Further evaluation of Binomial confidence intervals and difference intervals[†]

Sean Wallis, Survey of English Usage, UCL
s.wallis@ucl.ac.uk
February 2020

Abstract
Wallis (2013) provides an account of an empirical evaluation of Binomial confidence intervals and contingency test formulae. The main take-home message of that article was that it is possible to evaluate statistical methods objectively and provide advice to researchers that is based on an objective computational assessment.

In this article we develop the evaluation of that article further by re-weighting estimates of error using Binomial and Fisher weighting, which is equivalent to an 'exhaustive Monte-Carlo simulation'. We also develop an argument concerning key attributes of difference intervals: that we are not merely concerned with when differences are zero (conventionally equivalent to a significance test) but also accurate estimation when difference may be non-zero (necessary for plotting data and comparing differences).

Keywords: Binomial confidence interval, χ^2 test, difference interval, evaluation.

1. Introduction

All statistical procedures may be evaluated in terms of the rate of two distinct types of error.

- **Type I errors** (false positives): this is evidence of so-called 'radical' or 'anti-conservative' behaviour, i.e. *rejecting* null hypotheses which should not have been rejected, and
- **Type II errors** (false negatives): this is evidence of 'conservative' behaviour, i.e. *retaining* or *failing to reject* null hypotheses unnecessarily.

It is customary to treat these errors separately because the consequences of rejecting and retaining a null hypothesis are qualitatively distinct.

In classical experiments, whether in the lab or with corpora, researchers should err on the side of caution and risk Type II errors but not Type I errors. The premise is that it is safer to avoid investing research effort in a dead end – by yourself or others – rather than to find out later that you have wasted time and resources.

Note, however, that this is not a universal rule. If you were offering a potentially life-saving experimental drug to someone who is expected to otherwise die, you might risk Type I errors (that the drug had no significant effect, i.e. it did not work). This issue has arisen recently in clinical trial of the Ebola vaccine (Calain 2018). We must still attempt to weigh up the risk of side-effects.

Secondly, we need to decide on a 'gold standard' criterion. We need an independent measure of 'correctness'. A test evaluation can have one of four possible outcomes (Table 1).

	'Gold standard' test	
Test evaluation	True (significant)	False (non-significant)
True ('significant')	✓	Type I
False ('non-significant')	Type II	✓

Table 1. Comparing a test under evaluation with a 'gold standard' test.

Where the test we are evaluating is 'significant' and the gold standard test is significant, the tests are consistent, and using the evaluated test does not generate an error. Likewise, where the test evaluation and gold standard test both obtain a non-significant result, the methods perform equally.

[†] This article develops on Wallis (2013), Binomial confidence intervals and contingency tests. *Journal of Quantitative Linguistics 20*(3), 178-208, available from https://www.tandfonline.com.

But in other cases we have either Type I or Type II errors. The idea is we can add up these two types of error separately and thereby compare test performances.

This method is an effective one for evaluating tests. But it is not sufficient to evaluate *intervals*. This is because with an interval we also wish to know how far results *diverge*. We want to know that if we plot a confidence interval over many values, whether it is accurate for all values of p, and not misleading for certain ones (say, when p is close to 0 or 1). We turn to this question next.

2. Evaluating Binomial intervals

For a single proportion test or interval we will use *the exact Binomial distribution* as a gold standard. In this section we will recapitulate some of the method of Wallis (2013).

The Binomial intervals we consider are:

- the asymptotic 'Wald' Gaussian or single-proportion z interval (sometimes called 'standard error'),
- the Wilson score interval, which correctly inverts the Gaussian (see Wallis 2013, Newcombe 1998a), equivalent to a 2×1 χ^2 goodness of fit test,
- the same interval with a continuity correction applied (equivalent to a 2×1 Yates' χ^2), and
- an inverted log-likelihood G^2 interval.

Other suggested intervals have largely been dispensed with by earlier work, see, for example Newcombe (1998a).

To evaluate intervals, say the Wilson (w^-, w^+), we only need to evaluate one interval bound, thanks to the reflexivity principle $q = 1 - p$, so the lower bound of q is the upper bound of p, etc.

We will focus on the lower bound – w^- in this case – for any value of $p = f / n$. We substitute $P = w^-$ into the cumulative Binomial function (Equation (2)). The Binomial function for the probability of any particular value r from 0 to n is

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

where P is the population probability. The cumulative function simply adds up the Binomial distribution from $r = r_1$ to r_2 inclusive:

$$\textit{Cumulative Binomial probability } B(r_1, r_2; n, P) \equiv \sum_{r=r_1}^{r_2} B(r; n, P). \tag{2}$$

We want to calculate errors for the lower interval bound ($P = w^-$) of different formulae for intervals for observed p. Therefore we select the upper tail of the Binomial interval, i.e. the area from f to n for a true value P. This is calculated by

$$\textit{upper tail} = B(f, n; n, P).$$

Wallis (2013) notes that the Clopper-Pearson method employs a computational search procedure to sum the upper tail for $p = f / n$, by summing from f to n to find P where the following holds:

$$B(f, n; n, P) = \alpha/2. \tag{3}$$

where P is the lower bound of the interval under evaluation.

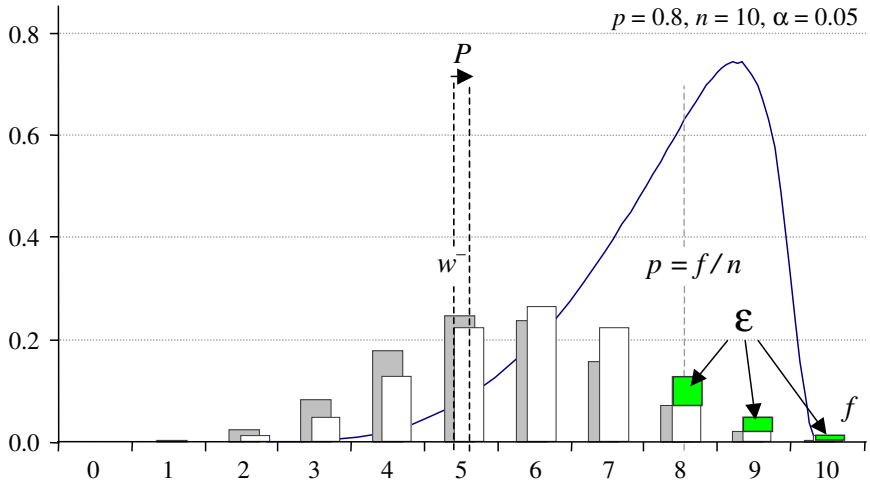

Figure 1. Error ε = difference area under tail when P has moved.

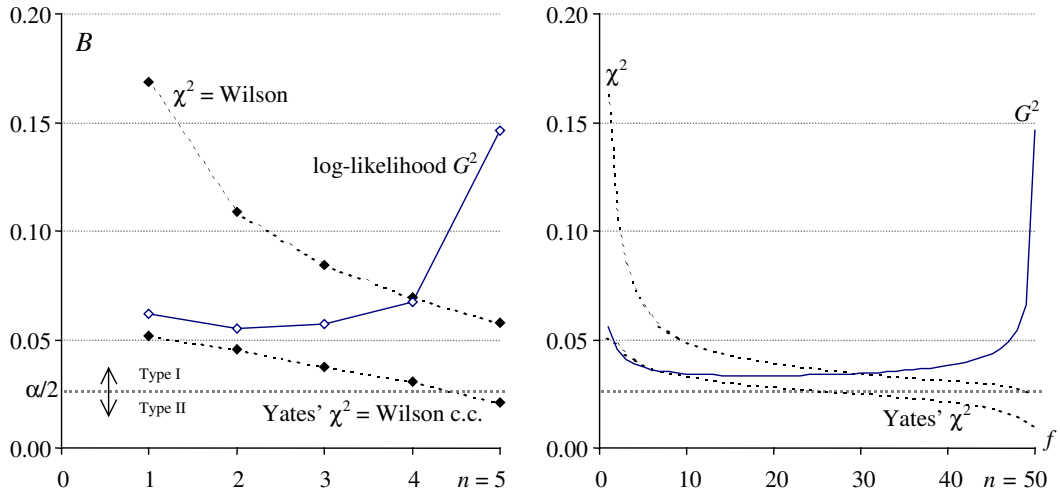

Figure 2. Binomial upper 'tail' area B for $f = 1$ to n, $n = 5$ and 50, $\alpha = 0.05$. Error $\varepsilon = B - \alpha/2$.

To measure the size of an error we use a similar method. We calculate an error term ε, representing the erroneous area relative to the correct tail B. All we need to do is rearrange the formula:

$$\varepsilon = B(f, n; n, P) - \alpha/2. \quad (4)$$

The error ε is the area between the two Binomial distributions in Figure 1. The first (front) is calculated with $P = w^-$, the second (shaded, behind), using the exact Clopper-Pearson interval bound.

The basic idea is that if w^- is at the correct Binomial location, ε will be zero. In practice w^- tends to be slightly anti-conservative, creating an error. Indeed, by plotting the actual distributions, Wallis (2018) shows that the Clopper-Pearson and continuity-corrected Wilson perform very similarly.

2.1 Estimating errors from single interval curves

Let us first attempt to see the scale of interval performance errors. We employ Equation (4) to obtain a Binomial error rate B relative to the target value of $\alpha/2$ (here, 0.025). We plot B for $n = 5$ and $n = 50$ in Figure 2. We mark the ideal error rate in the same graph (the straight line marked $\alpha/2$).

Positive differences above the dotted line in Figure 2 represent the probability of a Type I error (accepting a false alternate hypothesis, see above). Negative differences represent the chance of a Type II error (retaining a false null hypothesis). The graphs tell us that if we know f (or p) we can identify the functions that perform best for any rate $p = f / n$.

The next step is to aggregate these errors to obtain a single error score.

One way we could do this is to simply take the arithmetic mean of each error. A simple average assumes that *the prior chance of an error occurring* is constant for all values of p.

However, the probability of P being less than p is not constant. It is more probable that $P < p$ if $p = 1$ than if $p = 0.5$(say).

A better approach calculates a *weighted average*, with each term weighted by p or f, as in Equation (5). We weight each term by the simple combinatorial probability, $B(f, n, 0.5)$. This assumes we know nothing about why $p = f / n$ was selected but recognises that there are more combinations that can yield scores towards the middle of the range.[1]

[1] An alternative method is to take the combinatorial probability of $P < p$ into account, using the cumulative Binomial $B(0, f\text{-}1; n, 0.5)$ instead of f / n $(n+1)$. This does not make a difference to the order however, because the differences in performance are quite clear.

$$\textit{Type I error } \varepsilon_{\mathrm{I}} = \frac{\sum f \min(\varepsilon_f, 0) B(f, n, 0.5)}{n(n+1)},$$

$$\textit{Type II error } \varepsilon_{\mathrm{II}} = \frac{\sum f \min(-\varepsilon_f, 0) B(f, n, 0.5)}{n(n+1)}. \quad (5)$$

These weights do not sum to 1, so the overall scores are reduced, but they are reduced equally. In Table 2 we therefore divide by the sum of the weights. On this assessment, uncorrected 2 × 1 χ^2 or Wilson score interval performs the worst, followed by log-likelihood, with Yates' χ^2 or the continuity-corrected Wilson interval doing the best.

Here we evaluate confidence intervals in terms of their 'scalar accuracy', i.e. the degree to which they are measurably inaccurate on a probability scale. For this simple assessment we took into account the size of the errors as well as their chance of occurrence.

f	p	Binomial	χ^2	Yates'	G^2
0	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.2000	0.0050	0.0362	0.0105	0.0126
2	0.4000	0.0528	0.1176	0.0726	0.0807
3	0.6000	0.1466	0.2307	0.1704	0.1991
4	0.8000	0.2836	0.3755	0.2988	0.3718
5	1.0000	0.4782	0.5655	0.4629	0.6810
Error rates		Type I	0.0653	0.0129	0.0401
		Type II	0.0000	0.0002	0.0000

Table 2. Lower bounds obtained by search for Binomial, χ^2, Yates' χ^2 and log-likelihood G^2, and scaled weighted sum error rates, for $n = 5$, $\alpha = 0.05$.

2.2 Evaluating 2 × 1 tests and simple confidence intervals

Table 2 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 χ^2, or, where possible, by substituting the relevant Wilson formula.

Overall, log-likelihood is inferior to Yates' χ^2 for small p, because, as we have seen the lower bound has a large number of Type I errors as p approaches 1 (or $f \to n$).

With $n = 5$, using the combinatorial assessment, Yates' χ^2 underestimates the lower bound (and therefore the interval) on approximately 0.01% of occasions. This error falls for $n = 50$. Yates'

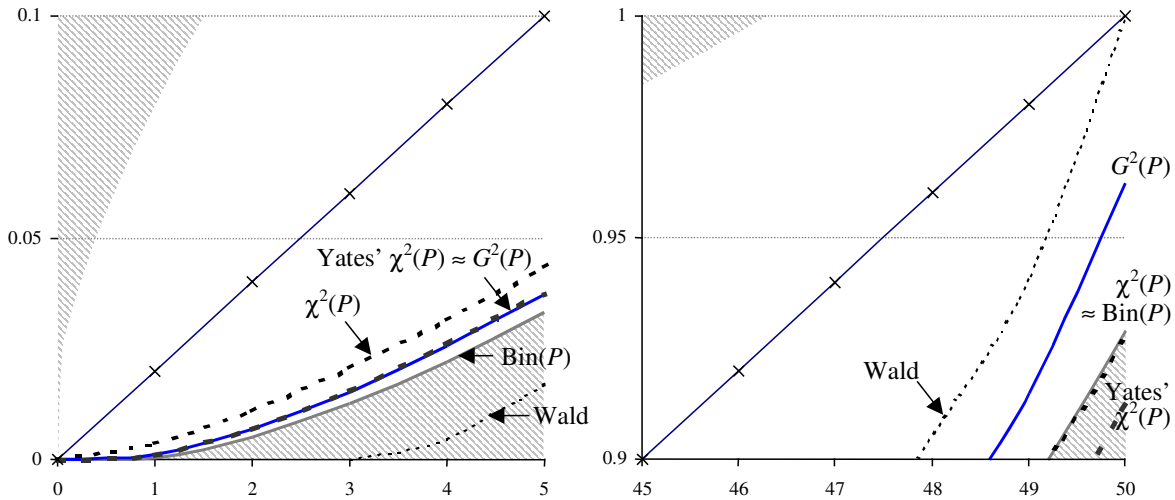

Figure 3. Plotting lower bound error estimates for extremes of p, $n = 50$, $\alpha = 0.05$.
Gloss: $\chi^2(P)$ = Wilson(p), Yates' $\chi^2(P)$ = Wilson(p) interval with continuity correction and $G^2(P)$ = log-likelihood interval for p. Bin(P) = Binomial(P) (the ideal result, shaded).

formula can exceed the Binomial interval at $x = n$, obtaining Type II errors, but this effect is minor.

These results reinforce two points. It is worthwhile employing continuity-corrected formulae with small n, and this method of interval estimation is robust. As we might expect, as n increases, the effect of (and need for) this correction reduces.

However, this still leaves open the question as to what happens at extremes of p. To get an idea of the behaviour of different formulae we can plot lower interval bounds at each end of the range for $n = 50$ (Figure 3). The target value is the Binomial, 'Bin(P)', which we have shaded.

In summary, the figure reveals:

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

Even with large n, the 'Wald' confidence interval is unreliable at probability extremes. Log-likelihood (labelled '$G^2(P)$') performs quite well for the lower bound of small p (Figure 3, left), but poorly for the lower bound of high p (which equals the upper bound for small p). So even if it may be considered acceptable for testing purposes, it is not reliable for estimating bounds at skewed p.

The rate of Type I errors for standard χ^2, Yates' χ^2 and log-likelihood are 0.0095, 0.0014 and 0.0183 respectively, maintaining the same performance distinctions we found for small n. Yates' χ^2 has a Type II error rate of 0.0034, a three-fold increase from $n = 5$. In Section 3.2, we evaluate intervals against the exact Binomial for $n = 1$ to 100 (see Figure 6), counting errors assuming intervals are independent. This confirms the pattern identified above.

3. Evaluating 2 × 2 tests and difference intervals

So far we have evaluated the performance of confidence intervals for a single proportion, mirroring a 2 × 1 'goodness of fit' contingency test where the population value is known. We next consider the performance of confidence intervals in combination, i.e. 2 × 2 tests. We will evaluate the following intervals and tests.

- 2 × 2 χ^2 homogeneity test, equivalent to a two-independent proportion z interval based on a pooled probability estimate $\hat{p}$ (Sheskin, 2011: 655),
- Yates' continuity-corrected 2 × 2 χ^2 test for homogeneity,
- 2 × 2 log-likelihood interval G^2 and interval,
- Sheskin's (2011: 658) z test for two independent proportions from independent populations test and interval,
- Newcombe's (1998b) 'Method 10' Wilson-based interval, and
- Newcombe's 'Method 11' Wilson-based interval with a continuity correction.

The last three intervals are derived from single proportion intervals (the Wald, Wilson, and continuity-corrected Wilson interval respectively) and employ a Bienaymé approximation. See Equation (8) below.

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

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

Table 3. 2 × 2 table and notation.

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

We can divide 2 × 2 tests into two different sub-tests: those where each probability is obtained from samples drawn from the same population (Section 3.1 below) and from independent populations (Section 3.2). Section 4 compares the performance of these baseline tests and discusses the implications of these two models.

As well as counting every possible cell combination (obtaining a 'cell count' error rate), we can count *every way of obtaining each cell combination*. Just as two coin tosses can obtain a single head by either first throwing a head, then a tail, or by throwing a tail, then a head, any cell frequency total can be obtained by different Binomial sequences. (The obvious exception is all-heads or all-tails, each of which can only be arrived at by one sequence.) The second type, based on a combinatorial calculation, uses the Fisher statistic, which we come to next.

3.1 Evaluating 2 × 2 tests against Fisher's test

Fisher's exact test (Sheskin 2011: 649) uses a combinatorial approach to compute the exact probability of a particular observed 2 × 2 table occurring by chance.

$$p_{Fisher}(a, b, c, d) = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{n!a!b!c!d!}, \tag{6}$$

where a, b, c and d represent the values in the 2 × 2 table (Table 3) and $n = a+b+c+d$. The probability p_{Fisher} is the chance of a *particular* pattern occurring. A χ^2 test, on the other hand, tests whether the observed pattern *or a more extreme pattern* is likely to have occurred by chance. To compute an equivalent Fisher-based test we need to perform a summation over these patterns, to obtain p_{FSum} from Equation (7).

$$p_{FSum}(a, b, c, d) = \begin{cases} \sum_{i=0}^{\min(b,c)} p_{Fisher}(a+i, b-i, c-i, d+i) & \text{if } \frac{a}{a+b} > \frac{c}{c+d} \\ \sum_{i=0}^{\min(a,d)} p_{Fisher}(a-i, b+i, c+i, d-i) & \text{otherwise.} \end{cases} \tag{7}$$

Sheskin notes that the Fisher test assumes that 'both the row and column sums are predetermined by

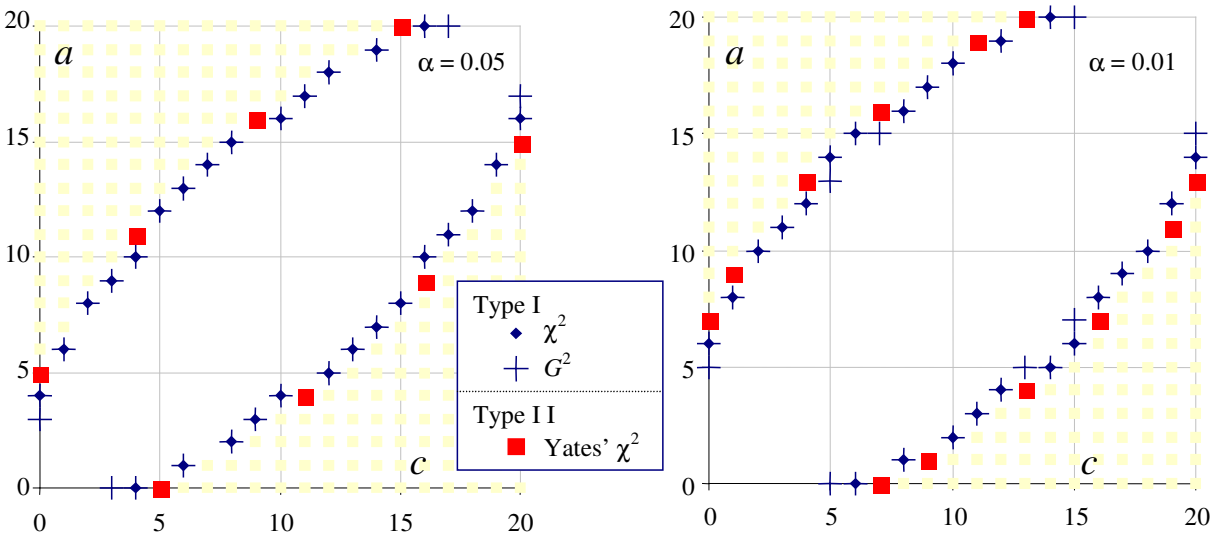

Figure 4. Evaluating χ^2, Yates' χ^2 and log-likelihood G^2 against Fisher's sum for error levels $\alpha = 0.05$ (left) and $\alpha = 0.01$ (right). The area outside the curve is considered significant by all tests, only discrepancies are marked. (After Wallis 2013).

the researcher.' Both column totals $a + b$ and $c + d$, and row totals $a + c$ and $b + d$, are constant, thereby legitimating this summation.

In *ex post facto* corpus analysis, this corresponds to a situation where samples are taken from the same population and the independent variable (as well as the dependent variable) represents a free choice by the speaker. Either value of the independent variable (IV) may be uttered by the same speaker or appear in the same source text. Alternative tests are the 2×2 χ^2 test (including Yates' test) and log-likelihood test. These tests can be translated into confidence intervals on the difference between p_1 and p_2.

It can also be used where samples are drawn from different populations (or subject groups) *but we want to test if they have the same population mean.* So the Fisher model can be used in most straightforward 2×2 comparisons. Where it does not work well is where we assume as part of the test that populations have a different mean – for example to compare an observed difference in proportions with a constant difference, D. The same applies to some meta-testing conditions (Wallis 2019). We discuss this in the next section.

Wallis (2013) evaluated tests by counting Type I and II errors for conditions where the tests do not agree with the result obtained by Fisher's sum test. Figure 4 plots a map of all tables of the form $[[a, b] [c, d]]$ for all integer values of a, b, c, d where $n_1 = a + b = 20$ and $n_2 = c + d = 20$.

We can see that in both cases, there are slightly more errors generated by G^2 than χ^2. Yates' χ^2 has no Type I errors but allows some Type II errors. For experimental purposes, this means that Yates' test performs best of all.

We first repeat the evaluation of Wallis (2013). We use a cell count error rate: the proportion of cell combinations that generate an error. We calculate tables for a given α and obtain the error rate. Figure 9 on page 13 plots error rates for evenly balanced patterns ($n_1 = n_2$) up to 100, testing 174,275 unique points. Yates' test has the lowest overall discrepancies – and these are solely Type II errors.[2]

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

However, counting cells has an obvious weakness. *Not every cell combination is equally likely to occur*. There is only one way of obtaining a cell with a cell frequency of 0 out of 10, but 10 ways of obtaining 1 out of 10. Ideally, instead of counting cells, it would be preferable to weight each erroneous cell by the chance that it would occur.[3]

Since we are examining a 2×2 table, we can use the Fisher statistic to estimate the chance of a particular frequency pattern occuring. For every erroneous cell we add the Fisher probability, p_{Fisher} (Equation (6)), and then divide the total by the Fisher sum, n+1 (where n is the total cell frequency, $n_1 + n_2$). This step is analogous to weighting scores by the simple Binomial combinatorial statistic (see Section 3 above).

The result of this evaluation is plotted in Figure 10 in the Appendix. It confirms the rank ordering observed in Wallis (2013), but it separates the performance of each test evaluation more clearly and consistently.

Sheskin's independent-population z test performs particularly badly; log-likelihood G^2 is outperformed by unmodified χ^2; and Newcombe-Wilson is better still. The Newcombe-Wilson test

[2] The jagged nature of each line is due to the fact that each table consists of a discrete matrix, but the interval estimators are continuous.

[3] One can approximate the same result with a 'Monte Carlo' method, see Newcombe 1998b. This uses a random number generator to pick combinations in a table. Our evaluation is exhaustive: we do not generate a random subset of cases for evaluation purposes, we generate all of them.

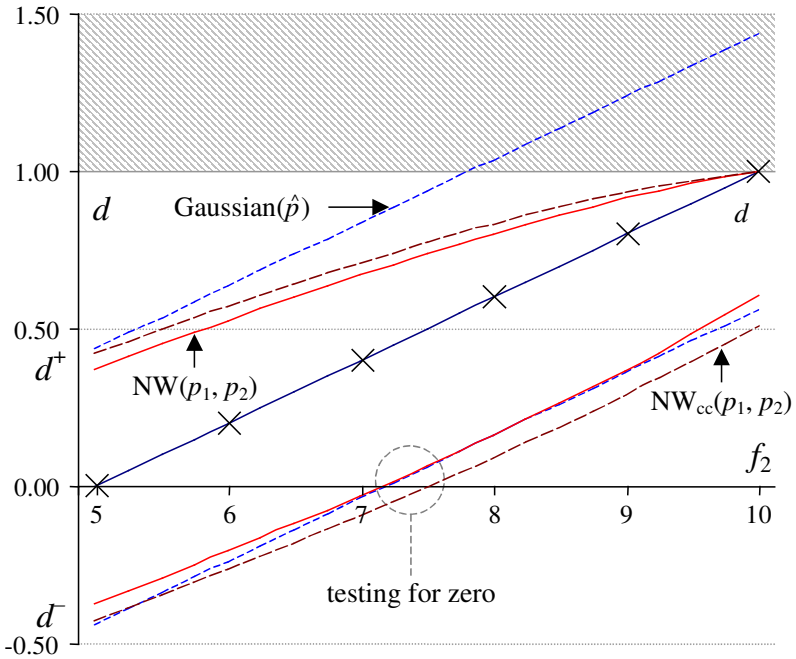

Figure 5. Plotting Gaussian and Newcombe-Wilson intervals (corrected for continuity and uncorrected) for positive $d = p_2 - p_1$ with frequencies $f_1 = 10 - f_2, f_2 \in (5, 10)$, $n_1, n_2 = 10$ and $\alpha = 0.05$. From Wallis (2021).

with continuity-correction approaches the performance of Yates' χ^2 test (in terms of total errors), even though Yates' test is designed as a substitute for Fisher's test.

The resulting curves correspond to additional error rates, so Type I errors correspond to tests that would be considered 'significant' but should be non-significant. In effect, they increase the actual error rate by the same amount. Thus, examining Fisher-weighted error rates (Figure 10), the Newcombe-Wilson error for $n_2 > 10$ never exceeds 0.01 ($n_1 = n_2$), whereas for $n_1 = 5n_2$, the additional error rate declines more steadily from slightly in excess of 0.0125.

As we shall see in Section 7, much of this cost is attributable to the difference between Fisher and Binomial models (see Figure 8).

3.2 Evaluating 2 × 2 tests against a paired exact Binomial test

An alternative baseline test is the paired exact Binomial test.

Why should we consider this option? First, the exact Binomial is the optimum baseline test for tests based on the *single proportion* or interval (Section 2). This approach has many advantages for the purposes of generalisation and visualisation.

Possible implementations include Sheskin's (2011) z test for two independent population proportions and Newcombe's Wilson-based interval and test (Newcombe 1998b). Wallis (2021) demonstrates that whereas the Fisher model is fine for evaluating differences from *zero*, it fails entirely should we wish to compare a difference in proportions d with an arbitrary non-zero proportion D.

Figure 5 shows that using the Gaussian interval based on a pooled probability $\hat{p}$, i.e. employing the same model as a standard 2 × 2 χ^2, we obtain a fixed-width interval that overshoots the difference range $d \in [-1, 1]$. Adding a continuity correction would be a waste of time!

By contrast, both of Newcombe's intervals perform well. The Binomial difference interval performs similarly.

Although our testing regime is intended to compare the performance of difference intervals, this section can also be thought of as an exhaustive evaluation of the single interval, because both the interval under evaluation and the control (Binomial) undergo the same Bienaymé interval manipulation method. See Zou and Donner (2008) for more discussion on this.

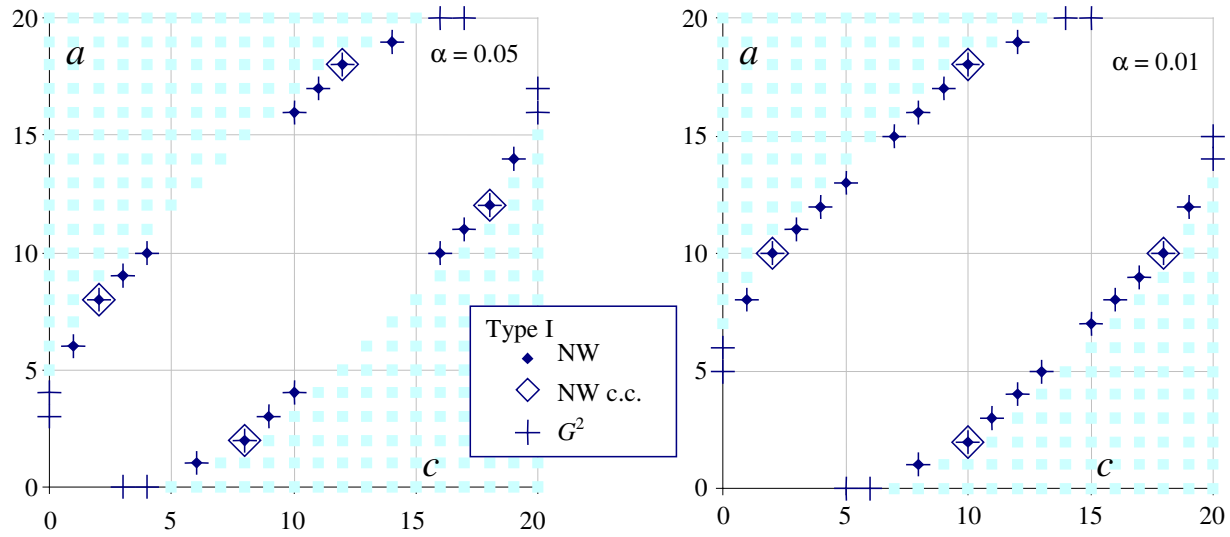

Figure 6. Evaluating the Newcombe-Wilson test, with and without continuity correction, and log likelihood G^2, against a difference test computed using the 'exact' Binomial interval, for error levels $\alpha = 0.05$ (left) and $\alpha = 0.01$ (right). (After Wallis 2013.)

This set of tests compares the difference in two observed probabilities, p_1 and p_2, with a combined interval. To obtain this interval we employ $p_1 = a / n_1$ and $p_2 = c / n_2$, where $n_1 = a + b$ and $n_2 = c + d$ (see Table 3 for the notation).

The baseline interval for comparison is obtained first by calculating Clopper-Pearson intervals for P_1 and P_2 satisfying the exact Binomial formula (Equation (1)), where $x = a$ and c, and $n = n_1$ and n_2 respectively.

The interval is combined by the Bienaymé formula (Equation (8)) to obtain a paired baseline test.

$$\textit{Bienaymé interval width} = \sqrt{(P_1 - p_1)^2 + (P_2 - p_2)^2}\,, \tag{8}$$

where P_1 and P_2 represent the extreme values of the *inner* interval (i.e. if $p_1 > p_2$, P_1 is the lower bound of p_1). These interval widths are in proportion to the standard deviations at P_1 and P_2 by interval equality. This test is slightly less conservative than Fisher's (see Section 7 below), i.e. it accepts some outcomes as significant when Fisher's exact test does not (Type I errors).

To combine other intervals (Wald z, Wilson, etc.) we also employ Equation (8), this time substituting the relevant inner interval points for P_1 and P_2. The Newcombe-Wilson interval can be computed by applying Equation (8) to the Wilson score interval formula, substituting $P_1 = w_1^-$ and $P_2 = w_2^+$ if $p_1 > p_2$, and $P_1 = w_1^+$ and $P_2 = w_2^-$ otherwise. In other words, we calculate the interval substituting for in the direction of change.

To include a continuity correction, we do the same, but employ Equations (8) and the continuity-corrected Wilson formula.

Figure 6 plots the result of comparing Newcombe-Wilson tests, with and without continuity correction, and the log-likelihood test, against the paired Binomial test. Figures 11 and 12 plot cell frequency error counts and the Fisher-weighted chance of error.

In the Appendix, to allow closer inspection, we have reproduced these graphs at a higher resolution than was possible in the journal article.

Overall they confirm the conclusions of Wallis (2013). The continuity-corrected Newcombe-Wilson test produces fewer errors than Yates' test in both conditions for larger n (in this case once the smaller sample $n_2 > 15$). Yates' test is excessively conservative.

The additional Type I error rate for the Newcombe-Wilson interval with continuity-correction falls to zero for most of the time for $n_1 = n_2$, and is below 0.0033 with the unevenly-weighted $n_1 = 5n_2$ case. Even for small n the interval performs well.

4. In conclusion: comparing Fisher and Binomial tests

We drew a distinction between two types of 2 × 2 tests. Fisher's 'exact' test (Section 3.1) is computed by summing Fisher scores for more extreme values *diagonally*, assuming that row and column totals are constant (Equation (6)). It assumes both independent and dependent variables are free to vary and samples are taken from the same population. The idea is that if any utterance by any speaker could be accounted for in any cell in the table, then the summation should be performed in both directions at the same time.

But what if the data is drawn from different populations, each with their own mean, such as any two arbitrary proportions?

In the original paper (Wallis 2013), we suggested that *in principle*, the paired Binomial model should be used in preference to Fisher's test (or any other contingency test) where data is drawn from different populations. This argument raised a few eyebrows among fellow statisticians. The main objection they raised is that the fact that data is drawn independently for values of the independent variable (i.e. for p_1 and p_2) does not prevent variation to be *considered* to be pooled in both directions in the test itself (i.e. that an equally associative test is still appropriate).

However the underlying pooling assumption here is that the mean, or sum, of two Binomial distributions about P_1 and P_2 is itself Binomial. However this assumption is not always appropriate, leading to a loss of generality. If there is a difference D between P_1 and P_2, and we add the two distributions together, the result will be 'bimodal' (it will have twin peaks). As we saw in Figure 11, Sheskin's z test (Sheskin 2011: 658) performs poorly compared to a Newcombe-Wilson version of the same method.

There is not universal agreement that the Fisher or Yates' test is always optimal. Zou and Donner (2008: 1694) argue that a centrally-pooled estimate of variance (i.e. around $\hat{p}$) 'does not reflect the asymmetry of the underlying sampling distributions' – unlike an estimate based on inner interval limits (w_1^+, w_2^- etc).

We saw that the paired Binomial test (and the Wilson approximation to it) is essential for conditions where there is a known difference D. The distinction is graphically illustrated in Figure 5, which shows that the Gaussian interval/Fisher model *is only correct for* $D = 0$, and using it for an interval on d is as problematic as using the Wald interval for the single proportion.

The paired Binomial Bienaymé model is thus necessary in meta-testing conditions (Wallis 2019), combining intervals, etc. The question is, how robust is it? Is precision at $D = 0$ sacrificed

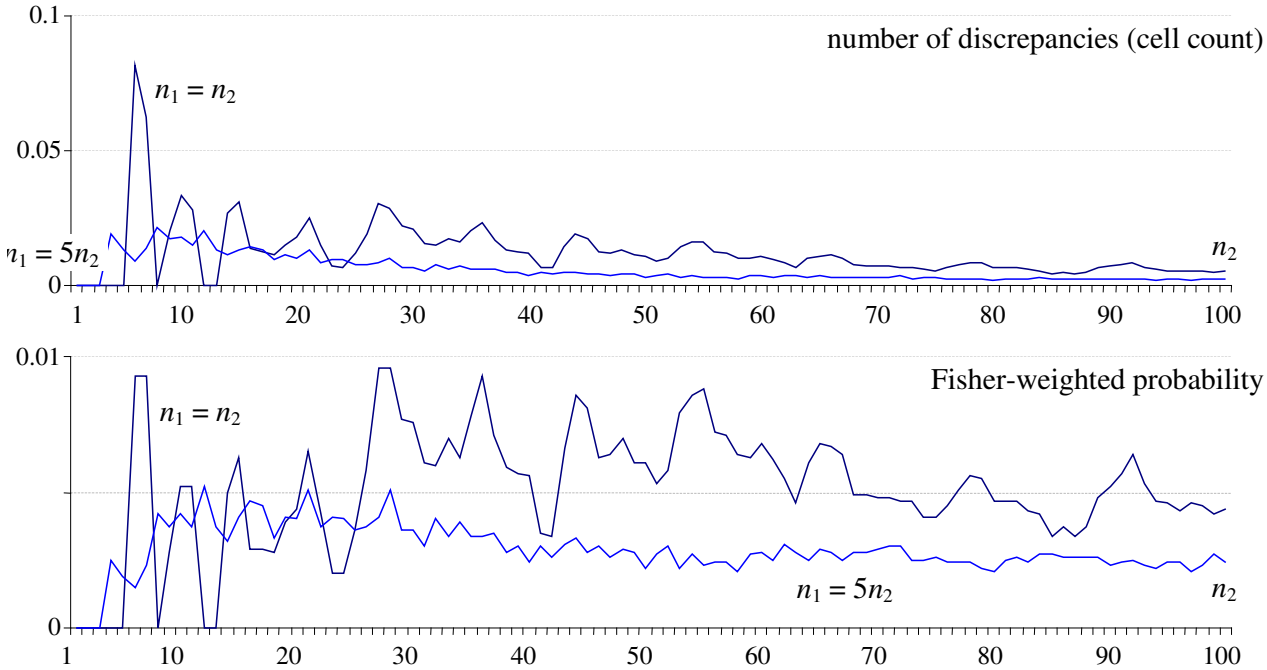

Figure 7. The effect of population independence: plotting the probability that the independent-population test is significant in cases where the same-population test is not ($\alpha = 0.05$).

for extensibility, and if so, to what extent?

The Fisher test calculates exact probabilities allowing for variation in both directions (across both DV and IV). The paired Clopper-Pearson test calculates exact intervals for two values of the IV and then employs the Bienaymé approximation to combine them. This second step introduces two potential errors – a 'smoothing' error (conventionally addressed through a continuity correction), and a 'Cartesian mapping' error (due to using Pythagoras' theorem on curved probability space).

We may compare the performance of the Clopper-Pearson and Fisher tests by the method we have used throughout this paper. We identify table configurations where one test obtains a significant result and the other does not. For $n_1 = n_2$ up to 100 and $n_1 = 5n_2$ we compare the results of tests in all possible configurations and calculate the probability of both types of errors independently.

The results show that the Fisher test is slightly more conservative than the paired Binomial test. Figure 7 plots the probability that the independent population test obtains a significant result when the dependent sample (Fisher) does not, using both estimates of error. There are no cases where Fisher's test is less conservative than the paired Binomial. The absolute differences (lower graph in Figure 7) are below 1% of all cases.

If we set $\alpha = 0.05$, the act of combining two 'exact' intervals could actually increase the error rate to 0.06, but tending to 0.055 over time. For $\alpha = 0.01$, the actual error rate may be as high as 0.015: 50% more than intended. This general method of combining intervals is due to Zou and Donner (2008). Figure 7 may be thought of as the additional cost the method obtains, evaluated by exact methods.

However, this is theoretical, as one would not employ the combined Binomial to substitute for Fisher! The real question concerns what this means in practice. Taking Fisher as the 'gold standard', i.e. even considering cases where population difference $D = 0$, the Newcombe-Wilson interval with a correction for continuity outperforms the 'exact' paired Binomial once $n > 20$ or so.

The additional continuity correction is conservative and 'absorbs' some erroneous cases. We replot the relevant plot lines from Figure 10 on the same axis in Figure 8.

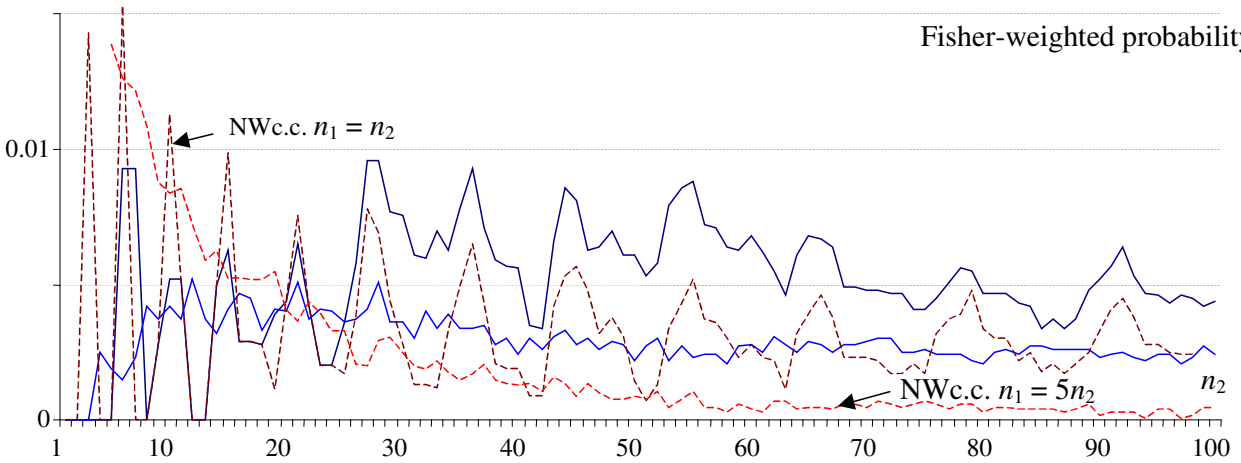

Figure 8. Type I errors generated by the continuity-corrected Newcombe-Wilson and exact Binomial intervals compared to the Fisher test, weighted by prior probability of occurrence.

References

Calain, P. (2018). The Ebola clinical trials: a precedent for research ethics in disasters. *Journal of Medical Ethics 44*:3-8. doi:10.1136/medethics-2016-103474.

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. (2018). Plotting the Wilson distribution, *corp.ling.stats*. London: Survey of English Usage. https://corplingstats.wordpress.com/2018/09/25/plotting-the-wilson-distribution.

Wallis, S.A. (2019). 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. (2021). *Statistics in Corpus Linguistics Research*. New York & Abingdon: Routledge.

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

Appendix: Evaluation of performance of tests against Fisher and paired Binomial tests

For figures, see overleaf.

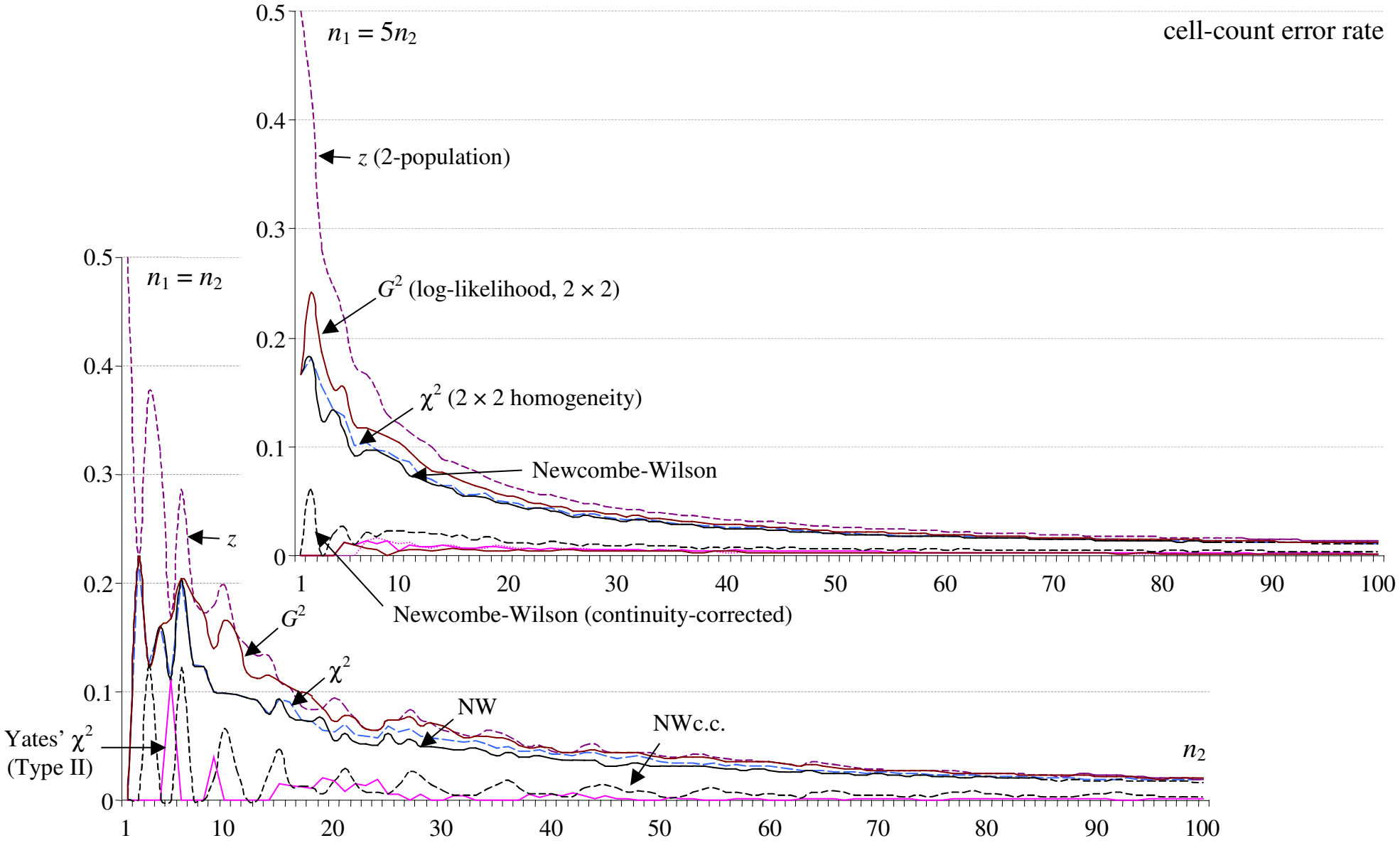

Figure 9. Cell-count error rates calculated against Fisher's test, $\alpha = 0.05$ for $n_1 = n_2$ up to 100 (lower) and $n_1 = 5n_2$ (upper). Errors are Type I unless otherwise indicated. (After Wallis 2013: 200.)

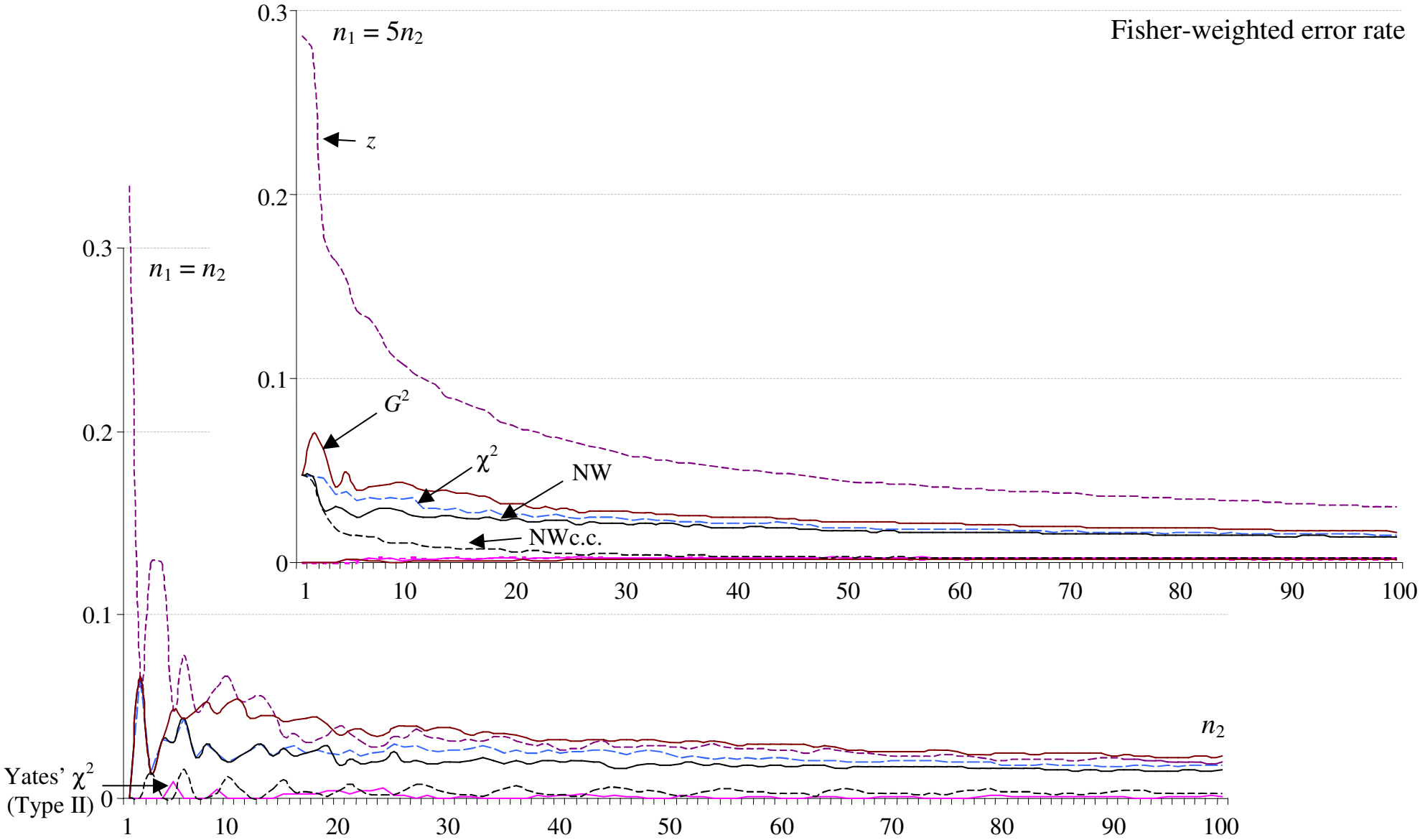

Figure 10. Fisher-weighted error rates calculated against Fisher's test. This weighted error takes into account the chance that a particular cell frequency can occur. In other words, it plots the probability of an erroneous result. The incorrect Wald z interval (top line) is much more likely to obtain erroneous results.

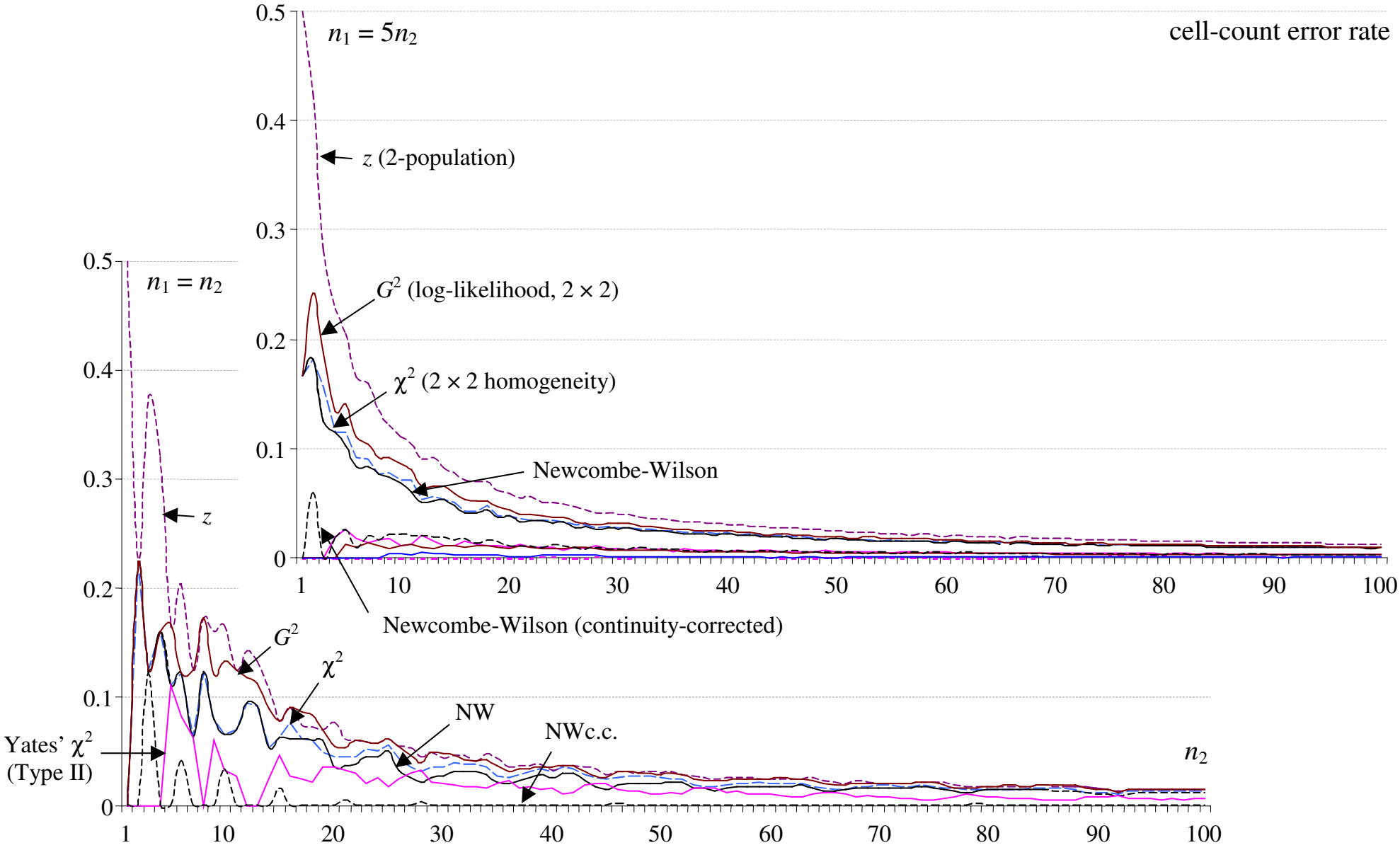

Figure 11. Cell-count error rates against the Binomial difference interval, α = 0.05. (After Wallis 2013:202).

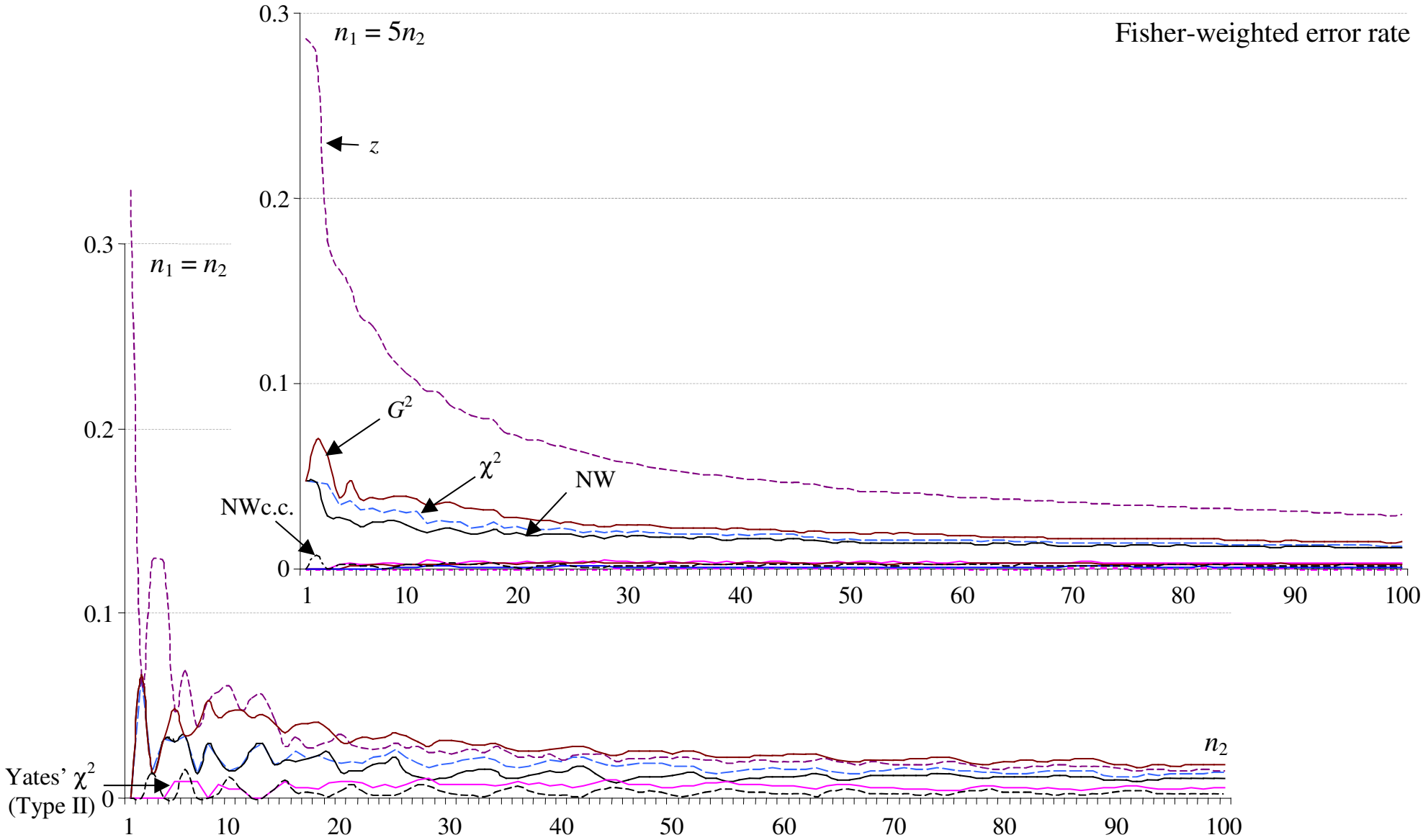

Figure 12. Fisher error rates against the Binomial difference interval, calculating the overall probability of error. The Type I error of the continuity-corrected Newcombe Wilson interval (dashed black line) is very small.