Accurate confidence intervals on Binomial proportions, functions of proportions, algebraic formulae and effect sizes

Sean Wallis
Survey of English Usage, UCL

Abstract

In many fields, confidence intervals are growing in popularity, and citation is becoming mandatory in some journals. Plotting data and citing scores with confidence intervals conveys a model of sampling uncertainty to the reader absent from traditional approaches where plotting data and conducting analysis are separated. Researchers may also compare sampled scores for significant difference or against a given benchmark.

However, statistical sources commonly quote formulae based on *standard error*. This assumes that the probable true value of an observed parameter is Normally distributed, an assumption often untrue for small samples or observations near numeric bounds. This method generates intervals not consistent with standard statistical test procedures, and occasionally produces wholly implausible results.

In this paper we discuss a superior approach to constructing intervals for a wide range of properties. This builds on the *Wilson score interval* for the simple proportion p, which is robust on the probability scale $\boldsymbol{P} = [0, 1]$ and may be corrected for continuity and sampling. We demonstrate how we may compute intervals for properties that are functions of p (such as $\ln(p)$, $\text{logit}(p)$ and p^2), and, by employing Zou and Donner's *interval difference theorem*, for algebraic combinations of independent proportions p_1, p_2, etc. (such as $p_2 - p_1$, Σp_i, p_1 / p_2 and $p_1^{p_2}$). These methods are efficient to calculate, robust, and perform consistently with standard tests, while being capable of extension to novel statistical test procedures.

Key words: standard error, confidence interval, Wilson score interval, interval equality principle, interval difference theorem, risk ratio, odds ratio, effect size, interval evaluation

1. Introduction

There is a growing interest among practising researchers in plotting data with *confidence intervals* (sometimes termed 'credible intervals' or 'compatibility intervals') due to their explanatory power. However, statistics compendia typically offer limited coverage of confidence interval methods, and cited formulae often involve a common mathematical error.

This issue is of particular concern for linguists. Numerous research problems that linguists address engage Binomial or Multinomial statistics, i.e. the statistics of simple choice proportions. However, the treatment of Binomial intervals and derived properties (e.g. Zar 2010: 85, Sheskin 2011: 286, 661) tends to be weak.

In linguistics, a number of additional properties, scores and effect sizes are commonly cited. In Section 3 we show how one can give Gries's ΔP score (Gries 2013) an interval due to Newcombe (1998b). However, we need a *general* algebraic method for calculating confidence intervals for any linguistic property. This is the subject of this paper. We show how Zou and Donner's (2008) method can be used to create intervals for a wide range of properties, and evaluate intervals computed over differing numerical scales against a Fisher 'exact' test.

1.1 The standard error

A common, incorrect method for calculating confidence intervals employs (asymptotic) *standard error*. The model is extremely pervasive, finding its way into tests, algorithms and specialist treatises (e.g. Bishop, Fienberg and Holland 1975).

The model can be expressed simply as follows: given an observation of a variable x, assume that variation (uncertainty), scaled by a standard deviation $S(x)$, is Normally distributed about x.

$$\textit{standard error interval for } x,\ (e^-, e^+) = x \pm z_{\alpha/2}.S(x), \tag{1}$$

where $z_{\alpha/2}$ is the two-tailed standard Normal deviate for an error level α.[1] For the purposes of computing an interval, α is constant (say, 0.01 or 0.05), so $z_{\alpha/2}$ is also constant.

This standard deviation term may measure the variation of observed values within a sample, within-sample standard deviation, $s(x)$. In this case the interval models the *scatter* of the values observed within a sample.

However, inferential statistics concerns the sampling of an observation x from a population with true value, X. Here we are interested in *the standard deviation of sample means*. In plain English, we identify a standardised measure of the variability of observed *averages* (means) when they are sampled. Such a mean might be Real (e.g. the mean pitch of n utterances), Interval (the mean length of n clauses), or Binomial (the proportion of n clauses, phrases or words in a corpus with a particular feature). In this paper we will focus on *Binomial intervals* predicting a population proportion P. These intervals have the greatest utility to linguists. Binomial proportions may represent linguistic alternation rates, or observed rates derived from multiple choices, such as semasiological shares (Wallis 2021: 77) or standardised type-token ratios.

Engaging a mathematical model relies on making assumptions (*requirements*) that the data conforms to certain parameters. In the case of the Binomial model, these include (i) sampled instances should be drawn independently and randomly from the population, and (ii) instances must be free to vary, so that proportions (rates) can range from 0 to 100%. We also assume (iii) that the population is infinite, or much larger than the sample (see also Section 2.2 below).[2]

Equation (1) assumes that the distribution of uncertainty is Normal (Gaussian) and symmetric. For small samples of Real or Interval variables, $z_{\alpha/2}$ is replaced by the equivalent critical value of the (symmetric) t-distribution.

However, a symmetric interval is incompatible with *bounded* variables. Suppose x is an observed Binomial proportion, p. Typical examples are the proportion of phrases, clauses or sentences with a feature, an alternation rate or meaning share.

This property is bounded by the probabilistic scale $\boldsymbol{P} = [0, 1]$. The desired confidence interval may be written $p \in (p^-, p^+) = p \pm z_{\alpha/2}.S(p)$.

Phenomena that linguists study are often rare. For example, *somebody* in the written component of ICE-GB (Nelson *et al*. 2002) is infrequent (4 cases in 423,581 words), but its alternate, *someone*, is less so (82 cases).

written	*somebody f*	*someone*	total *n*	proportion *p*
non-printed	0	24	24	0.0000
printed	4	58	62	0.0645

Table 1. Alternation of *somebody/someone*, printed and non-printed ICE-GB subcorpora.

Written data is subdivided into print and non-print sources. In the non-printed data (Table 1), we find zero examples of *somebody*, i.e. $p(somebody) = 0$. One of these statements must be true.

1. $S(p) = 0$. The interval has zero width, $p^- = p = p^+$, and is thus symmetric. But this means the observation has *no error*. It is falsely certain.
2. $S(p) > 0$. The error interval has a non-zero width. The lower bound $p^- < 0$. It 'overshoots'. The model says there is a 50% chance that the true proportion, $P < 0$, which is also impossible.

Wallis (2021: 297) terms this presumption of Gaussian uncertainty the 'Normal fallacy'. A symmetric interval on a bounded variable cannot be correct.

[1] Sometimes, unhelpfully, the term 'standard error' is used as a substitute for 'standard deviation'.

[2] Assumption (ii) means that 'per million word' rates are unlikely to be Binomial proportions for most sampled linguistic phenomena. Alternative baselines should be considered (Wallis 2021a: 47). For text corpus samples, assumption (i) can be addressed by a 'random-text sampling' correction (Wallis 2021a: 277).

1.2 Population sampling intervals and confidence intervals

The conventional model used in χ^2 and z tests employs the Normal approximation to the Binomial population proportion P. This gives us a legitimate, albeit approximate, Normal interval about P.

$$\text{population standard deviation } S(P) \equiv \sqrt{P(1-P)/n} \text{ , and} \qquad (2)$$
$$\text{Gaussian interval } (E^-, E^+) = P \pm z_{\alpha/2}.S(P).$$

This *population interval* identifies the range of values a sampled proportion p will be expected to be found at a given error level, α, given P.

We may engage in 'what if' reasoning. Suppose we thought the true rate of the *somebody/someone* alternation in printed texts of the kind found in ICE-GB was 15 in 100, i.e. P = 0.15. In a sample of 62 cases, the observed proportion should fall within $(E^-, E^+) = 0.15 \pm 0.0889 =$ (0.0611, 02389) at the $\alpha = 0.05$ error level, i.e. on 19 out of 20 sampling attempts.

We can now compare our observed rate, $p = 0.0645$, with this interval. It is within the range, so the observed p is not significantly different from the hypothesised rate P.

This model is not perfect. If $P = 0$ or 1 the interval width becomes zero. It overshoots near the boundary (e.g. for $P = 0.01$ and $n = 62$, $E^- = -0.0148$). Fortunately, applying Yates's *correction for continuity* (Yates 1934) conservatively compensates for both problems, and the 'smoothing error' created by the approximation of the discrete Binomial distribution by a continuous Normal curve. The interval is moved away from P by half an instance on either side.

$$\text{Yates's Gaussian interval } (E^-_{cc}, E^+_{cc}) = P \pm (z_{\alpha/2}.S(P) + \tfrac{1}{2n}). \qquad (3)$$

Equations (2) and (3) are employed in the *z test for the single proportion* (Wallis 2013) to compare an observed proportion, p, with an expected one, P. They perform identically to their equivalent 2 × 1 χ^2 goodness of fit test (see Equation (12), below).

However, population intervals have limited utility. Usually we do not know P. Instead we wish to predict the most likely range of values of P based on the observed rate, p, and the error level, α. We need a *confidence interval* for p.[3]

For decades, researchers wishing to create confidence intervals were directed to employ Equations (2) or (3), substituting p for P. Thus the following would be used in place of (2).

$$\text{observed standard deviation } S(p) \equiv \sqrt{p(1-p)/n} \text{ , and} \qquad (4)$$
$$\text{Wald interval } (e^-, e^+) = p \pm z_{\alpha/2}.S(p).$$

In one statistics reference after another, we see standard error or 'Wald' intervals quoted. But they obtain results that are *inconsistent* with their equivalent z or χ^2 test.

Consider our earlier example where $P = 0.15$ and $n = 62$. Substituting $p = 0.0645$ into Equation (4), we obtain the interval $(e^-, e^+) = (0.0034, 0.1256)$, which excludes $P = 0.15$. The Wald interval has given us a different result!

This interval has zero-width behaviour for $p = 0$ or 1, and overshoots near the boundary. These problems are conventionally addressed by the '3-sigma rule', which rules out Equation (4) if $p \pm 3S(p)$ exceeds [0, 1] (i.e. for small samples and proportions close to 0 or 1). Yet the principal utility of inferential statistics concerns small samples, and many fields, linguistics included, frequently contend with low-frequency terms.

But arguably the worst problem is that the Wald interval obtains results inconsistent with the Gaussian model (Equation (2)), i.e. it rules results 'significant' when the equivalent z test does not,

[3] Confidence intervals should not be confused with 'replication' or 'resampling' intervals. A confidence interval predicts the range of values of the true mean P, whereas a resampling interval predicts the range of the mean of a second sample. See Wallis (2020).

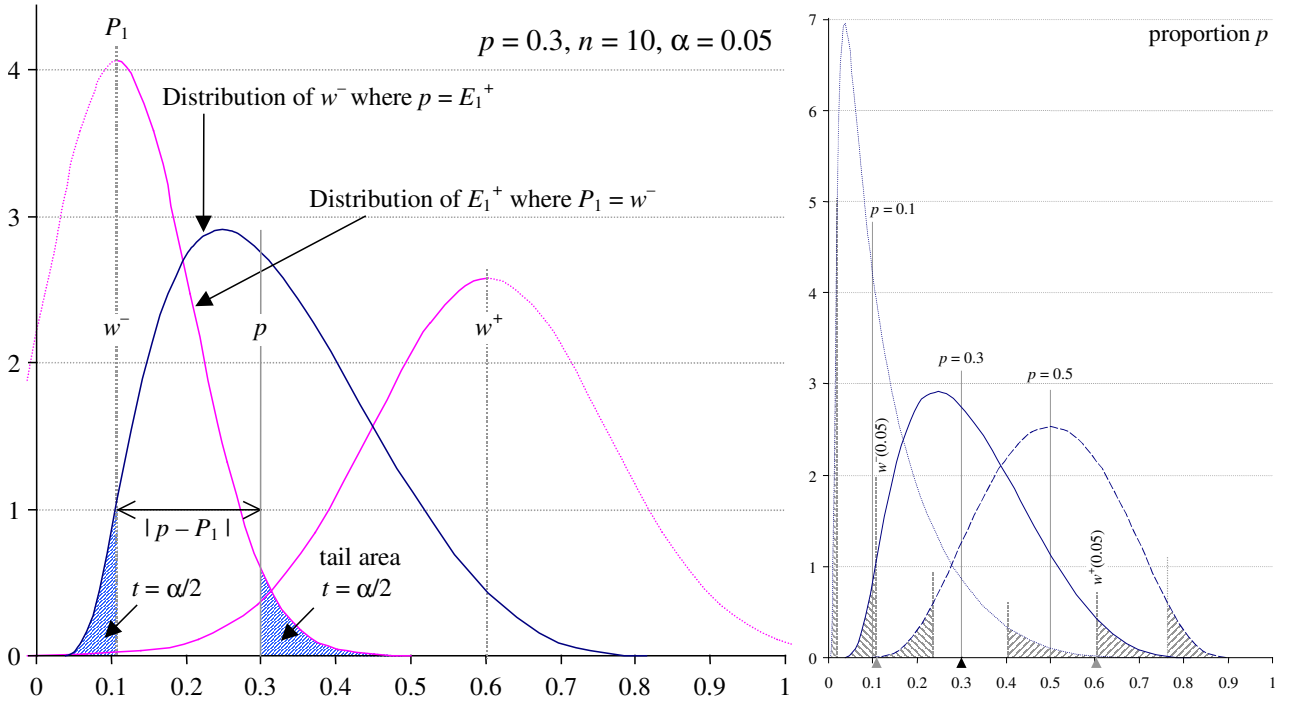

Figure 1. Left, the interval equality principle illustrated: example pdf distributions for the bounds of the Wilson interval (w^-, w^+), for $p = 3/10$ at $\alpha = 0.05$, with half-Gaussian distributions at each bound. Right: Wilson distributions for $p \in \{0.1, 0.3, 0.5\}$ and $n = 10$.

and vice-versa. This, we believe, is the source of the historic low status of confidence intervals. *Without a method for computing a confidence interval consistent with the equivalent significance test, confidence intervals cannot be 'proper' statistics.*

However, if we can address this problem, then confidence intervals become very powerful. Plotted intervals are visually intuitive and permit us to contrast observations on the same scale by eye without performing significance tests. See Figure 10.

This paper is set out as follows. In the next section we discuss the interval equality principle, which defines an interval by inverting an equivalent test procedure, guaranteeing consistency with the test. In Section 3 we introduce difference intervals and tests, and in Section 4 we generalise both approaches with mathematical functions and operators. We give examples for effect sizes in Section 5 and Section 6 is the conclusion.

2. The interval equality principle

The reasoning supporting the Wald formula is incorrect. The probability distribution of a population score P given an observed proportion p is *not* Normal. See Figure 1.

Wilson (1927) argued that confidence intervals on an observed Binomial proportion p should be obtained by mathematically *inverting* the population interval formula (Equation (2)), rather than substituting p for P (Equation (4)). This *interval equality principle* (Wallis 2021a: 319) is simply stated:

> when P is at a bound of p, p is at the opposite bound of P.

The difference between observed and expected proportions, $|\, p - P \,|$, must be the same, whether one measures it from p or from P (Figure 1, left, arrowed).

Suppose P_1 and P_2 represent the two potential values of P on either side of p just sufficiently distant to be deemed significant by a single-sample z test (Equation (2)). The following hold:

$$p = E_1^+ \equiv P_1 + z_{\alpha/2}.S(P_1), \text{ and} \qquad (5)$$
$$p = E_2^- \equiv P_2 - z_{\alpha/2}.S(P_2).$$

We wish to find two values: P_1, with its Normal upper bound at p, and P_2, with p at its lower bound. The interval (P_1, P_2) identifies the range of values not significantly different from p at the required error level (hence 'compatibility interval').

2.1 The Wilson score interval

We could employ a search procedure to find P_1 and P_2 (see Section 2.3). However, there is a more efficient method. Wilson solves Equation (2) to derive a direct formula for the interval:

$$\textit{Wilson score interval } (w^-, w^+) \equiv \left(p + \frac{{z_{\alpha/2}}^2}{2n} \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n} + \frac{{z_{\alpha/2}}^2}{4n^2}} \right) \Bigg/ \left(1 + \frac{{z_{\alpha/2}}^2}{n} \right). \qquad (6)$$

Recall our population interval for $P = 0.15$, $n = 62$. Substituting the 95% lower bound, $p = E^- = 0.0611$, into Equation (6) yields $(w^-, w^+) = (0.0235, 0.1500)$. The upper bound, $p = E^+ = 0.2389$, obtains (0.1500, 0.3582). The interval precisely inverts Equation (2).

We are used to seeing Normal distribution curves. What shape does the equivalent Wilson distribution have? Figure 1, left, plots the distribution of w^- and w^+ on either side of $p = 3/10$ for varying α. They meet at p where $\alpha = 1$.[4] This curve is not Normal, but has a well-defined relationship (the interval equality principle) with the inner half of Normal intervals at each tail. The tail areas, t, are equal for any value of α.

Figure 1, right, depicts example curves for $p = 0.1$, 0.3, and 0.5 for the same sample size. Apart from where $p = 0.5$, curves and intervals are asymmetric, tending towards the middle of the probabilistic range.

The interval has a number of important properties.

1. Distributions are entirely contained within $\boldsymbol{P} = [0, 1]$. At the limit where $\alpha \to 0$, $P \to 0$ (lower) or 1 (upper bound).
2. If n is large or p close to 0.5, the curve appears approximately Normal (this has long excused the 'Wald' standard error interval). But it is also well-behaved for small n and skewed p.
3. The Wilson model is related to the logistic ('S-curve') model. Except for $p = 0$ or 1, where the interval will be one-sided and logit(p) infinite, the *logit-Wilson* (the logit transform of Figure 1) is symmetric and tends to a Normal distribution (Wallis 2021a: 308, and Figure 8).

2.2 Adjusting the Wilson formula using functional notation

The formula may be corrected for continuity and adjusted for sampling.

For a two-tailed interval, let us define the functions

$$\text{WilsonLower}(p, n, \alpha/2) \equiv w^-, \text{ and}$$
$$\text{WilsonUpper}(p, n, \alpha/2) \equiv w^+, \qquad (7)$$

where w^- and w^+ are defined by Equation (6).

Adjustments to intervals are now straightforward. Wallis (2021a: 161-162) notes that both Yates's continuity correction and a finite population correction may be simultaneously applied to the Wilson interval.

[4] Equation (6) is the cumulative density function of the interval. We vary $\alpha \in (0, 1]$ in Equation (6) and compute the height (pdf) by delta approximation. See Wallis (2021a: 297-307).

The continuity-corrected Wilson score interval (w^-_{cc}, w^+_{cc}) is obtained by moving the first parameter away from p by Yates's continuity correction term, $\frac{1}{2n}$:

$$w^-_{cc} = \text{WilsonLower}(\max(0, p - \tfrac{1}{2n}), n, \alpha/2), \text{ and}$$
$$w^+_{cc} = \text{WilsonUpper}(\min(1, p + \tfrac{1}{2n}), n, \alpha/2). \quad (8)$$

The 'max' and 'min' functions restrict the first parameter to the probabilistic range $\boldsymbol{P}$. This interval is consistent with Equation (3).

Now, suppose we wish to employ a *finite population correction* (Wallis 2021a: 160). This factor reduces the variance when a sample is a subset of a finite population. It may be defined as

$$\textit{finite population correction } \nu = \sqrt{(N-n)/(N-1)}\,, \quad (9)$$

where N is the population size and n the sample size. It can be used to scale the variance, $S^2(P) = \nu^2 P(1-P) / n$. To scale the Wilson interval, we divide the weight of evidence, n, by ν^2.

$$w^- = \text{WilsonLower}(p, n / \nu^2, \alpha/2), \text{ and}$$
$$w^+ = \text{WilsonUpper}(p, n / \nu^2, \alpha/2), \quad (10)$$

and, likewise,

$$w^-_{cc} = \text{WilsonLower}(\max(0, p - \tfrac{1}{2n}), n / \nu^2, \alpha/2), \text{ and}$$
$$w^+_{cc} = \text{WilsonUpper}(\min(1, p + \tfrac{1}{2n}), n / \nu^2, \alpha/2). \quad (11)$$

Note how the continuity correction term, $\frac{1}{2n}$, is not altered by the finite population correction. The corrections are independent. Adjustments to variance to account for *random-text sampling* (Wallis 2021a: 277) may be incorporated in the same way.

2.3 Obtaining intervals by search

We can also invert a population interval function or 2 × 1 test procedure by search.[5] For example, goodness of fit χ^2 or log-likelihood may be calculated from two simple pairs:

$$\textit{observed } o_i \in [np, n(1-p)], \text{ and}$$
$$\textit{expected } e_i \in [nP, n(1-P)]. \quad (12)$$

Suppose we desire a log-likelihood interval (g^-, g^+) for a given p.

$$\textit{log-likelihood } G^2 \equiv 2\sum_{i=1}^{k} o_i \log(o_i / e_i)\,. \quad (13)$$

We search for the lower bound $g^- < p$:

$$\text{find } P = g^- \in [0, p) \text{ where } G^2 = \chi^2_{crit}(\alpha, 1) = z_{\alpha/2}{}^2. \quad (14)$$

The upper bound, g^+, may be found via a search for the lower bound for $1 - p$.

[5] Wallis (2021: 322) offers a binary search algorithm for a lower bound between 0 and p.

An alternative approach avoids the Normal approximation. The 'exact' *Clopper-Pearson interval* (Newcombe 1998a; Wallis 2013, 2021a: 147) is found by search using the cumulative Binomial function. Although more difficult to compute, it is useful for evaluation purposes.

2.4 Performance

Various researchers (e.g. Newcombe 1998a, Brown *et al.* 2001, Wallis 2013) have compared the performance of Binomial intervals. The Wilson score interval with continuity-correction improves over the uncorrected Wilson, log-likelihood and 'Wald' approaches (Figure 2). The closest fit to the 'exact' Clopper-Pearson (shaded) is the continuity-corrected Wilson interval.

Newcombe (1998a: 868) recommends that 'Wald' methods should be replaced, commenting that 'intervals calculated by these methods should no longer be acceptable for the scientific literature.' Wilson-based methods should be substituted.

Newcombe's assessment is that even without continuity correction, the Wilson score interval's average coverage probability is very close to the nominal value $1 - \alpha$, and, with a continuity correction, the method is 'nearly strictly conservative' (has almost no Type I errors) with minimum coverage probability 0.949 for a nominal 0.95. These intervals, including the Clopper-Pearson, tend to be slightly too 'mesial', i.e. too close to $p = 0.5$.

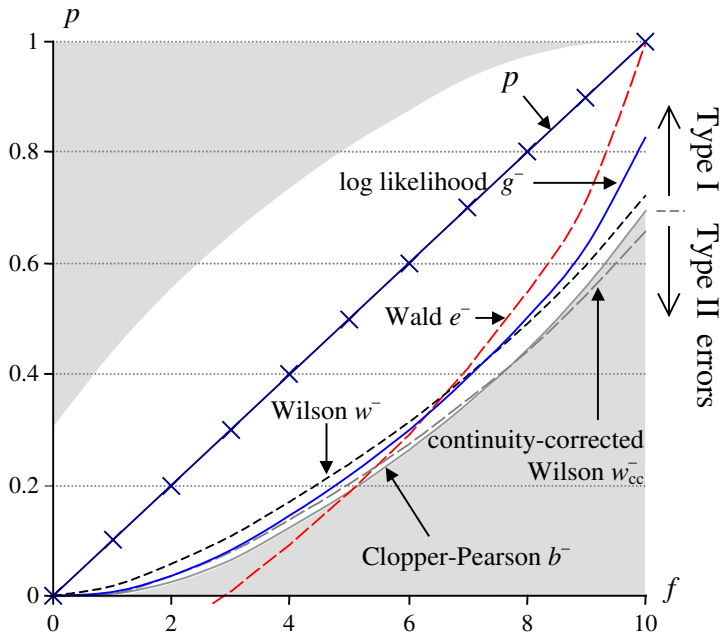

Figure 2. Estimates for the lower bound of p, $p = f / n$, $n = 10$, $\alpha = 0.05$, after Wallis (2013). We can clearly see the inconsistency of the 'Wald' interval, which is not computed by interval equality.

Figure 3 plots distributions of Wilson and Clopper-Pearson intervals for $p = 0.3$ and $n = 10$. Upper and lower bounds are computed separately, leaving a discontinuous space on either side of p for both Clopper-Pearson and continuity-corrected Wilson distributions.[6] All three methods are constrained by boundaries (as $P \to 0$ or 1, $\alpha \to 0$).

[6] (Wallis 2021a: 312) notes that distributions computed by the interval equality principle, while guaranteed consistent with the equivalent significance test, are not strictly equivalent to 'the probability distribution of P given p'. Instead they are best considered as independent distributions for each bound.

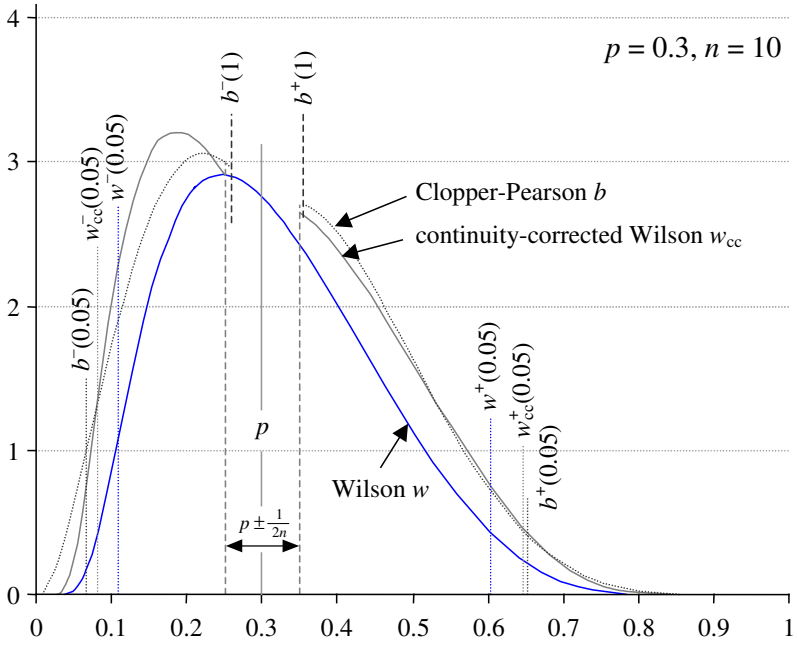

Figure 3. Continuity-corrected Wilson and Clopper-Pearson intervals, and their probability density distributions, unit scale, after Wallis (2021a: 311).

3. From contingency tests to difference intervals

We have evaluated the difference $d = p - P$. A related task is evaluating the difference between two independently observed proportions, $d = p_2 - p_1$, which has a general application. An interval for Stefan Gries's (2013) $\Delta P = p_1 - p_2$ can be easily obtained from the following.

3.1 A chi-square based interval

The 2 × 2 χ^2 test for homogeneity (Sheskin 2011: 643) is well-known:

$$\textit{chi-square}\ \chi^2 = \sum_{i=1..r} \sum_{j=1..c} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}}, \qquad (15)$$

where $o_{i,j}$ represent observed cell frequencies in Table 2, and the expected frequency $e_{i,j} = o_{i+} \times o_{+j} / o_{++}$. The sum is compared to the critical value $\chi^2_{crit}(\alpha, 1) = z_{\alpha/2}{}^2$.

This test can be reformulated as a *z test for two independent proportions* (Sheskin 2011: 655). We test if p_1 and p_2 are significantly different, assuming both are drawn from one population with a mean equal to the *pooled probability estimate*, $\hat{p}$.

We compare d to a Normal interval $(-e_d, e_d)$ centred on zero, obtained by

$$\begin{aligned} &\textit{probability estimate} && \hat{p} \equiv (n_1 p_1 + n_2 p_2) / (n_1 + n_2) = o_{1+} / o_{++}, \text{ and} \\ &\textit{standard deviation} && S(\hat{p}) \equiv \sqrt{\hat{p}(1-\hat{p})\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)}, \\ &\textit{standard error} && e_d = z_{\alpha/2}.S(\hat{p}). \end{aligned} \qquad (16)$$

	Outcome 1		Outcome 2		Total	
Condition 1	$o_{1,1}$	$= n_1 p_1$	$o_{2,1}$	$= n_1(1-p_1)$	o_{+1}	$= n_1$
Condition 2	$o_{1,2}$	$= n_2 p_2$	$o_{2,2}$	$= n_2(1-p_2)$	o_{+2}	$= n_2$
Total	o_{1+}	$= o_{1,1} + o_{1,2}$	o_{2+}	$= o_{2,1} + o_{2,2}$	o_{++}	$= n_1 + n_2$

Table 2. Contingency table comparing two proportions, p_1 and p_2 supported by n_1 and n_2. Row and column totals are indicated by '+' subscripts.

The formula performs identically to the χ^2 test and may incorporate a continuity correction and other adjustments. However, if repositioned about d (Sheskin 2011: 661), it performs poorly.

3.2 The Newcombe-Wilson interval

Newcombe (1998b) computes an alternative interval for $d = p_2 - p_1$ by summing variances independently estimated at inner interval bounds. Recall that the Wilson interval (w^-, w^+) assumes P has a Normal distribution at each bound (Figure 1).

Suppose $p_1 > p_2$. To find out if p_1 is significantly greater than p_2 we examine the intervals on the inner side of this difference, i.e. w_1^- and w_2^+. Figure 4 illustrates the idea. We assume intervals are independent and tangential, and independent variances may be summed (termed the Bienaymé theorem).[7]

$$S^2(d) = \begin{cases} S^2(w_1^-) + S^2(w_2^+) & \text{if } p_1 > p_2, \\ S^2(w_1^+) + S^2(w_2^-) & \text{otherwise.} \end{cases} \quad (17)$$

We know the following interval widths must be equal:

$$(p_1 - w_1^-) = z_{\alpha/2}.S(w_1^-),$$
$$(w_2^+ - p_2) = z_{\alpha/2}.S(w_2^+), \text{ etc.}$$

A zero-based difference interval is computed from Equation (18).

$$(w_d^-, w_d^+) = (-\sqrt{(p_1 - w_1^-)^2 + (w_2^+ - p_2)^2}\ ,\ \sqrt{(w_1^+ - p_1)^2 + (p_2 - w_2^-)^2}\ ). \quad (18)$$

This *Newcombe-Wilson interval* is asymmetric, unlike Equation (16). It may be used as a significance test for the difference between two Binomial proportions by simply testing if d lies outside it.

It may also be repositioned about d by simple subtraction:

$$d \in (d^-, d^+) = d - (w_d^-, w_d^+) = (d - w_d^+, d - w_d^-). \quad (19)$$

Whereas (w_d^-, w_d^+) has origin 0, this interval (d^-, d^+) is an interval for d, like the Wilson interval for p. We can now plot confidence intervals for d, and test if d is less than or greater than a given difference D. The interval may also be further generalised, for example, to compare two observed differences d_1 and d_2. See Section 5.3.

Figure 5 sketches some consequences of this formulation. The first sketch, left, reminds us that the Newcombe-Wilson test may be performed either by checking whether the zero-based interval (w_d^-, w_d^+) excludes d or if the difference-based interval (d^-, d^+) excludes zero. The central figure emphasises that the inner gradient is the shallowest gradient for d (conventionally tested against zero) and the outer gradient is the steepest.

The third sketch compares these gradients with Wilson intervals at their end points. The shallowest gradient (here d^+) falls within single interval bounds, so if two intervals have no overlap the difference $d = p_2 - p_1$ must be significantly falling or rising. The difference cannot be significant if either point is within the Wilson interval of the other.

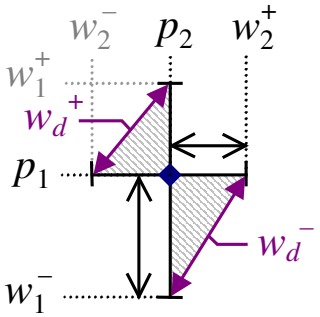

Figure 4. Calculating the bounds of the Newcombe-Wilson interval using the Bienaymé formula.

[7] Probability space ($\boldsymbol{P} \times \boldsymbol{P}$) is curved, and this formula introduces a small error. However this does not undermine the overall viability of the method. We explore the impact of this approximation over different scales in Appendix 1.

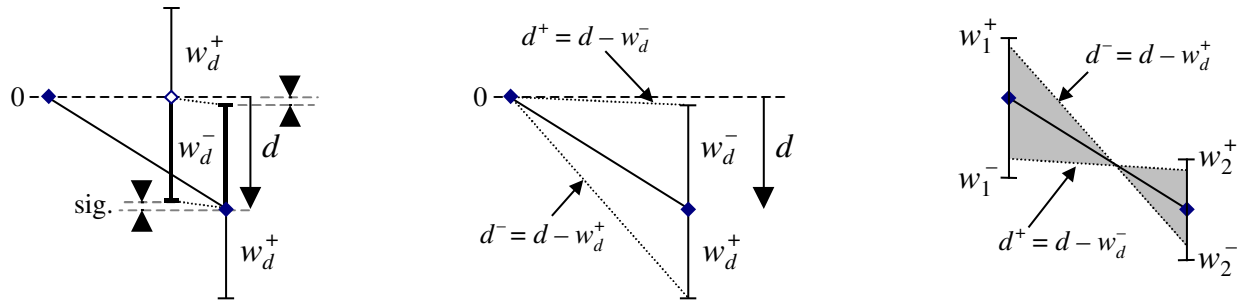

Figure 5: Geometry of Newcombe-Wilson difference intervals, for a significant fall $d < 0$, after Wallis (2021a). Left: repositioning the Newcombe-Wilson interval at d allows us to test if $d < w_d^-$ or $d - w_d^- < 0$. Middle: identifying slopes for $d \in (d^-, d^+)$. Right: with Wilson intervals for p_1 and p_2.

Wallis (2021a: 119) terms this system the *Wilson interval comparison heuristic*:

Wilson interval comparison heuristic: (20)

For any pair of points representing comparable observed proportions:

1. if two points' intervals **do not overlap**, the points *must* be significantly different;
2. if one point is **inside the interval** of the other, the points *cannot* be significantly different;
3. **otherwise**, carry out a statistical test to decide whether they are significantly different.

Provided that p_1, p_2, etc. are plotted on the same scale, one can perform a visual assessment as a first pass, eliminating demonstrably significant and non-significant cases. Only where intervals partially overlap is a significance test required.

Although referred to as the 'Wilson' interval comparison heuristic, thanks to Zou and Donner's generalisation (see 4.2) it can supplement many interval comparisons.

3.3 Performance

The left graph in Figure 6 was constructed by iterating cell frequencies in Table 2. As one proportion increases, the other declines: $p_1 = 1 - p_2$, $p_2 \in [0, 1]$, so d increases.

Whether or not a continuity-corrected is employed, the Newcombe-Wilson interval is well-behaved, neither overshooting nor collapsing to zero-width. But a repositioned Gaussian interval overshoots. Figure 6, right, examines intervals for a constant difference $d = 0.5$.

We can also examine probability density curves for these intervals. Figure 7 plots the

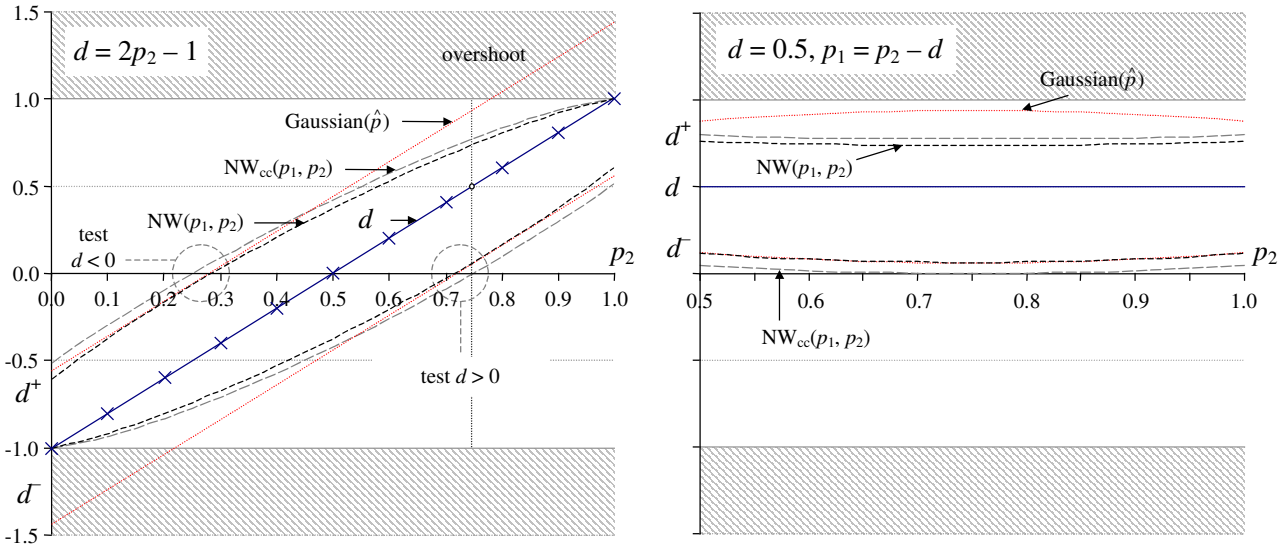

Figure 6. Estimates for 95% intervals on difference $d = p_2 - p_1$ in a 2 × 2 matrix with $n_1 = n_2 = 10$. Left: $p_1 = 1 - p_2$, $d \in [-1, 1]$. Right: a constant difference $d = 0.5$, and $p_1 = p_2 - d$.

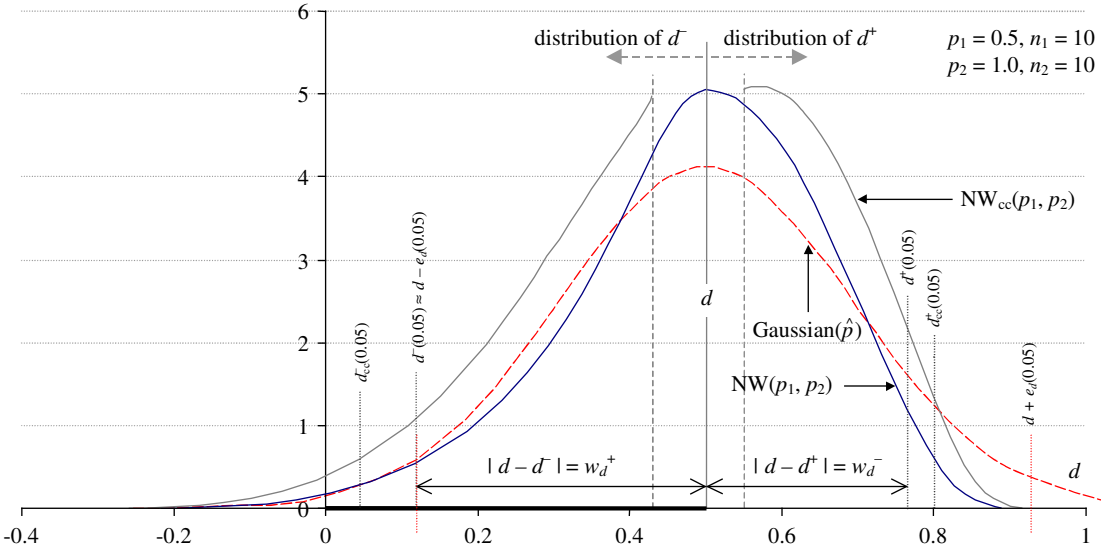

Figure 7. Newcombe-Wilson distributions for d^- and d^+, unit scale, where $p_1 = 0.5$, $p_2 = 1.0$, $n_1 = n_2 = 10$. For comparison, the (erroneous) repositioned Gaussian interval is shown.

distributions of interval bounds where $p_1 = 0.5$, $p_2 = 1.0$, and $n_1 = n_2 = 10$.

These methods are consistent for testing against zero (the archetypal 2 × 2 test scenario). But Newcombe's method is clearly superior when repositioned.

4. Confidence intervals for other properties

Although Binomial proportions are ubiquitous in linguistic research problems, we often wish to compute intervals for other properties.

4.1 Functions of the Binomial proportion

We can simply obtain confidence intervals for *monotonic functions* of p (Wallis 2021a: 175). Monotonic functions always either increase or decrease over the parameter's range and have a unique solution when inverted.

For any function $fn(p)$ of a Binomial proportion p that is monotonic over $p \in \boldsymbol{P} = [0, 1]$, the transformed Wilson score interval is

$$\textit{transformed Wilson}\ (w_t^-, w_t^+) = \begin{cases} (fn(w^-), fn(w^+)) & \text{if } fn \text{ increases with } p, \text{ or} \\ (fn(w^+), fn(w^-)) & \text{otherwise.} \end{cases} \qquad (21)$$

For example, the *logit* (log odds) function, $\text{logit}(p) \equiv \ln(p / (1 - p))$, is monotonic and increasing, so the logit Wilson interval is simply $(\text{logit}(w^-), \text{logit}(w^+))$. The *reciprocal* function, $1/p$, monotonically decreases, so its interval, $(1/w^+, 1/w^-)$, has interval bounds reversed. To compute an interval for mean clause length, $\bar{l}$ = words/clauses, we can use p = clauses/words.

Probability density distributions for selected functions of $p \in \{0.1, 0.3, 0.5\}$ with $n = 10$ are shown in Figure 8. Exceptionally, the logit Wilson is symmetric and approximately 'Normal' (Wallis 2021a: 307), but note in passing how the others have very different distributions!

Intervals subject to non-monotonic transforms require us to identify *turning points* (local minima or maxima). Suppose that a is a turning point within an interval (w^-, w^+). The lower bound is simply $\min(fn(w^-), fn(w^+), fn(a))$, and the upper bound is the maximum of the same sequence. See Section 5.1.

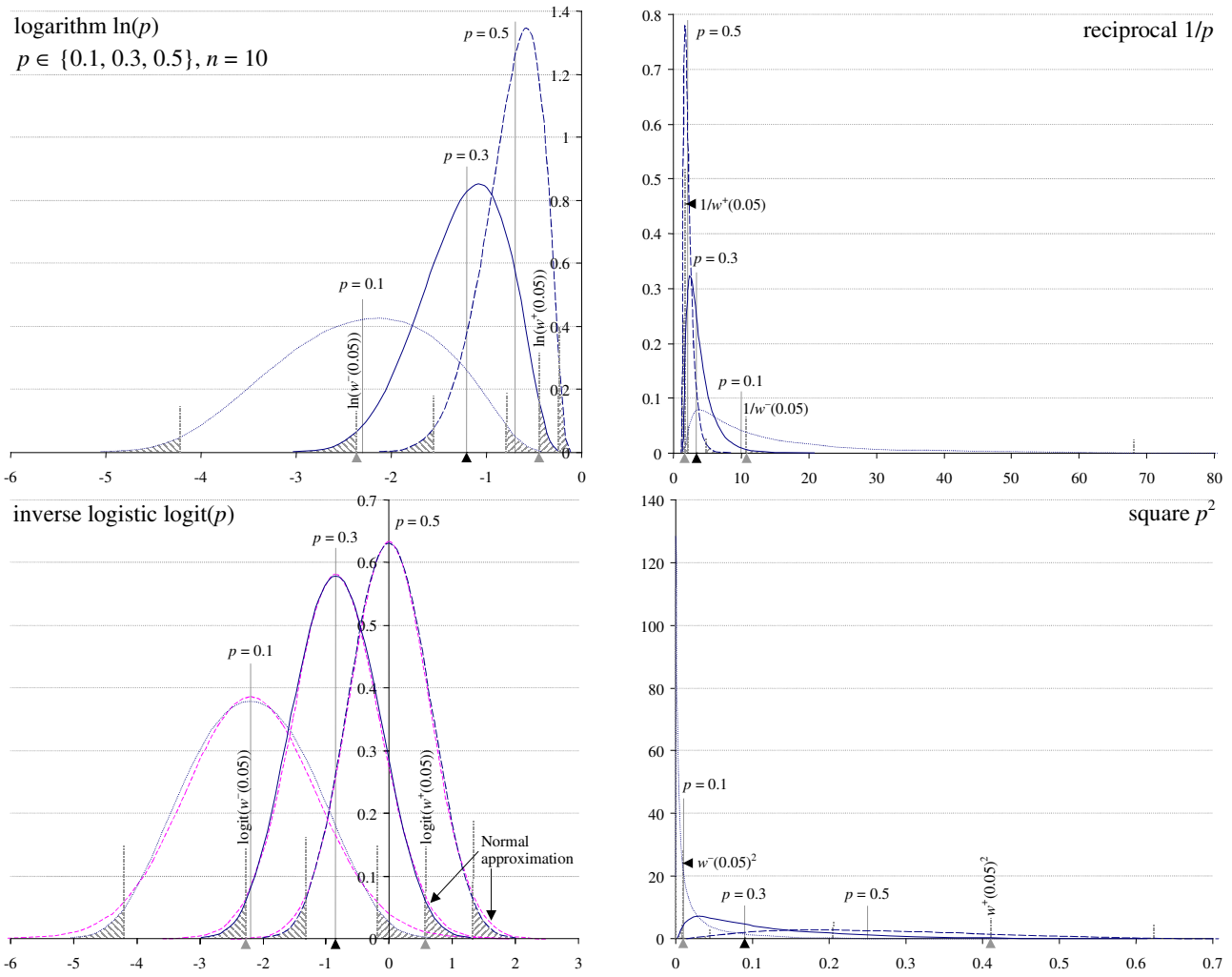

Figure 8. Unit probability density distributions of selected functions of Binomial proportions, $p \in \{0.1, 0.3, 0.5\}$, $n = 10$, tail areas at $\alpha = 0.05$ are labelled for $p = 0.3$. From upper left, clockwise: natural logarithm $\ln(p)$, reciprocal $1/p$, square p^2 and $\text{logit}(p)$, which is approximately Normal.

4.2 Functions of two or more independent proportions

Functions with multiple parameters may be generalised from the Newcombe-Wilson difference interval using a theorem proposed by Zou and Donner (2008). We apply the Bienaymé theorem to inner interval variances (cf. Figure 4) but on different numeric scales.

Zou and Donner's *interval difference theorem* quotes an interval (L, U) for the difference between two independent parameters, $\hat{\theta}_1$ and $\hat{\theta}_2$, each with intervals (l_i, u_i).

$$L = \hat{\theta}_1 - \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - l_1)^2 + (u_2 - \hat{\theta}_2)^2}\ , \text{ and}$$
$$U = \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{(u_1 - \hat{\theta}_1)^2 + (\hat{\theta}_2 - l_2)^2}\ . \qquad (22)$$

If we substitute $\hat{\theta}_1 = p_2$, $\hat{\theta}_2 = p_1$ with respective Wilson intervals, we obtain the Newcombe-Wilson difference interval about d (Section 3.2).

Zou and Donner (2008: 1695) say this equation may be applied to other parameters even if the underlying distribution for each parameter is not Normal, provided that each has 'separate confidence limits that have coverage levels close to nominal.' This is a strong claim worthy of evaluation.

	function	lower bound	upper bound	scale
proportion	p	w^-	w^+	$\boldsymbol{P}$
alternate	$q = 1 - p$	$1 - w^+$	$1 - w^-$	$\boldsymbol{P}$
weighted	kp	kw^-	kw^+	$\boldsymbol{P} \times k$
plus constant	$p + k$	$w^- + k$	$w^+ + k$	$\boldsymbol{P} + k$
reciprocal	$1/p$	$1/w^+$	$1/w^-$	$\boldsymbol{P^{-1}}$
square	p^2	$(w^-)^2$	$(w^+)^2$	$\boldsymbol{P^2}$
logit	$\text{logit}(p)$	$\text{logit}(w^-)$	$\text{logit}(w^+)$	$\mathfrak{R}$
increasing	$fn(p)$	$fn(w^-)$	$fn(w^+)$	
decreasing		$fn(w^+)$	$fn(w^-)$	
non-monotonic		$\min(fn(w^-) \ldots fn(w^+))$	$\max(fn(w^-) \ldots fn(w^+))$	
two independent proportions				
difference (NW)	$p_2 - p_1$	$p_2 - p_1 - \sqrt{(w_1^+ - p_1)^2 + (p_2 - w_2^-)^2}$	$p_2 - p_1 + \sqrt{(p_1 - w_1^-)^2 + (w_2^+ - p_2)^2}$	
sum	Σp_i	$\Sigma p_i - \sqrt{\Sigma (p_i - w_i^-)^2}$	$\Sigma p_i + \sqrt{\Sigma (w_i^+ - p_i)^2}$	
ratio	p_1 / p_2	$\exp(\ln(p_1 / p_2) - \sqrt{(\ln(p_1) - \ln(w_1^-))^2 + (\ln(w_2^+) - \ln(p_2))^2})$	$\exp(\ln(p_1 / p_2) + \sqrt{(\ln(w_1^+) - \ln(p_1))^2 + (\ln(p_2) - \ln(w_2^-))^2})$	
product	$p_1 \times p_2$	$\exp(\ln(p_1 \times p_2) - \sqrt{(\ln(p_1) - \ln(w_1^-))^2 + (\ln(p_2) - \ln(w_2^-))^2})$	$\exp(\ln(p_1 \times p_2) + \sqrt{(\ln(w_1^+) - \ln(p_1))^2 + (\ln(w_2^+) - \ln(p_2))^2})$	

Table 3. Example confidence intervals derived from the Wilson score interval (w^-, w^+).

Parameters $\hat{\theta}_1$ and $\hat{\theta}_2$ may be any monotonic function of p_1 and p_2. Thus we obtain an interval for the *risk ratio*, $r = p_1 / p_2$, since a ratio is a difference on a log scale:

$$\begin{aligned} &\textit{ratio } r = p_1 / p_2, \\ &\textit{log ratio } \ln(r) = \ln(p_1) - \ln(p_2), \\ &(w_r^-, w_r^+) = (-\sqrt{(\ln(w_1^+) - \ln(p_1))^2 + (\ln(p_2) - \ln(w_2^-))^2}\,, \\ &\qquad \sqrt{(\ln(p_1) - \ln(w_1^-))^2 + (\ln(w_2^+) - \ln(p_2))^2}\,). \end{aligned} \tag{23}$$

Finally, we reverse the transformation and reposition the interval by Equation (19):

$$\textit{interval for } r\ (r^-, r^+) = \exp(\ln(r) - (w_r^-, w_r^+)) = (\exp(\ln(r) - w_r^+), \exp(\ln(r) - w_r^-)).$$

The theorem has multiple applications. Trivially, negation is monotonic, so $(\hat{\theta}_1, \hat{\theta}_2) = (p_1, -p_2)$ obtains the interval for the *sum*, $p_1 + p_2$, which may be generalised to a series of terms. The interval for the *product*, $p_1 \times p_2$, may also be obtained via the log transform. See Table 3.

This is just the beginning. In sum and difference formulae, p_1 and p_2 may be replaced by Real parameters. For products and ratios, parameters must be positive.

Thus the *odds ratio* is the ratio of two odds, where $\text{odds}(p) = p / (1 - p)$. This function is monotonic, increasing and yields a positive Real. To obtain an interval for the odds ration, substitute $\text{odds}(p_i)$ for p_i, $\text{odds}(w_i^-)$ for w_i^-, etc. in Equation (23).

Wallis (2021b) derives intervals for *power*, $p_1^{p_2}$, and *logarithm*, $\log_{p_2}(p_1)$, functions. Just as multiplication becomes addition on a log scale, power becomes multiplication:

$$\textit{power } p_1^{p_2} = \exp(\ln(p_1) \times p_2) = \exp(-\exp(\hat{\theta}_1 - \hat{\theta}_2)), \tag{24}$$

where

$$\hat{\theta}_1 = \ln(-\ln(p_1)),\ (l_1, u_1) = (\ln(-\ln(w_1^+)), \ln(-\ln(w_1^-))), \text{ and}$$
$$\hat{\theta}_2 = -\ln(p_2),\ (l_2, u_2) = (-\ln(w_2^+), -\ln(w_2^-)),$$

which we substitute into Equation (22).

Similarly, the logarithm interval is obtained by rewriting it as a ratio of negated logs, $-\ln(p_i)$. The bounds must be swapped for each negative monotonic transform (see Equation (21)).

4.3 Performance

Examining their risk ratio interval, Zou and Donner (2008: 1697) comment that it behaves more consistently to χ^2 than traditional 'delta' methods (Altman *et al*. 2000). However, this is a poor comparison. Delta methods assume that variance is Normal on a logarithmic scale (a 'standard error' problem), and consequently obtain infinite intervals at extremes. Figure 9 plots intervals for risk ratio r and odds ratio o using both approaches. The graph reveals such poor performance for 'delta' methods that they are best retired.

A better method for evaluating interval performance is to compare it with a classical 2 × 2 contingency test. If a repositioned difference interval excludes 0 or a ratio interval excludes 1, the parameters p_1 and p_2 are significantly different. (If $p_1 \neq p_2$ then $p_2 - p_1 \neq 0$ and $p_1 / p_2 \neq 1$.) Intervals for risk and odds ratios, and logarithm ('log ratio'), may be evaluated by this method.

This evaluation is an *inner interval* comparison. Only the interval bound nearest to 0 or 1 is tested. We compare the result with a 'gold standard' Fisher 'exact' test. Appendix 1 plots the performance cost of computing the Bienaymé sum-of-variances theorem (see Figure 4) on these different number scales.

Each ratio method has an additional Type I error cost compared to the simple difference, which also introduces errors. However, these errors are marginal compared to those introduced by not employing a continuity correction, and overall performance is comparable to the χ^2 test.

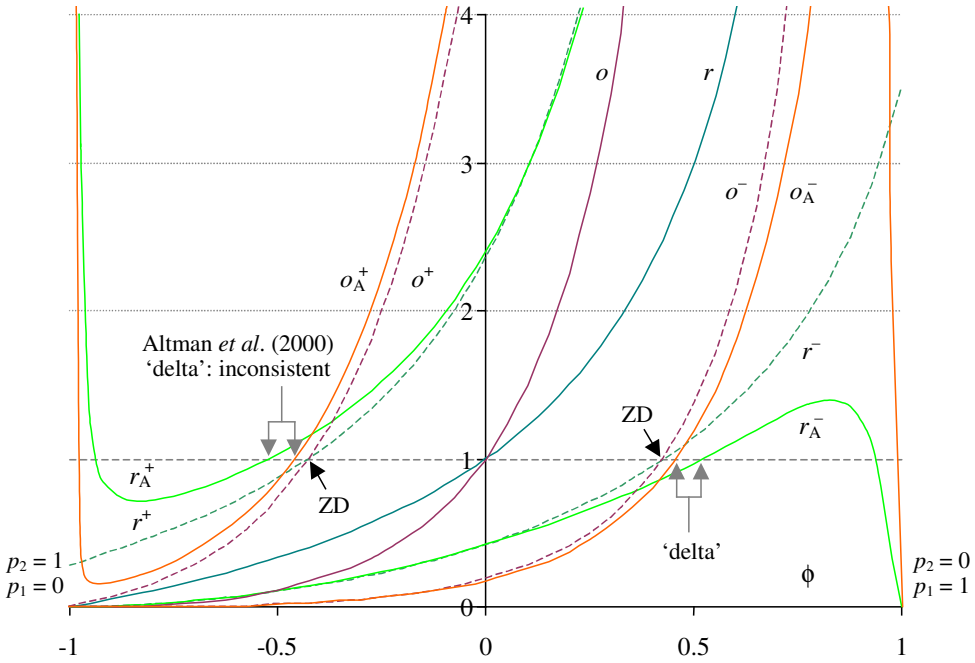

Figure 9. Comparing risk and odds ratio intervals and tests, $n_1 = n_2 = 10$, $\alpha = 0.05$, $\phi = p_1 - p_2$, interpolating over a diagonal matrix. Traditional methods cited by Altman *et al*. (labeled r_A and o_A) are more conservative, catastrophically so where p_1 or $p_2 \to 0$. They are also visibly inconsistent where bounds cross 1, unlike Zou and Donner's methods ('ZD', arrowed).

Interval performance may be ordered from worst to best:

odds ratio (ln(odds(***P***))) < logarithm (ln(–ln(***P***)) < risk ratio (ln(***P***)) < difference (***P***).

In other words, computing intervals by this method tends to introduce more errors when employed on a logarithmic scale (risk ratio) than a probabilistic one, and so on.

4.4 Analytic reduction

These error rates are sufficiently small to permit us to propose a general algebra of interval calculations based on Zou and Donner's theorem, the Wilson score interval (with continuity correction and, potentially, sampling adjustments) and monotonic transformations.

However, one further note of caution is required. The theorem requires that observed parameters are *independent*. If two parameters in a formula are not independent, we must rewrite the formula.

Example 1. Difference of alternate proportions

Suppose we want an interval for d where $p_2 = 1 - p_1$. Since p_2 is determined by p_1, the interval is trivial:

$$d = 1 - 2p_1 \in (1 - 2w_1^+, 1 - 2w_1^-).$$

Example 2. Percentage difference $d^{\%}$

Consider the widely-cited percentage difference:

$$d^{\%} = d / p_1 = (p_2 - p_1) / p_1. \tag{25}$$

If we apply Zou and Donner's theorem first to the difference $d = p_2 - p_1$, and then to the ratio d / p_1, we obtain a poor interval that assumes uncertainty about p_1 twice. See Wallis (2021d).

The solution is to rewrite Equation (25) in a canonical form where each parameter appears once only:

$$d^{\%} = p_2 / p_1 - 1.$$

We compute the ratio interval for p_2 / p_1 with Equation (23), and subtract 1.

5. Effect sizes and meta-tests

More complex applications of this interval algebra can be found in the derivation of intervals for *effect sizes*, which are properties of a contingency table or vector. Traditionally, effect sizes were quoted to estimate the absolute scale of differences in observed distributions, without confidence intervals. Only 'small', 'medium' or 'large' effects were cited (Sheskin 2011: 676).

We will consider two examples, each with multiple applications.

5.1 Unweighted goodness of fit ϕ_p

Unweighted error ϕ_p (Wallis 2021: 229) is a simple 'root mean square' goodness of fit error measure. It sums the difference between observed and expected proportions p_i and P_i. For two-valued tables, it equals the difference $| p_1 - P_1 |$. A small score means a close fit.

$$\textit{unweighted } \phi_p = \sqrt{\sum (p_i - P_i)^2 / 2}\,. \tag{26}$$

Wallis (2021e) obtains an interval for ϕ_p in two steps. First, we obtain an interval for each summed term,

squared difference $\mathrm{sqd}(p_i) = (p_i - P_i)^2/2$.

This function is *non-monotonic*, with a local minimum, 0, at P_i (considered 'given', i.e. constant). We obtain a conservative interval (d_i^-, d_i^+) from

$$\mathrm{sqd}(p_i) \in (d_i^-, d_i^+) = \begin{cases} (\mathrm{sqd}(w_i^-), \mathrm{sqd}(w_i^+)) & \text{if } w_i^- > P_i \\ (\mathrm{sqd}(w_i^+), \mathrm{sqd}(w_i^-)) & \text{if } w_i^+ < P_i \\ (0, \max(\mathrm{sqd}(w_i^-), \mathrm{sqd}(w_i^+))) & \text{otherwise.} \end{cases} \quad (27)$$

Second, we use a modified version of the *sum* formula in Table 3 to account for the fact that a *k*-valued goodness of fit table has *k*-1 degrees of freedom.[8]

$$\phi_p^2 \in (L, U) = (\phi_p^2 - \sqrt{\kappa \Sigma(\mathrm{sqd}(p_i) - d_i^-)^2}\ ,\ \phi_p^2 + \sqrt{\kappa \Sigma(\mathrm{sqd}(p_i) - d_i^+)^2}\ ), \quad (28)$$

where $\kappa = k/(k-1)$. Finally we take the square root to obtain an interval for ϕ_p.

Bowie, Wallis and Aarts (2013) employed goodness of fit ϕ scores to estimate the correlation between the present perfect construction and present and past-marked verb phrases in the *Diachronic Corpus of Present-day Spoken English* (DCPSE), over different genre subcategories. Without an interval formula, they could only observe that present and past scores appeared numerically distinct, but could not identify if they were significantly different.

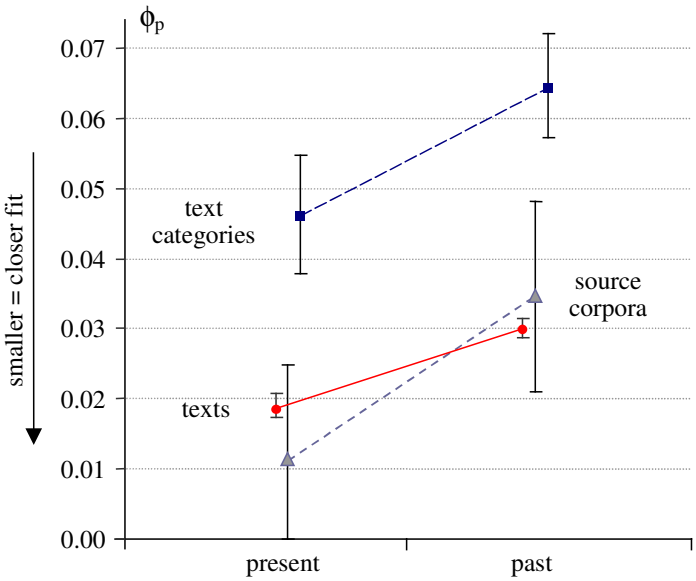

Figure 10. Comparing goodness of fit scores for the present perfect against present and past-marked verb phrase baselines across three different genre categorisations in DCPSE, with 95% confidence intervals obtained from Equation (28). See Wallis (2021e).

Figure 10 plots ϕ_p scores over three different genre categories in the corpus: 2 source corpora (simple 'time': LLC and ICE-GB), 10 text categories (of divergent sizes), and 280 texts. There is a significant difference between each pair of scores. Notably, intervals are smaller where more subcategories are employed.

[8] Both series $\Sigma p_i = 1$ and $\Sigma P_i = 1$. They are not independent. For $k = 2$, $p_2 = 1 - p_1$, and we may employ *signed* $\phi_p = p_1 - P_1$ with interval $(w_1^- - P_1, w_1^+ - P_1)$. Equation (28) obtains the same result if the signed interval is greater than zero.

5.2 Cramér's 2 × 2 ϕ

A more complex derivation is required for Cramér's 2 × 2 ϕ. This is a well-known metric closely related to the chi-square statistic and simple difference (*d* or Δ*P*). Unlike difference, ϕ is bidirectional ('associative'), i.e. the same score is obtained when dependent and independent variables are reversed.

Wallis (2021a: 225) demonstrates how the score can be applied to a grammatical priming analysis. ϕ measures the association between two choices, A and B, in the same utterance or text. Confidence intervals permit us to investigate whether distance attenuates this association, and to compare association ('priming strength') across different grammatical relationships.

A signed ϕ score for 2 × 2 tables (ϕ ∈ [-1, 1]) may be calculated by

$$signed\ 2 \times 2\ \phi = \frac{o_{1,1}o_{2,2} - o_{2,1}o_{1,2}}{\sqrt{o_{1+}o_{2+}o_{+1}o_{+2}}}, \tag{29}$$

using the notation in Table 2.

A confidence interval for ϕ ∈ (ϕ^-, ϕ^+) is found by observing that $\phi^2 = d(x_1) \times d(y_1)$, where $d(x_1)$ is the difference in proportions of x_1 out of X (along the y axis), etc. Differences $d(x_1)$ and $d(y_1)$ are monotonically related, and not independent. Zou and Donner's method is therefore inappropriate (see 4.4 above). Instead, an interval is obtained from the signed geometric mean of the differences:

$$\begin{aligned}\phi^- &= -\text{sign}(d^+(y_1))\sqrt{d^+(y_1) \times d^+(x_1)}\text{, and}\\ \phi^+ &= -\text{sign}(d^-(y_1))\sqrt{d^-(y_1) \times d^-(x_1)},\end{aligned} \tag{30}$$

where $d^+(x_1)$ is the upper Newcombe-Wilson bound of $d(x_1)$, etc. The initial term '$-\text{sign}(d^+(y_1))$', etc. reinstates the sign of ϕ.[9] This interval outperforms standard error based estimates (Bishop, Fienberg and Holland 1975: 387) for reasons similar to those identified in Section 3.3.

5.3. Meta-tests for differences between scores

Zou and Donner's (2008) theorem can be employed to compare any pair of independent scores for significant difference. Provided that we have a good-coverage interval for each parameter, we can substitute them into Equation (22) and create a difference interval and test.

Wallis (2019b; 2021a: 233-260) discusses a series of *meta-tests* that evaluate differences between scores. So-called 'difference of differences' tests may be employed in replication studies, or to examine the impact of changes in experimental design.

A simple example is the *Newcombe-Wilson gradient meta-test*, for comparing two observed differences (gradients). Employing Zou and Donner's theorem we have

$$difference\ of\ differences\ (w_d^-, w_d^+) = (-\sqrt{(w_{d_1}^+)^2 + (w_{d_2}^-)^2}, \sqrt{(w_{d_1}^-)^2 + (w_{d_2}^+)^2}), \tag{31}$$

where $(w_{d_i}^-, w_{d_i}^+)$ is the zero-based Newcombe-Wilson interval width for d_i (Equation (18)). To test if the two differences are significantly different, we compare the difference of differences $d = d_2 - d_1$ with this interval. (We may also reposition the interval about d for plotting purposes.)

We can use this method to compare any two independent properties, such as risk ratios or effect sizes (Figure 10). It may also be generalised across multiple degrees of freedom using χ^2.

[9] The equation fails if terms within the square root have different signs, which can occur near 0. Zero or the negated arithmetic mean may be substituted.

As a corollary, we can employ the Wilson interval comparison heuristic (Equation (20)) to visually compare any two properties for which this test is legitimate. Note that in Figure 10, two out of the three interval pairs do not overlap.

6. Conclusions

Lab researchers learn that measurement is embodied with an estimate of accuracy. Likewise, observed means, proportions and probabilities are necessarily qualified estimates, their accuracy determined by sample size and method. When we plot or cite statistics derived from data we should account for sampling uncertainty.

The solution is to deploy *confidence intervals* identifying the most likely range of the true population value. This is not to be confused with 'scatter', i.e. observed within-sample spread.

Accurate confidence intervals on a Binomial proportion $p \in \boldsymbol{P} = [0, 1]$ are asymmetric, except at $p = 0.5$, due to the presence of bounds at 0 and 1. We have previously shown that in computing Binomial intervals, the Wilson score interval with continuity-correction performs almost as well as 'exact' methods. The principal advantage of Wilson-derived methods is not merely efficient computation. They can be readily corrected for continuity, finite population and random-text sampling, by first adjusting the Wilson formula for each term.

Armed with a reliable model for estimating the sampling error of a single proportion, p, a wide range of possibilities emerge through algebra and some basic theorems. We have shown that we may

- transform the interval to other mathematical scales, such as $1/p$, $\ln(p)$, or $\text{logit}(p)$,
- compute intervals for differences between independent proportions $p_2 - p_1$, and other mathematical relations (sum, ratio, product, power, etc.),
- create intervals for other properties derived from these, such as the odds ratio, percentage difference and effect sizes, and
- create meta-tests, such as the difference in differences gradient test.

We demonstrated that despite the fact that summation of variance is performed on different scales, their inner intervals have comparable performance to classical contingency tests (χ^2, Fisher).

Confidence intervals have long been seen as 'not proper' statistics, because their conventional treatment depended on a fundamentally erroneous standard error model. Even when a Normal model is legitimate (such as testing for $d \neq 0$ with a Gaussian interval on $\hat{p}$), it has limited generality. It is time to restore confidence intervals to their rightful place.

The following are recommended. Graphed proportions, differences and other scores should be plotted with confidence intervals wherever feasible, and variables should be cited with bounds at selected error levels. This approach should replace the long-criticised practice of 'p value' citation. Aside from the logical error involved in comparing error levels (p values) between experiments, they do not encourage an appreciation of the uncertainty of observations.

Given the errors we have seen with difference, ratio and ϕ intervals, it would be surprising if other formulae and algorithms were immune to the 'standard error'. Their internal workings should be reviewed, and where this error has appeared, substituted with Wilson-based alternatives.

References

Altman, D.G., Machin, D., Bryant, T.N. & Gardner, M.J. (2000). *Statistics with Confidence* (2nd ed.). BMJ Books: Bristol.

Beyene, J. & Moineddin, R. (2005). Methods for confidence interval estimation of a ratio parameter with application to location quotients. *BMC Medical Research Methodology, 5*(32). doi:10.1186/1471-2288-5-32.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Bowie, J., S.A. Wallis & B. Aarts (2013). The perfect in spoken English. In B. Aarts, J. Close, G. Leech & S.A. Wallis (eds.) *The Verb Phrase in English*. Cambridge: Cambridge University Press.

Brown, L.D., Cai, T.T. & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science 16*(2), 101-133. doi:10.1214/ss/1009213286.

Clopper, C.J. & Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the Binomial, *Biometrika 26*(4), 404-413. doi:10.2307/2331986.

Gries, S.Th. (2013). 50-something years of work on collocations: What is or should be next… *International Journal of Corpus Linguistics 18*(1). 137–165.

Nelson, G., B. Aarts & S.A. Wallis (2002). *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Varieties of English Around the World series. Amsterdam: John Benjamins.

Newcombe, R.G. (1998a). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine, 17*, 857-872. doi:10.1.1.408.7107.

Newcombe, R.G. (1998b). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, *17*, 873-890. doi:10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0

Sheskin, D.J. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures* (5th ed.). Boca Raton, Fl: CRC Press.

Wallis, S.A. (2013). Binomial confidence intervals and contingency tests. *Journal of Quantitative Linguistics 20*(3), 178-208. doi:10.1080/09296174.2013.799918.

Wallis, S.A. (2019a). Confidence intervals on pairwise ϕ statistics. *corp.ling.stats*. London: Survey of English Usage. Available from https://corplingstats.wordpress.com/2019/12/17/conf.

Wallis, S.A. (2019b). Comparing χ^2 tables for separability of distribution and effect. Meta-tests for comparing homogeneity and goodness of fit contingency test outcomes. *Journal of Quantitative Linguistics 26*(4), 330-355. doi:10.1080/09296174.2018.1496537.

Wallis, S.A. (2020a). Confidence intervals and replication intervals. *corp.ling.stats*. London: Survey of English Usage. Available from https://corplingstats.wordpress.com/2020/10/01/rep.

Wallis, S.A. (2020b). Plotting the Clopper-Pearson distribution. *corp.ling.stats*. London: Survey of English Usage. Available from https://corplingstats.wordpress.com/2020/04/25/plotting.

Wallis, S.A. (2021a). *Statistics in Corpus Linguistics Research: a new approach*. New York and Abingdon: Routledge.

Wallis, S.A. (2021b). Confidence intervals on powers and logs. *corp.ling.stats*. London: Survey of English Usage. Available from https://corplingstats.wordpress.com/2021/11/29/powers.

Wallis, S.A. (2021c). Evaluating the performance of risk and odds ratio tests. *corp.ling.stats*. London: Survey of English Usage. Available from https://corplingstats.wordpress.com/2021/11/15/evaluating.

Wallis, S.A. (2021d). Confidence intervals on percentage difference – a cautionary tale. *corp.ling.stats*. London: Survey of English Usage. Available from https://corplingstats.wordpress.com/2021/09/26/percentage-diff.

Wallis, S.A. (2021e). Confidence intervals on goodness of fit ϕ scores. *corp.ling.stats*. London: Survey of English Usage. Available from https://corplingstats.wordpress.com/2021/09/08/gof-phi-intervals.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22*(158), 209-212. doi:10.1080/01621459.1927.10502953.

Yates, F. (1934). Contingency tables involving small numbers and the chi-square test. *Journal of the Royal Statistical Society*, *1*(2), 217-235. doi:10.2307/2983604.

Zar, J.H. (2010). *Biostatistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Zou, G.Y. & Donner, A. (2008). Construction of confidence limits about effect measures: A general approach. *Statistics in Medicine, 27*(10), 1693-1702. doi:10.1002/sim.3887.

Appendix 1

To evaluate their consistency with classical contingency testing, we can treat difference and ratio intervals as a type of 2 × 2 test. In essence, such tests compare if $p_1 \neq p_2$, where '≠' is interpreted as 'is significantly different from'. A difference test compares if $d = p_2 - p_1 \neq 0$, whereas a ratio test compares if the risk ratio $r = p_1 / p_2$, logarithm $l = \ln(p_1) / \ln(p_2)$, or odds ratio $o = \text{odds}(p_1) / \text{odds}(p_2) \neq 1$.

For $n_1 = n_2$ from 1 to 200, we enumerate all discrete matrix combinations and compare test performance. Where the test differs from the Fisher 'exact' test, we sum the additional error, weighted by the prior probability of that cell combination occurring (the *Fisher weight*). See Wallis (2021c). We repeat the exercise for $n_1 = 5n_2$.

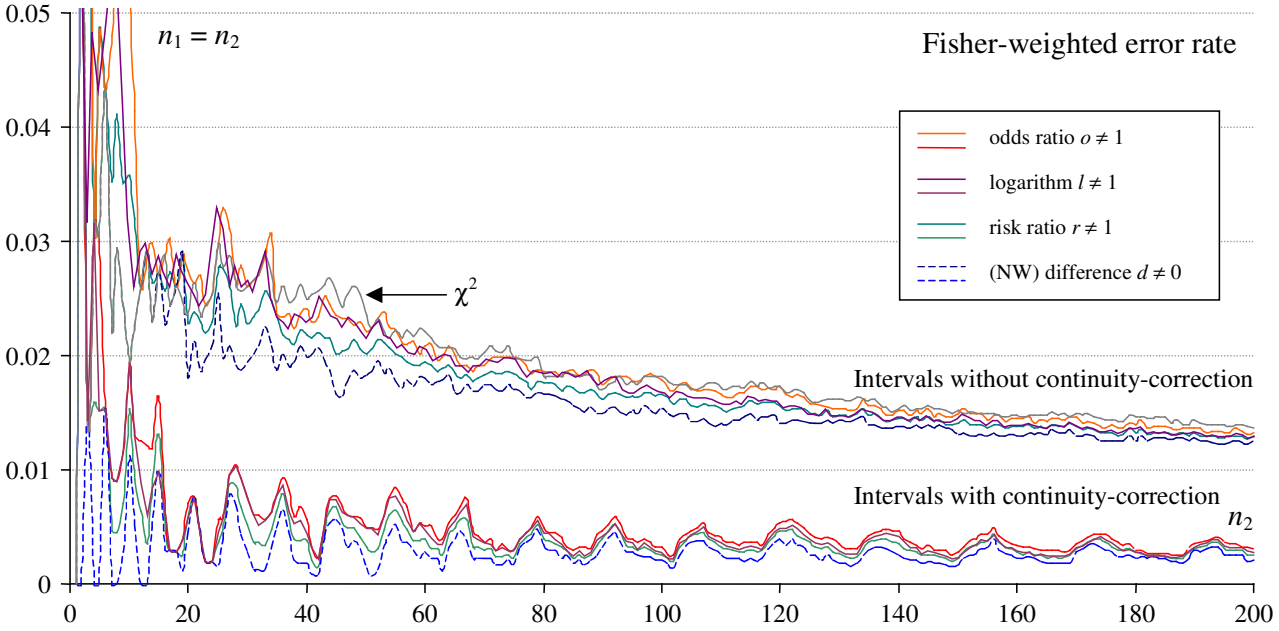

Figure A1. Fisher-weighted error rates for Type I errors against the Fisher 'exact' test, computed for values of $n_2 \in \{1, 2, \ldots 200\}$, $\alpha = 0.05$, with equal-sized samples. Yates's χ^2 test obtains no Type I errors in this case.

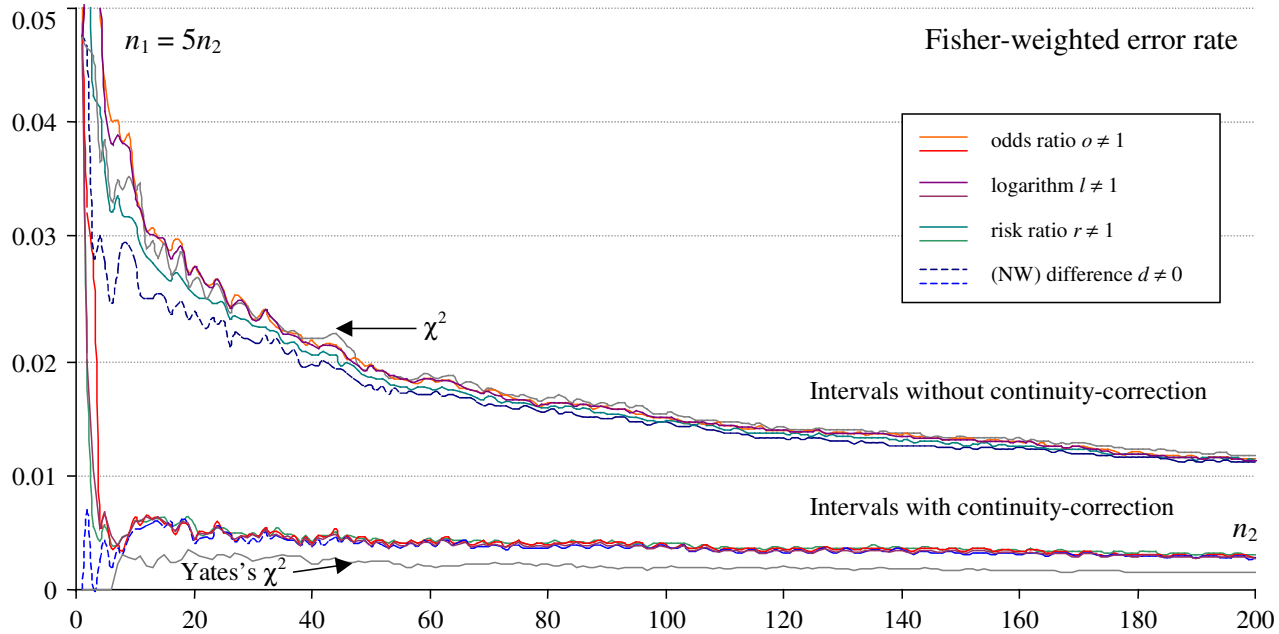

Figure A2. Fisher-weighted error rates for Type I errors against the Fisher 'exact' test, computed for values of $n_2 \in \{1, 2, \ldots 200\}$, $\alpha = 0.05$, with unequal-sized samples ($n_1 = 5n_2$).