Wilson score interval* (Wilson, 1927) and may be written as

$$\text{Wilson score interval } (w^-, w^+) \equiv \left( p + \frac{z^2}{2n} \pm z \sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}} \right) / \left( 1 + \frac{z^2}{n} \right). \quad (3)$$

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

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

2) a corrected *standard deviation*  $s' = \sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}} / \left( 1 + \frac{z^2}{n} \right)$

<sup>2</sup> There are a number of possible formulae, but we will come to this later.

such that  $w^- = p' - z.s'$  and  $w^+ = p' + z.s'$ .

The  $\chi^2$  test checks for the sample probability falling within Gaussian intervals on the population distribution,  $E^- < p < E^+$ . 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 shown diagrammatically in Figure 3.

Employing the Wilson interval on the sample probability does not itself improve on the  $2 \times 1$  goodness of fit  $\chi^2$  test. The improvement is in estimating the confidence interval around  $p$  on either side.

How do these calculations differ in practice? Well,  $p'$  is pushed towards the centre of the distribution. The total width of the interval is twice  $z.s'$  (i.e. proportional to  $s'$ ).

The graphs in Figure 4 plot the standard deviation for Normal and Wilson expressions for different  $p$  values. Note that  $s'$  never reaches zero for low or high  $p$ , and the differences between curves diminishes with increasing  $n$ .

Newcombe (1998) 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.

Figure 4 illustrates why Wilson-based tests **outperform** Gaussian-based ones (e.g. those implied by standard chi-square or log-likelihood tests). Note the middle of the  $p$  curve where the standard deviation for the Wilson interval falls below the Gaussian, indicating that even for non-skewed tests the Gaussian is more conservative. This is in addition to the observation that Wilson-based tests do not fail for skewed data sets (i.e. where  $p$  tends to zero).

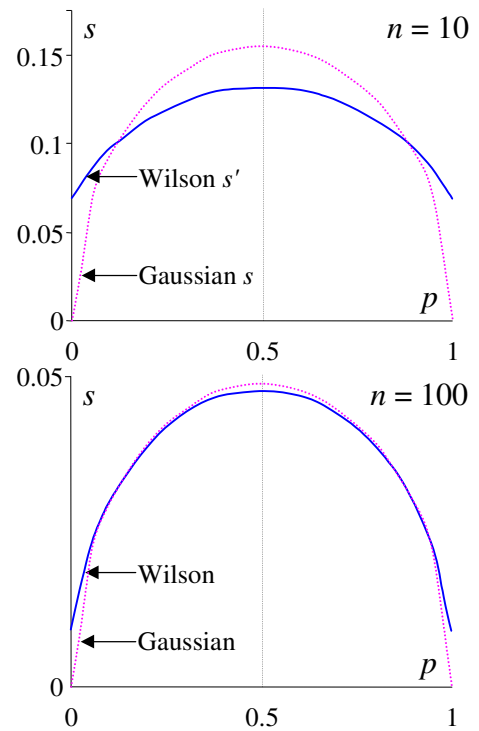
### 3. Exact Binomial intervals

So far we have employed the Binomial *approximation* to the Normal distribution. However this is still inaccurate for small samples. In order to evaluate formulae against an ideal distribution we need to find a way to calculate  $P$  values using the exact 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 cases illustrated by Figure 3 where  $P < p$ . There are two interval boundaries on each probability, but the argument is symmetric (i.e. we could apply exactly the same calculation substituting  $q = 1 - p$ , etc. in what follows). We have already noted that the upper interval for the goodness of fit  $\chi^2$  for an expected probability  $P$  is the lower bound ( $w^-$ ) of the Wilson score interval for  $p$  at an error level  $\alpha = 0.05$ .

Imagine 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 population Binomial



**Figure 4.** Gaussian and Wilson standard deviations  $s, s'$  for  $p = [0, 1]$  and  $n = 10$  and  $n = 100$ .

distribution is defined as a series of discrete probabilities for  $r$ , where the height of each column is defined by the following expression (Sheskin, 1997: 115):

$$\text{Binomial probability } B(r; n, P) = nCr \cdot P^r (1 - P)^{(n-r)}, \quad (4)$$

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>3</sup>, and the probability of each single pattern of  $r$  heads and  $(n - r)$  tails, based on the probability of a head being  $P$ . We may then obtain a cumulative probability by summing over a range  $x_1$  and  $x_2$  inclusive:

$$\text{Cumulative Binomial prob. } B(x_1, x_2; n, P) = \sum_{r=x_1}^{x_2} B(r; n, P) = \sum_{r=x_1}^{x_2} nCr \cdot P^r (1 - P)^{(n-r)}.$$

Again, note that this formula is defined for  $P$ . To find an exact upper bound for  $p = x / n$  we need to find  $P$ , employing a search procedure to find where the following holds:

$$B(x, n; n, P) = \alpha/2. \quad (5)$$

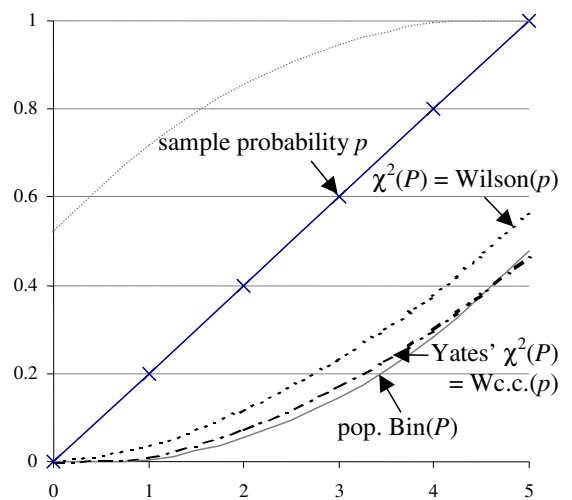
This obtains an exact result for any integer  $x$ :  $P$  can be modified until the formula converges on the single error ‘tail’  $\alpha/2$ . We then report  $P$ .<sup>4</sup>

The 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 before, *we do not know* the true population value  $P$  but we expect that data would be Binomially distributed around it.

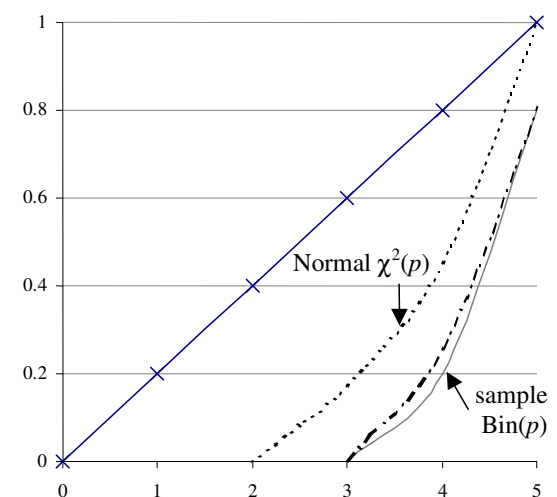
Figure 5 shows the result of computing the lower bound for  $p = P$ , employing this Binomial formula. We also plot the Wilson formula, with and without an adjustment termed the ‘continuity correction’. As we have noted, these formulae are equivalent to the standard goodness of fit  $\chi^2$  and Yates’ corrected  $\chi^2$  respectively.

Figure 5 shows that all three distributions obtain lower confidence intervals on  $p$  which tend towards zero at  $x = 0$ , but do not converge to zero at  $x = n$ . The dotted line at the top of Figure 5 is the upper bound for the exact population Binomial interval, which flips this around. At the extremes are highly skewed intervals, as we expected.

Were further evidence required, we can see the effect of the error of mischaracterising the interval about  $p$  in Figure 6. This effectively swaps the axes of  $n$  and  $p$ . The intervals tend towards zero at  $x = n$  but are very



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



**Figure 6.** Erroneous sample-centred lower bounds for  $p$ .  $\chi^2(p) = \text{Normal}(1)$ .

<sup>3</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>4</sup> This method is Newcombe (1998)’s method 5 (‘Clopper Pearson’) using exact Binomial tail areas. In Figure 6 we estimate the interval for the mean  $p$  by summing  $B(0, r; n, p) < \alpha/2$ .

large (and become negative) for small  $x$ . The Binomial ‘curve’ for  $p$  here is discrete – it consists of rationals  $r/n$  – and conservative, because the sum is *less* than  $\alpha/2$  rather than exactly equal to it.

#### 4. Continuity correction and log-likelihood

We have addressed the major 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 standard goodness of fit  $\chi^2$  test underestimates the error compared to the Binomial interval.

We can predict, therefore, that the uncorrected chi-square test may find some results ‘significant’ which may not be deemed significant if the exact Binomial test was performed. The area between the two curves represents this tendency to make so-called ‘Type 1’ errors – where results are stated as significant when the evidence does not support this.

We can now proceed to evaluate 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 8), 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:

$$\text{Yates' } \chi^2 = \sum \frac{(O - E - 0.5)^2}{E}, \quad (6)$$

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 converges to the continuity corrected version of the Wilson interval calculated using (7) below. We have already seen in Figure 5 the improved performance that this obtains.

$$w^- \equiv \frac{2np + z^2 - \{z\sqrt{z^2 - \frac{1}{n} + 4np(1-p)} + (4p-2) + 1\}}{2(n + z^2)} \text{ if } p > 0, \text{ otherwise } 0, \text{ and}$$

$$w^+ \equiv \frac{2np + z^2 + \{z\sqrt{z^2 - \frac{1}{n} + 4np(1-p)} - (4p-2) + 1\}}{2(n + z^2)} \text{ if } p < 1, \text{ otherwise } 1. \quad (7)$$

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 as we have noted is often claimed to be an improvement on goodness of fit  $\chi^2$ . The most common form of this function is given as

$$\text{log-likelihood } G^2 = 2 \sum O \ln \left( \frac{O}{E} \right), \quad (8)$$

where  $\ln$  is the natural logarithm function, and any term where  $O$  or  $E = 0$  simply returns zero.

#### 5. Evaluating overall performance

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 following error term  $e$  defined as shown in Figure 7:

$$e = B(x, n; n, P) - \alpha/2, \quad (9)$$

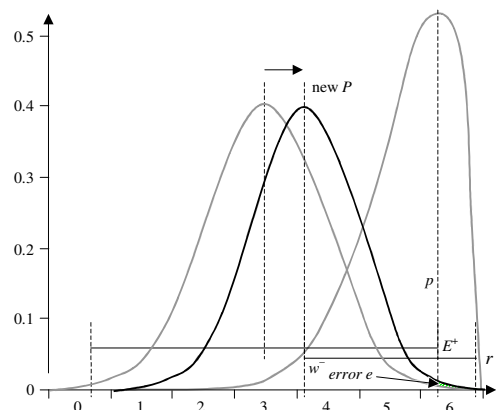
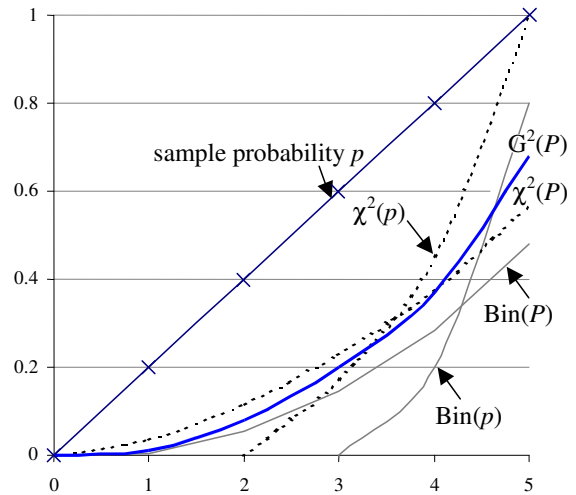


Figure 7. Error  $e$  = difference area under tail when  $P$  has moved.

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).

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

- **Type I errors**, or false positives: this is so-called ‘conservative’ behaviour, i.e. *retaining* null hypotheses unnecessarily, and
- **Type II errors**, or false negatives: ‘anti-conservative’ behaviour, i.e. *rejecting* null hypotheses unnecessarily.



**Figure 8.** Log-likelihood vs. Binomial: a compromise between intervals on  $p$  and  $P$ ?

It is customary to treat these errors separately because the consequences of rejecting a null hypothesis and retaining a null hypothesis unnecessarily are different. We include graphs of the Binomial area in Appendix 1. To calculate the overall rate of an error we perform a weighted sum because the probability of  $P$  being less than  $p$  depends on  $p$  (note that when  $p = 0$ ,  $P$  cannot be less than  $p$ ):

$$\text{Type I error } e_I = \frac{\sum x \min(e_x, 0)}{n(n+1)/2} \text{ and Type II error } e_{II} = \frac{\sum x \min(-e_x, 0)}{n(n+1)/2}. \quad (10)$$

Table 1 summarises 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, by substituting the relevant Wilson formula.

$r$	$p$	Binomial $P$	$P(\chi^2)$	$P(\text{Yates' } \chi^2)$	$P(G^2)$
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
<b>Error rates:</b>		Type I	0.0554	0.0084	0.0646
		Type II	0.0000	0.0012	0.0000

Table 1: Population  $P$  for Binomial,  $\chi^2$ , Yates'  $\chi^2$  and log-likelihood  $G^2$  ( $n=5$ ), obtained by search.

Table 1 shows that, compared with the exact Binomial figure, log-likelihood underestimates the error, although it improves on uncorrected  $\chi^2$ . This remains true as  $n$  increases. Log-likelihood performs less well than uncorrected  $\chi^2$  for small  $r$ , because the lower bound has a large number of Type I errors as  $r$  approaches  $n$ , as Figure 8 indicates (see also Appendix 1).

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

Clearly it is valuable to employ continuity-corrected formulae, and this type of interval estimation is robust. As we might expect, as  $n$  increases, the effect of (and need for) this correction reduces.

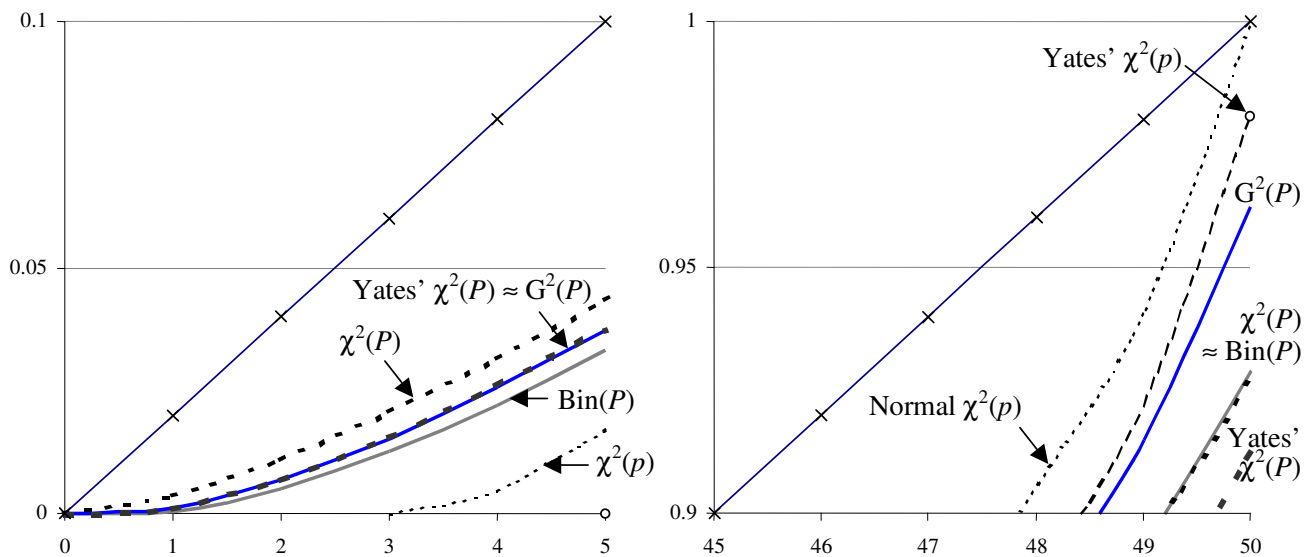


Figure 9. Plotting lower bound error estimates for extremes of  $p$ ,  $n = 50$ ,  $\alpha = 0.05$ .

## 6. High skew with larger samples

So far we have concentrated on the performance of these measures for small samples, e.g. with  $n = 5$ . For relatively unskewed data as  $n$  increases, all measures tend to the Gaussian about  $p$ . As we saw in Section 1, however, this still leaves the question as to what happens at extremes of  $p$ . Since linguists, like medical researchers, are often interested in changes in low frequency events, this is not an unimportant question!

Consider Figure 9 which plots lower interval measures at extremes for  $n = 50$ .

- **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.

We observe that even with quite large  $n$ , the assumption that the correct confidence interval is Normal (the ' $\chi^2(p)$ ' curve) is not supportable at probability extremes.

Log-likelihood performs quite well for the lower bound of small  $p$  (Figure 9, 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  $n = 5$ .

## 7. Conclusions

We have demonstrated that the assumption that the confidence interval around a sample observation is Normal is both **incorrect** and **inaccurate**.

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 mean that the interval about  $p$  is Binomial.
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 sample confidence interval we have examined is the Wilson interval with continuity correction. The log-likelihood test does not improve performance for small samples or skewed values, indeed it underperforms compared to the uncorrected  $\chi^2$  test (Wilson score interval).

We have demonstrated that, as Sheskin (1997) puts it, employing a continuity correction 'obtains a close estimate of the exact Binomial probability'. Our results are similar to those of Newcombe (1998: 868), who, by testing against a large computer-generated random sample, found in practice 95.35% sample points within the uncorrected 95% Wilson confidence interval.

It has become *de rigueur* in medical statistics (for example) to cite confidence intervals rather than exact values. I recommend quoting  $p$ ,  $w^-$  and  $w^+$  in tables and plotting the observation  $p$  with the corrected Wilson interval in graphs.<sup>5</sup>

The Wilson method described here allows a  $z$  test for a single sample to be carried out and obtain the same result as the goodness of fit  $\chi^2$  test (= the  $z$  test for the population proportion). Wallis (2010) introduces Newcombe's extension of the Wilson formula to compare two proportions. This allows us to create a test equivalent to the  $2 \times 2$   $\chi^2$  test which is robust for low frequency, highly skewed data.

Another potential application of Wilson's interval is in line fitting.

A standard optimum method is to fit a formula  $f(x)$  to data points  $x(x)$  using the method of least variance. This means finding a formula  $f(x)$  computationally, where the error  $\Sigma(f(x) - x(x))^2/s(x)^2$  is at a minimum. A variety of functions, such as linear or power functions, may be substituted into  $f(x)$  and function constants varied until a good fit is found.

Core to this method is the assumption that data about  $x$  is Normally distributed. As we have seen, the Wilson interval is skewed, and therefore the error on either side of the mean will differ. However Newcombe points out that if we employ a *logit scale*, i.e. transform  $p$  values to  $\text{logit}(p)$ , where  $\text{logit}(p) \equiv \log(p / 1 - p)$ , then, provided  $p \in (0, 1)$ ,  $\text{logit}(p) - \text{logit}(w^-) = \text{logit}(w^+) - \text{logit}(p)$ .

Since this holds for any critical threshold  $z$  or sample size  $n$ , the distribution about  $\text{logit}(p)$  must be symmetric. We could then employ the following estimators.

$$\begin{aligned} \text{mean } x &\equiv \text{logit}(p), \text{ and} \\ \text{standard deviation } s &\approx (\text{logit}(p) - \text{logit}(w^-)) / z. \end{aligned}$$

This implies an improved method for fitting by least variance minimising  $\Sigma(\text{logit}(f(x)) - \text{logit}(p(x)))^2 / (\text{logit}(p(x)) - \text{logit}(w^-(x)))^2$ . This seems to provide a good estimator although it tends to underestimate the error for small  $n$  or extreme  $p$ .

## References

- NEWCOMBE, R.G. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17**: 857-872.
- SHESKIN, D.J. 1997. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, Fl: CRC Press.

---

<sup>5</sup> For plotting  $p$  it can be useful to obtain corrected y-axis error margins  $Y^+$  and  $Y^-$  as follows:  $Y^- = z.s' + (p - p')$ , and  $Y^+ = z.s' - (p - p')$ .

WALLIS, S.A. 2010. *z*-squared: The origin and use of  $\chi^2$ . London: Survey of English Usage, UCL.  
[www.ucl.ac.uk/english-usage/staff/sean/resources/z-squared.pdf](http://www.ucl.ac.uk/english-usage/staff/sean/resources/z-squared.pdf)

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

### Appendix 1. Error curves

As noted in Section 5 we employ equation (9) to obtain an error rate relative to the target value of  $\alpha/2$  (here 0.025). Figure A1 plots the Binomial area for  $x > 0$ . The graphs plot the deviation from the ideal value of these functions for a particular value of  $x$ .

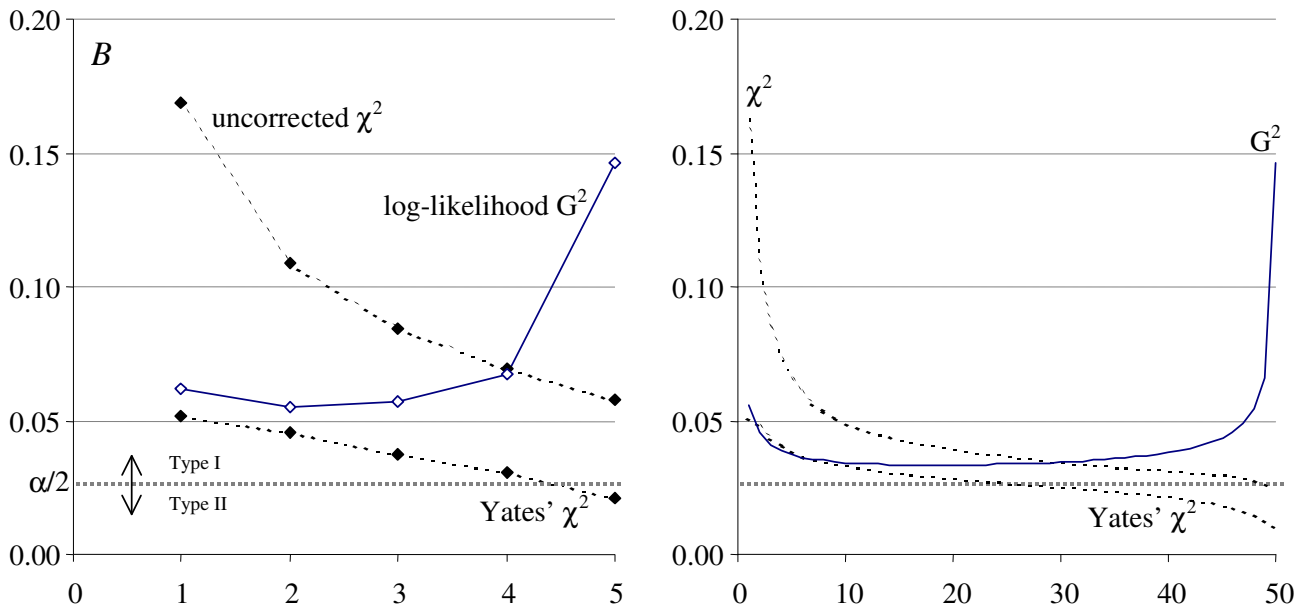


Figure A1. Binomial 'tail' area  $B$  for values of  $x$ : left,  $n = 5$ , right  $n = 50$ ;  $\alpha = 0.05$ .

Positive differences above the dotted line in Figure A1 represent the probability of a Type I error (accepting a false alternate hypothesis). Negative differences represent chances 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.

In averaging these errors it may appear that we should simply take the arithmetic mean of each error. If we do this log-likelihood improves on uncorrected  $G^2$ , in the same ratio as the area under the curves in Figure A1. Using a simple average to assess performance assumes that the chance of each error occurring is equal.

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 properly we need to calculate a **weighted average**, with each term weighted by  $p$  or  $x$ , which is what equation (10) does.