Plotting the Wilson distribution

Sean Wallis, Survey of English Usage, University College London
September 2018[1]

1. Introduction

We have discussed the Wilson score interval at length elsewhere (Wallis 2013a, b). Given an observed Binomial proportion $p = f / n$ observations, and confidence level 1-α, the interval represents the two-tailed range of values where P, the true proportion in the population, is likely to be found. Note that f and n are integers, so whereas P is a probability, p is a proper fraction (a rational number).

The interval provides a robust method (Newcombe 1998, Wallis 2013a) for directly estimating confidence intervals on these simple observations. It can take a correction for continuity in circumstances where it is desired to perform a more conservative test and err on the side of caution. We have also shown how it can be employed in logistic regression (Wallis 2015).

The point of this paper is to explore methods for computing *Wilson distributions*, i.e. the analogue of the Normal distribution for this interval. There are at least two good reasons why we might wish to do this.

The first is to shed insight onto the performance of the generating function (formula), interval and distribution itself. Plotting an interval means selecting a single error level α, whereas visualising the distribution allows us to see how the function performs over the range of possible values for α, for different values of p and n.

A second good reason is to counteract the tendency, common in too many presentations of statistics, to present the Gaussian ('Normal') distribution as if it were some kind of 'universal law of data', a mistaken corollary of the Central Limit Theorem. This is particularly unwise in the case of observations of Binomial proportions, which are strictly bounded at 0 and 1.

As we shall see, the Wilson distribution diverges from the Gaussian most dramatically as it tends towards the boundaries of the probabilistic range, i.e. where the interval approaches 0 or 1. By contrast, the Normal distribution is unbounded, and continues to plus or minus infinity.

The Wilson score interval (Wilson 1927) may be computed with the following formula.

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

Let us first consider cases where P is less than p. At the lower bound of this interval ($P = w^-$) the upper bound for the Gaussian interval for P, E^+, must be equal to p (Wallis 2013a).

We can carry out a test for significant difference between p and P by either

a) calculating a Gaussian interval at P and testing if p is greater than the upper bound, or
b) calculating a Wilson interval at p and testing if P is less than the lower bound.

[1] This paper summarises performance obtained with a spreadsheet by the author, www.ucl.ac.uk/english-usage/statspapers/wilson-dist.xls. Experimenting with different values of p and n is recommended.

To consider cases where P is greater than p, we simply reverse this logic. We test if p is smaller than the lower bound of a Gaussian interval for P, or P is greater than the upper bound of the Wilson interval for p. The Gaussian version of the test is called **the single proportion z test**. It can also be calculated as a **goodness of fit χ^2 test** (Wallis 2013a, b).

2. Plotting the distribution

We can define the Wilson *distribution* as follows:

- the distribution of the predicted probability of the true value P, based on an observation p, where P has a known relationship to p, computed using the Wilson score interval.

More precisely, we might consider it as the sum of two distributions:

- the distribution of the Wilson score interval lower bound w^-, based on an observation p and
- the distribution of the Wilson score interval upper bound w^+.

2.1 Obtaining values of w^-

First, we calculate the lower bound w^- from Equation (1) above for a series of values of α. In practice, we obtain a reasonably accurate initial plot by computing $z_{\alpha/2}$ and thus w^-, for $\alpha \in \mathbf{A}$ where $\mathbf{A} = \{0.0002, 0.05, 0.1, 0.15 \ldots 0.95, 1\}$, i.e. for intervals of 0.05 but excluding zero.

$$w^-(\alpha) = \left(p + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n} + \frac{z_{\alpha/2}^2}{4n^2}} \right) \Bigg/ \left(1 + \frac{z_{\alpha/2}^2}{n} \right). \qquad (2)$$

Note that $z_{\alpha/2}$ for $\alpha = 1$, $z_{1/2} = 0$, and for $\alpha > 1$, $z_{\alpha/2} = -z_{(1-\alpha/2)}$ and $w^-(\alpha/2) = w^+(1-\alpha/2)$. For $\alpha > 1$, calculating w^- computes percentage points above the observation p (i.e. w^+). So to compute w^+ we can simply extend $\mathbf{A}$ beyond 1, to include $\{1, 1.05 \ldots 1.9, 1.95, 1.9998\}$.

By inspection we note that the limit region below 0.05 (and above 1.95), is likely to see gradient change as α approaches zero. In other words, we cannot assume the line between these points is a straight line. Therefore we add points to $\mathbf{A}$ covering successive fractions $\{1/40, 1/80, \ldots 1/640\}$, and $\{2 - 1/40, 2 - 1/80, \ldots 2 - 1/640\}$.

Equation (2) obtains a position on a horizontal probability scale, w^-, computed for a given *cumulative* probability α. In other words, for $w^- < p$, the formula tells us that there is a probability of α that the true value is below w^-.

2.2 Employing a delta approximation

The next stage is to convert this cumulative probability into a column height. To do this we employ a *delta approximation*, a trick familiar to students of calculus.

The simplest method is to calculate Equation (2) for two points, α and $\alpha - \delta$. We approximate the area between the resulting values of w^- to a column δ wide and h high, we can compute h = *width* / *area*, which we can plot over w^-.

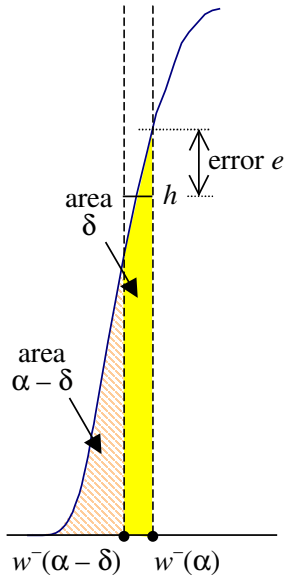

Figure 1. Estimating the height of $w^-(\alpha)$, $h(\alpha)$, using a onse-sided delta approximation. As $\delta \to 0$, error $e \to 0$.

Copyright © Sean Wallis 2018. All rights reserved.

To plot w^- for areas below p, $\alpha < 1$, we can use the following formula.

$$\begin{aligned} h(\alpha) &= 0 && \text{if } w^-(\alpha) = w^-(\alpha - \delta) \\ &= \frac{\delta}{w^-(\alpha) - w^-(\alpha - \delta)} && \text{otherwise.} \end{aligned} \quad (3)$$

The first test deals with cases where $p = 0$, which obtain a situation where all values of $w^-(\alpha) = 0$. We can continue this approximation for $\alpha \geq 1$. But if we want symmetric results for $p = 0.5$, we can take a delta above α for all cases for $\alpha \geq 1$.

$$\begin{aligned} h(\alpha) &= 0 && \text{if } w^-(\alpha) = w^-(\alpha + \delta) \\ &= \frac{\delta}{w^-(\alpha + \delta) - w^-(\alpha)} && \text{otherwise.} \end{aligned} \quad (4)$$

Finally, we set $h(1) = 0$ when $p = 0$ or 1.

Equation (3) converges to the correct value as $\delta \to 0$. It follows that δ should be as small as possible.

By experimentation, we find that if δ is below 0.0001, in some versions of Excel™ results become unreliable. Approximations in the computation of $z_{\alpha/2}$ seem to be the culprit. This leaves us with a small error in the calculation. We can see this error in that Equations (3) and (4) do not obtain exactly the same results. At the scale of the graph, this error is small, but perceivable.

To minimise this error, we average heights estimated using delta approximations above and below α. This improves the estimate for any monotonic region[2] $(\alpha - \delta, \alpha + \delta)$, and does not substantially worsen it if α represents a peak value.

$$h(\alpha) = \tfrac{1}{2} \left(\frac{\delta}{w^-(\alpha) - w^-(\alpha - \delta)} + \frac{\delta}{w^-(\alpha + \delta) - w^-(\alpha)} \right). \quad (5)$$

Although the distribution may be computed with a single formula over $\alpha \in (0, 2)$, recall that the Wilson distribution is really the sum of two distributions, each with a unit area of 1. The first of these areas is the distribution for the upper bound w^+, the second the distribution for the lower bound w^-. (This distinction will become important later on.)

To scale these distributions to the same scale as the equivalent Binomial or Normal distribution above and below p, we can divide both upper and lower bound distributions by $2n$.

3. Example plots

3.1 An initial example

To begin with, let us hold $n = 10$. This is a small sample size, but not *so* small as to present particular issues. First, we will consider $p = 0.5$. We obtain an interval that appears at first sight to match the Normal distribution.

[2] A region where the gradient is always increasing or decreasing, i.e. everywhere except where the peak value is within the range.

Copyright © Sean Wallis 2018. All rights reserved.

For the purposes of comparison we have also plotted Normal distributions centred at $P = w^-(0.05)$ and $w^+(0.05)$, divided by n. These distributions therefore have the same area (ignoring boundary clipping) as each corresponding area of the Wilson distribution below and above p.

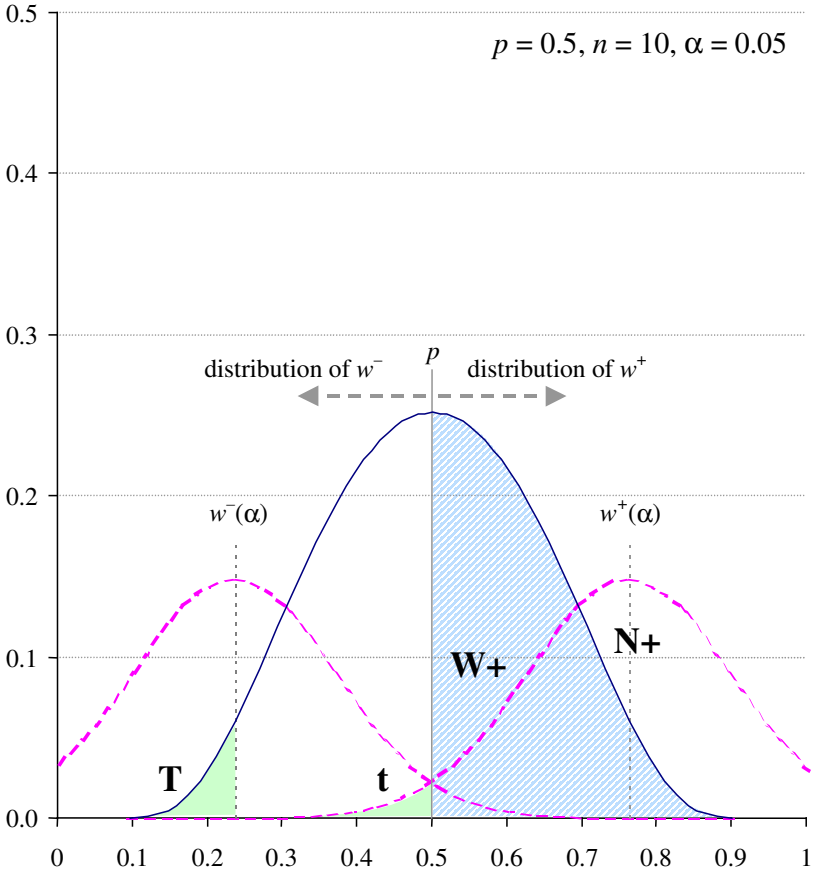

Figure 2. Plot of Wilson distribution (centre), with tail areas highlighted for $\alpha = 0.05$, plotted $p = 0.5$, $n = 10$; with Normal distributions centred at w^- and w^+.

3.2 Properties of the Wilson distributions

In this figure, the area under the Wilson distribution for w^+ (where $p > 0.5$), **W+**, has the same area as the area under the complementary Normal distribution **N+** (assuming that the Normal distribution is unclipped). In this case, area(**W+**) = area(**N+**) = $1/n$. It also has the same area as the Wilson distribution for w^-.

Provided that $p \in (0, 1)$ (i.e. it is not at the extremes), the interval will be two-sided, area(**W+**) = area(**W–**) and have a total area of $2 \times$ area(**W+**) = $2/n$.

The tail areas of the Wilson distributions represent 0.05 of the area under the curve above and below p respectively, in the same way as the equivalent tail areas of the Normal distribution.

The tail areas of both distributions on either side of p are also 0.05 of the area under those curves above and below these centres. For small n, the Normal distribution is visibly clipped by the probability range, but we can disregard the clipped section of these distributions for testing purposes, as our observation p is always on the inner side of these distributions.

The tail areas for the Normal, area(**t**) = area(**N+**) $\times \alpha/2$. Both tail areas for the Wilson interval, below $w^-(\alpha)$ and above $w^+(\alpha)$, are $\alpha = 0.05$ of each separate distribution. Thus in Figure 2, area(**T**) = area(**W–**) $\times \alpha$ (i.e., $\alpha/2$ of the total area). This obtains a two-tailed test when p is not at the extremes, but a one-tailed test when p is at 0 or 1.

Copyright © Sean Wallis 2018. All rights reserved.

3.3 Varying *p*

As p tends to 0, we obtain increasingly skewed distributions (Figure 3). The interval cannot be easily approximated by a Normal interval, and the sum of the two distributions is decidedly not Gaussian ('Normal').

In Figure 3, note how the mean p is no longer the most likely value (mode).

In plotting this distribution pair, the area on either side of p is projected to be of equal size, i.e. it treats as a given that the true value P is equally likely to be above and below p. This is not necessarily true! Indeed we might multiply both distributions by the probability of the prior. But this fact should not cause us to change the plot.

Note how, thanks to the proximity to the boundary at zero, the interval for w^- becomes increasingly compressed between 0 and p, reflected by the increased height of the curve.

The tendency to express the distribution like an exponential decline on the least bounded side reaches its limit when $p = 0$ or 1. The 'squeezed interval' is uncomputable and simply disappears.

3.4 Small *n*

What happens if we reduce n?

All else being equal we should expect that *the smaller the sample size, the larger the confidence interval*.

In the figures that follow we have plotted Wilson distributions for $p = 0$ and $p = 0.5$ for $n = 2$. Recall also that p must be a true fraction of n, so, for example, for $n = 2$, $p = 0.2$ would not be possible in practice.

The interval for $\alpha = 0.05$ now spans most of the range between 0 and 1. The boundaries 'squeeze' the interval close to 0 and 1. We obtain the 'wisdom-tooth' shape in Figure 4 and an undulating curve in Figure 5.

Note that the areas are larger because we are now scaling by 2/2 = 1 instead of 2/10 = 1/5.

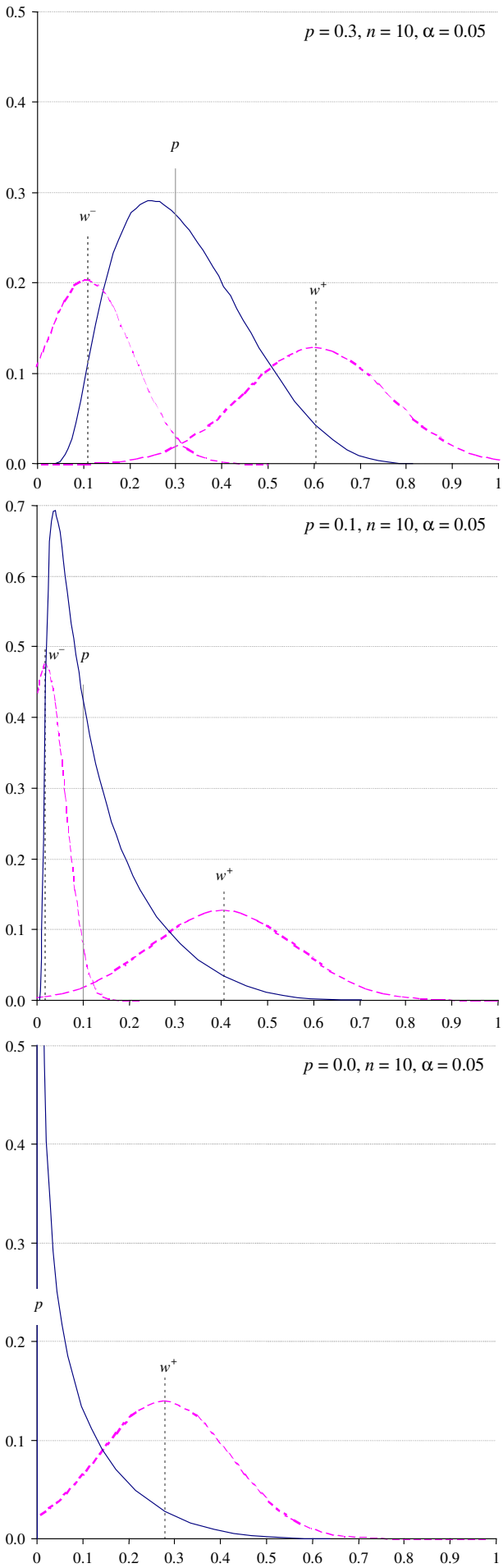

Figure 3. Plots of Wilson distributions for $p = 0.3$, 0.1 and 0.0.

Copyright © Sean Wallis 2018. All rights reserved.

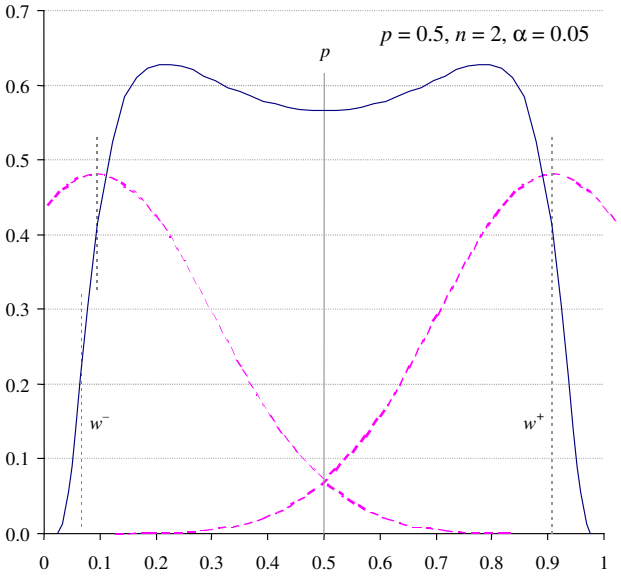

Figure 4. Plot of Wilson interval for $p = 0.5$ and $n = 2$. With such a large confidence interval, the boundaries at 0 and 1 cause the area to 'bulge' on either side.

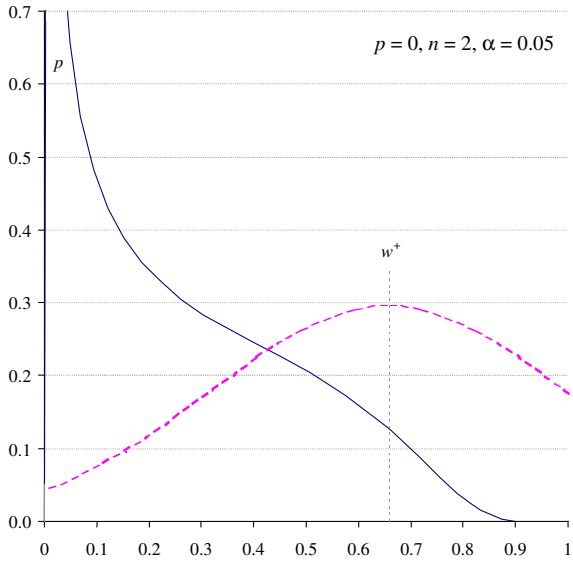

Figure 5. With $p = 0$, and $n = 2$, the gradient close to $w^+(0.05)$ is also affected by the boundary at 1, causing the gradient to undulate.

4. Further perspectives on the distribution

4.1 Percentiles of the Wilson distributions

We can plot percentiles of the distributions, as in Figure 6. The set **A** includes ten-percentile points, and we have simply plotted dividing lines to partition the area at each point.

Figure 6 contains two distributions, containing twenty areas in total, each equal in area.

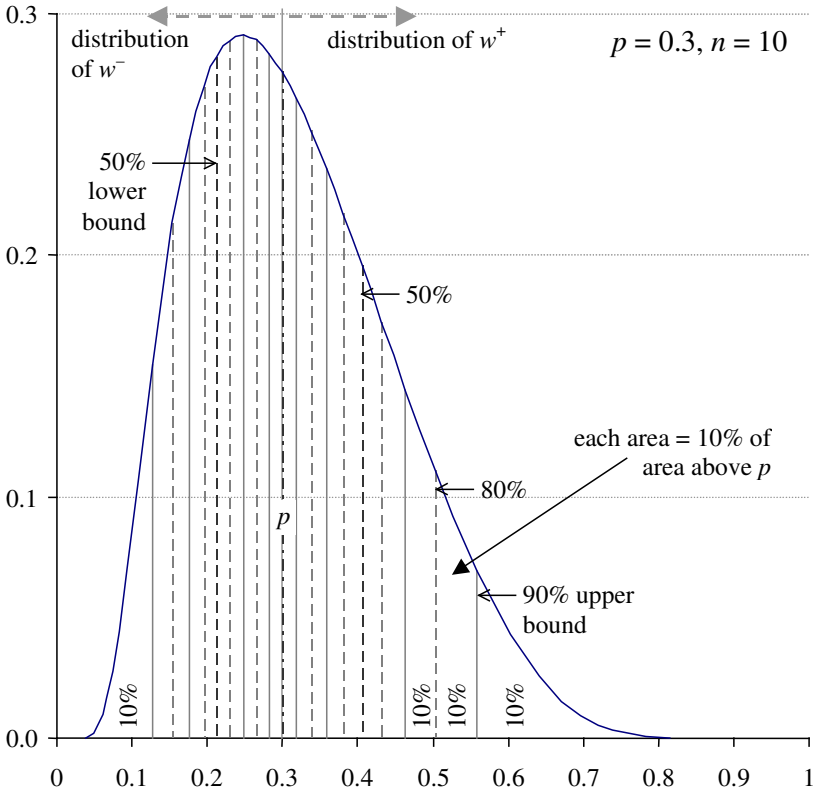

Figure 6. Ten-percentiles of the Wilson lower and upper distributions. Each area marked '10%' is of equal area. This is not always easy to see, particularly with respect to the tails.

Copyright © Sean Wallis 2018. All rights reserved.

4.2 The logit Wilson distribution

We earlier noted Robert Newcombe's observation (Newcombe, 1998) that – save when $p = 0$ or 1 – Wilson's score interval is symmetric on a logit scale.

Our method for logistic line fitting (regression) uses an estimate of variance based on the Wilson interval expressed on an inverse logistic, or 'logit' scale (Wallis 2015). Regression over variance relies on an assumption that the model of variance employed is Normal. In other words, it assumes the logit of the Wilson distribution resembles a Normal distribution.

We are now in a position to explore that assumption. We calculate logit(w^-) using Equation (2) and (6):

$$\text{logit}(p) \equiv \log(p) - \log(1 - p), \tag{6}$$

where log is the natural logarithm. Figure 7 plots the resulting distribution obtained by delta approximation, and (for comparison purposes) a closely-matching Gaussian distribution.

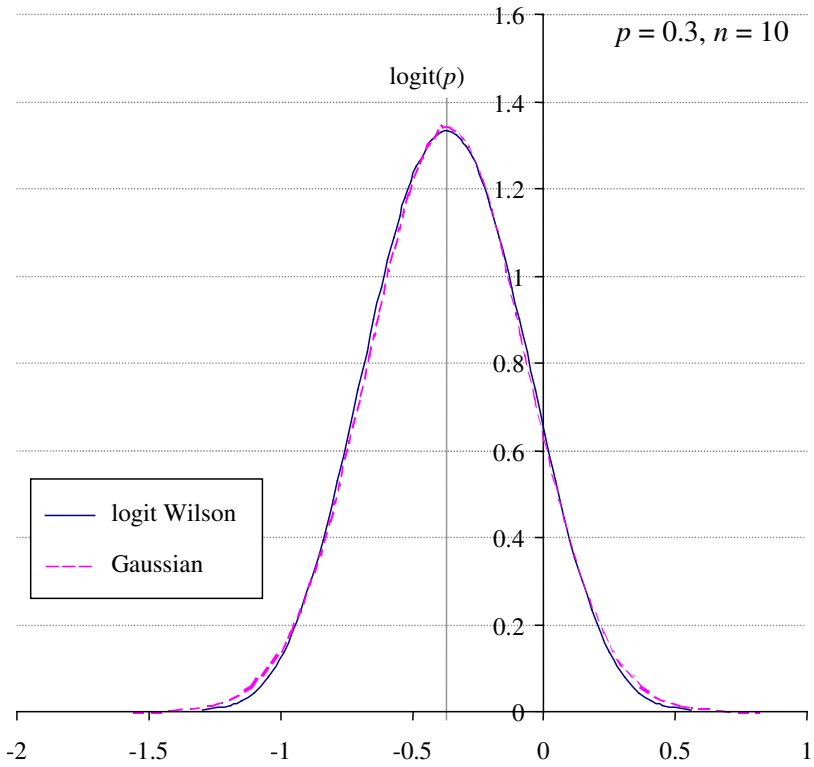

Figure 7. Logit Wilson distribution, i.e. the Wilson score interval on a logit scale, transformed into a distribution. This closely resembles a Gaussian (Normal) distribution centred on logit(p).

It turns out that, with the exception of when p is at boundaries 0 or 1 (which we exclude from fitting), the distribution closely matches a Normal distribution estimated by the following.

$$\begin{aligned}&\text{mean } \mu = \text{logit}(p),\\ &\text{standard deviation } \sigma = (\text{logit}(p) - \text{logit}(w^-(\alpha/2)) \,/\, z_{\alpha/2}.\end{aligned} \tag{7}$$

Figure 7 shows, by way of comparison, the Normal distribution estimated using $\alpha = 0.5$ in this formula. This approximation improves with increasing centrality and increasing n.

The approximation is not perfect, but it is considerably less prone to error than approximating the Normal to the Wilson interval on the probability scale (also known as the 'Wald' interval), or even the generally accepted approximation of the Normal to the Binomial distribution.

Copyright © Sean Wallis 2018. All rights reserved.

4.3 Continuity-corrected Wilson distributions

As we noted, the approximation from the discrete Binomial distribution to the Normal introduces an error that is conventionally mitigated with a continuity correction originally due to Yates (1934). In the case of the Normal distribution around P, this widens the interval by adding $0.5/n$ to the upper bound and subtracting this term from the lower bound.

Newcombe (1998) presents a formula for computing the equivalent Wilson score interval with continuity correction. The equation initially appears forbidding but it includes common terms that can be pre-calculated.

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

$$w^+ \equiv \min\left(1, \frac{2np + {z_{\alpha/2}}^2 + \{z_{\alpha/2}\sqrt{{z_{\alpha/2}}^2 - \frac{1}{n} + 4np(1-p) - (4p-2)} + 1\}}{2(n + {z_{\alpha/2}}^2)}\right). \quad (8)$$

This is the continuity-corrected version of Equation (1).

Earlier we emphasised that 'the Wilson distribution' was really two different distributions: one for w^- and one for w^+. Thanks to the continuity correction, these two formulae do not obtain the same result for $\alpha = 1$, unlike Equation (1), which converges to a midpoint.

This means we calculate intervals and heights separately.

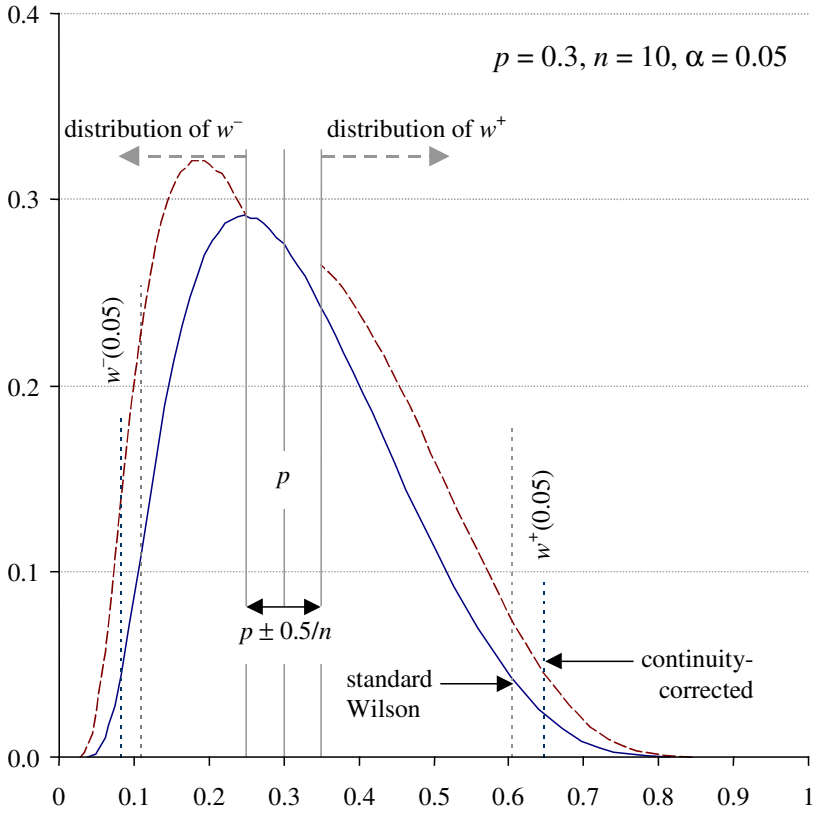

Figure 8. Uncorrected Wilson distribution (solid line) with continuity-corrected distributions for upper and lower bounds (dashed).

We can see the effect of the continuity correction on the intervals, rendering them more conservative (moving them further out from p), at the same time as causing the interval to be compressed even further within the probabilistic range [0, 1].

Copyright © Sean Wallis 2018. All rights reserved.

Conclusions

The Wilson score interval is a member of a class of confidence intervals that correctly characterise expected variation about an observation of a Binomial proportion, $p \in [0, 1]$. These intervals include the Clopper-Pearson interval, calculated by finding roots of the Binomial distribution for a given α, and the Wilson interval with continuity-correction that we document here. All three behave similarly, with the Clopper-Pearson falling between the two Wilson interval distributions depicted in Figure 8. See Newcombe (1998) and Wallis (2013a) for a comparison of competing intervals.

Common to this class of intervals is the fact that they are affected by boundary conditions at 0 and 1. In discussing the logistic curve, Wallis (2010) pointed out that the inverse logistic or 'logit' function maps a probabilistic range p to an unbounded Real dimension y by effectively folding space as it approaches the boundary. Figure 9 shows the idea.

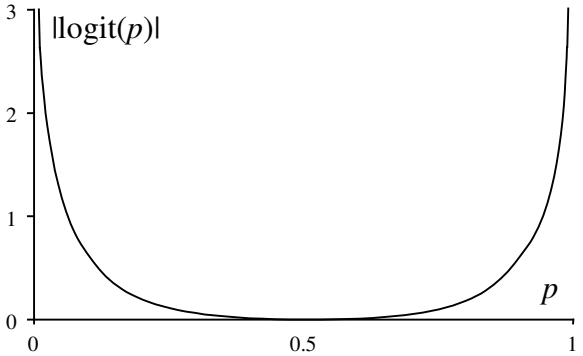

Figure 9. Absolute logit cross-section folding an infinite plane into a probabilistic trench. After Wallis (2010).

It is this folding of the interval into probability space that explains two aspects of the Wilson distribution we observe.

1. As p approaches 0 or 1, the distribution between the boundary and p becomes increasingly compressed and is pushed up, in some cases above the distribution at p. Meanwhile the interval on the 'open' side increasingly resembles a decay curve. This explains the shape of the distributions in Figure 3.

2. In Figure 4 and 5, we examined what happens to the distribution for small n. This appeared to generate what at first sight seems an even more baffling result, namely that for $p = 0.5$ and $n = 2$, the distribution had two peaks (it was 'bimodal'). A small n causes the distribution to spread over most of the probability range. The boundaries 'distort' what would otherwise be a declining interval. We see a similar but less dramatic effect for $p = 0$.

The logit transformation of the same interval for $p = 0.5$ and $n = 2$ obtains a 'bell curve' approximating to a Normal distribution about 0. We showed that provided that p was not at 0 or 1, not only is the logit Wilson interval *symmetric* as Newcombe (1998) pointed out, it resembles a Normal distribution. With increased n, the approximation improves, and for $n = 10$ the approximation is very close indeed (see Figure 7). This distribution is centred at logit(p), with a standard deviation that may be obtained from the width of the Wilson interval on a logit scale. This observation is support for the generalised logistic regression method described in Wallis (2015).

Our final comment relates to a point we made by way of introduction. It is often important to plot distributions to help us conceptualise the performance of what otherwise may appear to be

Copyright © Sean Wallis 2018. All rights reserved.

dry algebraic functions. Statistical distributions are not experienced directly. They represent the aggregated sum of experiences, and statistical reasoning is necessarily an act of imagination. The 'bell curve expectation' is the ideological predisposition to expect that variation around observations of any kind is Normal and symmetric. This expectation appears in the 'Wald' interval or presentations of 'standard error' for observed proportions or probabilities.

As we have shown, the predicted distribution of future observations based on a single observation of a Binomial proportion cannot be Normal. Where the observation is supported by a large *n* and the distribution is tightly spread, and/or where the observation is close to 0.5, the distribution may be approximately Normal.

But many types of data are highly skewed, and there are often good reasons why we might wish to work with small *n*. In the 1990s, medical statisticians started paying attention to this question.

Consider a clinical trial for a new heart drug for patients vulnerable to heart attacks. We have an expected rate of heart attacks for this group based on previous clinical data. We do not wish to recruit more subjects than necessary, so we must work with small *n*. The expected chance of a heart attack, *P*, over a short monitored period, *t*, is still small however, being close to zero.

A clinical trial manager must contend with two questions.

1. How many heart attack incidents would be significantly greater than would be expected by chance? In other words, does the lower bound of observed rate of heart attacks in the subject group, *p*, at a given time *t* exceed *P* sufficiently to be incapable of being explained by chance? The trial should stop immediately because the drug appears to be having a negative effect.

2. Following a trial period, is the drug working so well that further trials may be accelerated, more subjects recruited, etc.? To reach this conclusion we must examine the upper bound of our observed heart attack rate, p / t.

Either way, we are concerned with probabilities that are likely to be close to, but not equal to, zero, by observing proportions of events found in small samples. We need an accurate method for identifying when either stopping condition is reached without extending *t* longer than necessary. This is what the Wilson class of intervals obtains.

References

Newcombe, R.G. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17**: 857-872.

Wallis, S.A. 2010. Competition between choices over time. London: Survey of English Usage. http://corplingstats.wordpress.com/2012/03/31/competition-between-choices-over-time

Wallis, S.A. 2013a. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics* **20**:3, 178-208.

Wallis, S.A. 2013b. *z*-squared: the origin and application of χ^2. *Journal of Quantitative Linguistics* **20**:4, 350-378.

Wallis, S.A. 2015. Logistic regression with Wilson intervals. London: Survey of English Usage. http://corplingstats.wordpress.com/2015/04/24/logistic-regression

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

Yates, F. 1934. Contingency tables involving small numbers and the chi-square test. *Journal of the Royal Statistical Society*, **1**: 217-235.

Copyright © Sean Wallis 2018. All rights reserved.