

Comparing χ^2 tests for separability

Interval estimation for the difference between a pair of differences between two proportions

Sean Wallis, Survey of English Usage, University College London
First published August 2011. Revised April 2012.¹

1. Introduction

Researchers often wish to compare the results of their experiments with those of others. Alternatively they may wish to compare permutations of an experiment to see if a modification in the experimental design obtains a significantly different result. This question concerns an empirical analysis of the *effect* of modifying an experimental design on reported results, rather than a deductive argument concerning the optimum design.

Many researchers attempt this type of evaluation by employing statements about their results (citing t , F or χ^2 scores, error levels or “ p values”, etc), as benchmarks for the strength of their results, implying a comparison that is frequently misunderstood (Goldacre 2011).

The fact that one chi-square value or error level exceeds another *merely means that reported indicators differ*. It does **not** mean that the results are statistically separable, i.e. that the results are significantly different from each other, at a given likelihood of error. However if we wish to claim a difference in experimental outcomes between experimental ‘runs’, this is precisely what we must establish. In this paper we attempt to address how this question of separability may be evaluated.

We begin by focusing on comparing the results of two paired contingency tests:

- a) two 2×2 tests for homogeneity (independence) and
- b) two 2×1 goodness of fit tests.

The idea is that both dependent and independent variables are *matched* but not precisely identical, i.e., in both tests we attempt to measure the same quantities by different definitions, methods or samples. The new test then compares these test results for separability and tells us if the effect of the change in experimental design obtains a significantly different result.

Consider the example below, from Aarts, Close and Wallis (forthcoming). The two tables summarise contingency tests for two different sets of data. The results appear to be different, especially if we consider effect size measures ϕ and $d^{\%}$. The question is whether we can test if they are significantly different *from each other*.

(spoken)	<i>shall</i>	<i>will</i>	Total	$\chi^2(\textit{shall})$	$\chi^2(\textit{will})$	summary
LLC (1960s)	124	501	625	15.28	2.49	$d^{\%} = -60.70\% \pm 19.67\%$
ICE-GB (1990s)	46	544	590	16.18	2.63	$\phi = 0.17$
TOTAL	170	1,045	1,215	31.46	5.12	$\chi^2 = 36.58$

(written)	<i>shall+</i>	<i>will+ 'll</i>	Total	$\chi^2(\textit{shall+})$	$\chi^2(\textit{will+ 'll})$	summary
LOB (1960s)	355	2,798	3,153	15.58	1.57	$d^{\%} = -39.23\% \pm 12.88\%$
FLOB (1990s)	200	2,723	2,923	16.81	1.69	$\phi = 0.08$
TOTAL	555	5,521	6,076	32.40	3.26	$\chi^2 = 35.65$

Table 1: A pair of 2×2 χ^2 tables for *shall/will* alternation, after Aarts *et al.* (forthcoming): upper, spoken, lower: written, with other differences in the experimental design. Note that χ^2 values are almost identical but Cramér’s ϕ and percentage swing $d^{\%}$ are different.

¹ A spreadsheet is also available from www.ucl.ac.uk/english-usage/staff/sean/resources/2x2-x2-separability.xls.

The idea is summarised by Figure 1. There are two broad classes of test: those that distinguish results of goodness of fit tests (“separability of fit”) and comparing tests of homogeneity (“separability of independence”, cf. Table 1).

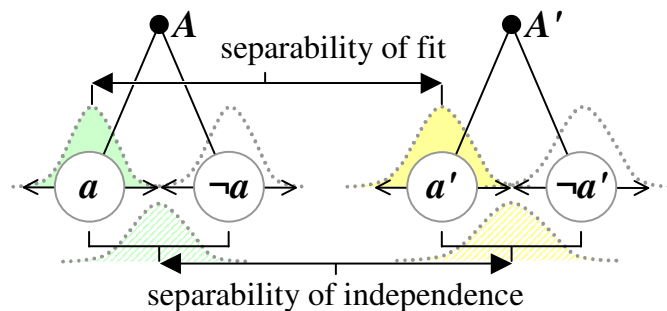


Figure 1: Visualising separability tests.

It is possible to employ a similar approach for evaluating pairs of larger “ $r \times c$ ” or “ $r \times 1$ ” tables (see section 4). However, we argue elsewhere (Wallis 2010) that it is good practice that such tables, which have many degrees of freedom (and therefore contain multiple potential axes of variation), should be analysed by subdivision into smaller tables to identify areas of significant difference. The simplest tests we describe here may therefore have the greatest utility.

The tests we describe here represent a kind of **meta-analysis**: they provide a method for comparing and summarising experimental results. Other tests for comparing contingency test results include McNemar and Cochran Q tests (Sheskin, 1997) which compare distributions, but not differences, and are known to be weak tests.

An alternative test computation to that described here is Zar’s (1999: 471, 500) **chi-square heterogeneity analysis**. We discuss Zar’s test in section 4. The null hypothesis is that ‘samples are from the same population’ and therefore justify pooling, rather than, as we have put it, that differences in experimental design and sampling fail to achieve a significantly different result.

Finally, note that in this paper we discuss contingency tests. There is a comparable procedure for comparing multiple runs of t tests (or ANOVAs) but it is rarely described as such. This is the **test for interaction in a factorial analysis of variance** (Sheskin 1997: 489) where one of the factors represents the repeated run.

1.1 Comparing proportions

The task of comparing two binomial *proportions* is a common one. The ubiquitous 2×2 χ^2 test and the z test for two independent proportions from the same population are mathematically equivalent (Wallis 2010, see also note 2). While the standard test is well-known, there is a known problem in the mathematical assumptions underpinning the test such that it fails on small datasets and skewed data, and (as generations of student statisticians have discovered), leads to advice regarding low frequency cells, Yates’ correction, etc. In recent years Wilson’s (1927) *score interval* has been rediscovered. A method based on this interval which corrects this error has been proposed and shown to outperform other methods (Newcombe 1998). See Wallis (2009) for a discussion.

The z test for two independent proportions is performed by comparing the difference between two proportions, $p_1 - p_2$, to determine whether this difference exceeds a confidence interval. The values p_1 and p_2 are binomial proportions of the form $p = f/n$ where f is the number of observed instances of a subtype of n cases, each assumed to be independent and free to take one subtype value or another. This interval is calculated by combining the confidence intervals for each single proportion separately.² The test has one degree of freedom, and the difference can range from [-1 to +1].

² The ‘Wald’ approximation assumes that given a rate of p observed over n cases, the standard deviation for p is $s = \sqrt{p(1 - p)/n}$. This obtains an interval $p \pm z.s$ where z is the critical value of the standard Normal deviate.

When we compare two tables for significant difference we assume that the observations come from **independent populations**.³ The difference between this test and the z test for two independent proportions from the **same population** (= the $2 \times 2 \chi^2$ test for homogeneity) concerns the formula for the standard deviation and therefore the confidence interval. With independent populations the variance is the sum of the individual variances (Sheskin 1997: 229):

$$\text{standard deviation } s_d \equiv \sqrt{s_1^2 + s_2^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \quad (1)$$

where s_j is the standard deviation obtained from subtest j . The symmetric confidence interval for the difference ($-e_d, e_d$) is simply $\pm z \cdot s_d$. The corresponding test is significant if $-e_d < p_1 - p_2 < e_d$ or, alternatively $|p_1 - p_2| < e_d$.

Unfortunately the ‘Wald’ approximation underpinning equation (1) has a known weakness. The interval around an observation p does not approximate well to the Gaussian. This assumption is mistaken in principle (Wilson 1927) and error-prone in practice (Newcombe 1998). (For an introduction to this question see Wallis 2009). To compute confidence intervals on p , Wilson proposes the asymmetric *score interval* in place of the Wald interval $p \pm z \cdot s$.

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

Newcombe then proposes a method (‘method 10’ in Newcombe 1998) for combining Wilson’s score interval to a new difference interval, (w^-, w^+) such that

$$w_d^- \equiv -z \sqrt{\frac{w_1^+(1-w_1^+)}{n_1} + \frac{w_2^-(1-w_2^-)}{n_2}}, \quad w_d^+ \equiv z \sqrt{\frac{w_1^-(1-w_1^-)}{n_1} + \frac{w_2^+(1-w_2^+)}{n_2}}, \quad (3)$$

where w_i^- and w_i^+ are the upper and lower bounds of the Wilson interval for each datapoint p_i and w_d^- and w_d^+ the bounds of the new interval. Again, the test is significant if $w_d^- < p_1 - p_2 < w_d^+$.

In this paper we compare solutions to the problem of comparing the results of two 2×2 contingency tests. It should go without saying that any statistical method is an adjunct to a process of experimental refinement based on underlying *theoretical* principles. Questions of ensuring that the phenomenon measured has real theoretical meaning, that baselines for comparison are meaningful and that observations are free to vary are not addressed by significance tests!

It should also be noted that what is being attempted is **not** a three dimensional chi-square test, e.g. a $2 \times 2 \times 2$ test (Zar 1999:506). This test has three degrees of freedom, one for each axis. The tests we discuss here have one degree of freedom only – that concerning the difference between two differences, and are therefore unambiguous in their interpretation. *Simply stated, the null hypothesis is that the two tests obtain the same result.*

1.2 Some example data

For the purposes of illustration, consider the following pair of contingency tables, each representing a dependent variable A (columns) and an independent variable B (rows). These could represent

³ This paper concerns cases where samples are assumed to be from independent populations to compare results. Note that the standard 2×2 test for homogeneity assumes that the samples are drawn from the same population, which is equivalent to performing the z test using the interval $\pm z \cdot s_d$ where $s_d = \sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}$ where n_1 and n_2 are the total number of cases underpinning each observation and \hat{p} is the *pooled probability estimate* for both observations, i.e. $(f_1 + f_2)/(n_1 + n_2)$. Newcombe’s equation (3) is a reworking of this formula for s_d . See Wallis (2010).

shall/will and 1960s/1990s respectively in Table 1. The z test in each case evaluates whether the probability of selecting the first column value in each case, $p_{j,i}$, is significantly affected by changes in the row variable B . In this case both 2×2 subtests are *individually* significant ($\chi^2 = 33.34$ and 11.06 respectively). The question we wish to test is whether they significantly differ *from each other*.

$f_{1,1} = 293$	$f_{1,2} = 113$	$F_1 = 406$
223	200	423
$n_{1,1} = 516$	$n_{1,2} = 313$	$N_1 = 829$

Contingency Table 1

$f_{2,1} = 20$	2	$F_2 = 22$
3	6	9
23	8	$N_2 = 31$

Contingency Table 2

The binomial proportion for selecting the first value in row i in subtest t is $p_{t,i} \equiv f_{t,i}/n_{t,i}$ (so, for example, $p_{1,1} = 293/516 = 0.5678$). We will employ the notation d_t to represent the difference in proportions in each subtest ($d_t \equiv p_{t,1} - p_{t,2}$), and s_t for the standard deviation of the difference.

The differences in proportions are thus $d_1 = 0.2068$ ($0.5678 - 0.3610$) and $d_2 = 0.6196$ ($0.8696 - 0.2500$). From this we obtain the difference in differences $D = d_1 - d_2 = -0.4128$. Difference measures can be easily visualised, as in Figure 2. The question we are concerned about is determining the appropriate confidence interval for D , and thereby a significance test.

2. Tests for differences between 2×2 tests (separability of independence)

2.1 Gaussian

We have already seen (equation 1) that the z test for two proportions sampled from independent populations employs the sum of variances rule to combine standard deviations. As the two subtests are also assumed to derive from different populations, we can employ the same formula to create an interval for evaluating the difference D between the two differences, d_1 and d_2 .

Employing the equivalence $e \equiv z.s$, the symmetric error interval about zero may be computed as

$$e_D = \sqrt{e_{d_1}^2 + e_{d_2}^2} = \sqrt{0.0702^2 + 0.3652^2} = 0.3719. \tag{4}$$

In our case the difference $D = -0.4128$ exceeds the interval $(-0.3719, 0.3719)$, and therefore the test is deemed **significant** at the error level determined by z . We can conclude that, allowing for the possibility that samples are taken from different underlying populations (e.g. different corpora) these results are separable.

2.2 Wilson

To more accurately estimate the interval around an observation p , Wilson (1927) offers the asymmetric score interval, equation (2). The fact of this asymmetry requires a little more care in combining intervals. Newcombe's equation (3) generalises the single sample Wilson interval to a test based on the difference interval by considering the *inner* sides of each interval.

Identifying the inner side (i.e. the side in the direction of the sign of D) is not as easy as

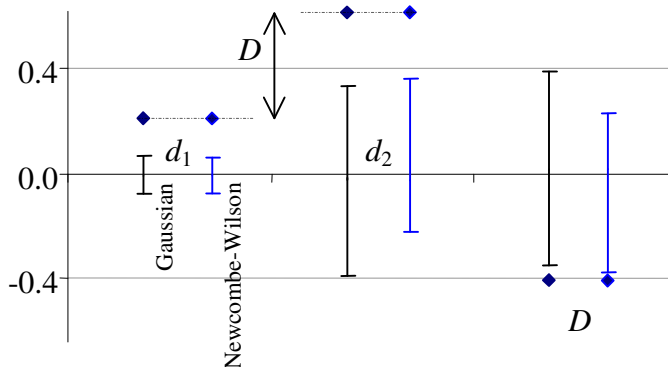


Figure 2: Gaussian and Newcombe-Wilson intervals about zero for d_1 , d_2 and D . Observations outside these intervals are significant.

might first appear. Interval points for differences $w_{d_1}^+$ (etc.) are centred on the zero axis. We must first relocate the intervals around the points d_1 and d_2 , which also requires inverting the interval (Figure 3). The inner pair of interval bounds may then be identified.

Although it is common to compute both inner and outer sides with Newcombe's formula, for testing purposes only the inner side needs to be considered. Applying equation (3) to the contingency tables we obtain the difference intervals $(-0.0663, 0.0696)$ and $(-0.1977, 0.3905)$ respectively.

We can apply the sum of variances rule to two instances of Newcombe's interval substituting appropriate pairs of Wilson w values for e_1 and e_2 (Zou and Donner 2008). The lower bound is the inner side of the interval when D is negative, respectively $w_{d_1}^-$ and $w_{d_2}^+$.

$$w_D^- = \sqrt{(w_{d_1}^-)^2 + (w_{d_2}^+)^2} = \sqrt{0.0663^2 + 0.3905^2} = 0.3961. \quad (5)$$

Similarly the upper bound $w_D^+ = 0.0696^2 + 0.1977^2 = 0.2096$, obtaining an interval of $(-w_D^-, w_D^+) = (-0.3961, 0.2096)$. Again, $D = -0.4128$ exceeds this range and the difference is significant.

2.3 ϕ -based tests

The methods we have employed have been based on estimating intervals for evaluating differences in differences (or differences in 'swings' to use Wallis' (2010) terminology). This test has only one degree of freedom, but different intervals may be applied depending on whether we assume that experimental samples are drawn from the same or independent populations. We will briefly consider an alternative approach.

Bishop, Fienberg and Holland (1975) offer an estimate of standard deviation for the signed 2×2 measure of association, ϕ , defined as

$$\phi \equiv \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{f_1(n_2 - f_2) - f_2(n_1 - f_1)}{\sqrt{n_1 n_2 F(N - F)}}. \quad (6)$$

The second formula adopts the notation in §1.2, discarding subtest indices for clarity. Using our contingency tables this formula obtains $\phi_1 = 0.2006$ and $\phi_2 = 0.5973$. The difference $\phi_1 - \phi_2 = -0.3967$. An alternative method for comparing 2×2 test outcomes would be, therefore, to evaluate the difference between these ϕ values.⁴

We calculate the standard deviation of ϕ assuming that ϕ is normally distributed. Bishop *et al* state that the standard deviation approximates towards the following expression:

$$s(\phi) \approx \frac{1}{2\phi N} \left\{ 4 \sum_{ij} \frac{P_{ij}^3}{P_{i+}^2 P_{+j}^2} - 3 \sum_i \frac{1}{P_{i+}} \left(\sum_j \frac{P_{ij}^2}{P_{i+} P_{+j}} \right)^2 - 3 \sum_j \frac{1}{P_{+j}} \left(\sum_i \frac{P_{ij}^2}{P_{i+} P_{+j}} \right)^2 \right\}$$

⁴ It is reasonable to consider differences in ϕ because it measures the degree of perturbation of the 2×2 matrix in a well defined linear manner (see Wallis 2010). However the test still employs the Gaussian approximation.

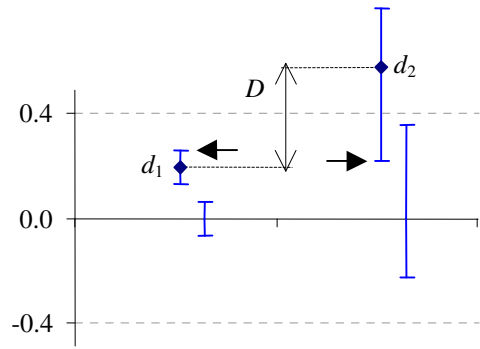


Figure 3: Identifying the inner interval (arrows) for D .

$$+ 2 \sum_{ij} \left[\frac{p_{ij}}{p_{i+} p_{+j}} \left(\sum_k \frac{p_{kj}^2}{p_{k+} p_{+j}} \right) \left(\sum_l \frac{p_{il}^2}{p_{i+} p_{+l}} \right) \right], \text{ for } \phi \neq 0 \quad (7)$$

where $p_{ij} = f_{ij}/N$ and p_{i+}, p_{+j} , etc. represent row and column (prior) probability sums.

Applying this formula obtains standard deviations of $s(\phi_1) = 0.0338$ and $s(\phi_2) = 0.1613$. In each case ϕ exceeds the respective $z.s$ error bar (± 0.0663 ; ± 0.3162), confirming that each 2×2 table is individually significant.

We can create an interval for comparing values of ϕ using equation (4), which assumes the samples are drawn from independent populations (as discussed previously).

$$e_d = \sqrt{e_1^2 + e_2^2} = \sqrt{0.0663^2 + 0.3162^2} = 0.3231.$$

Since $\phi_1 - \phi_2 = -0.3967$ exceeds this interval, the result is significant.

In principle this formula is extensible to comparing larger $r \times c$ tests and even tests of different design, although, with multiple degrees of freedom in each test it is not clear how meaningful the results would be. An alternative approach is described in section 4.

3. Tests for differences between 2×1 goodness of fit tests (separability of fit)

A 2×1 'goodness of fit' chi-square test evaluates whether the distribution at a subvalue is consistent with ('fits') the overall distribution. It can be computed using the chi-square statistic or rewritten as a single-sample z test for population proportions. Expressed as a z test, we must determine whether an observed proportion in column 1 (say), $p_{1,1} = f_{1,1}/n_{1,1}$, differs from the same proportion for the entire set, $p_1 = F_1/N_1$. From our contingency tables we obtain $p_{1,1} = 0.5678$ and $p_1 = 0.4897$; $p_{2,1} = 0.8696$ and $p_2 = 0.7097$.

The test can be computed using χ^2 or, alternatively, as a single-sample z test for a population proportion (Sheskin 1997: 118). This employs the 'expected' probability p_1 and sample size $n_{1,1}$. The difference d_1 is $p_{1,1} - p_1 = 0.0781$; $d_2 = 0.1599$.

Note that, just as the difference test for two 2×2 tests is not the same as a 3D ($2 \times 2 \times 2$) test, the difference of differences test is not the same as the $2 \times 2 \chi^2$ test. The two samples we are comparing are drawn from different populations. A different approach is required.

3.1 Gaussian

The standard deviation of the goodness of fit test is based on the population probability, so, employing our notation, $s_1 = \sqrt{p_1(1 - p_1)}/n_{1,1}$. This gives us $e_1 = z.s_1 = 0.0431$ at a 0.05 error level, which is below d_1 , so the goodness of fit test is significant. On the other hand, $e_2 = 0.1855$ and the second test is not significant.

The samples are from independent populations, so to evaluate the difference of these differences we employ the sum of variances rule (equation 4) to combine e_1 and e_2 . This gives us a confidence interval

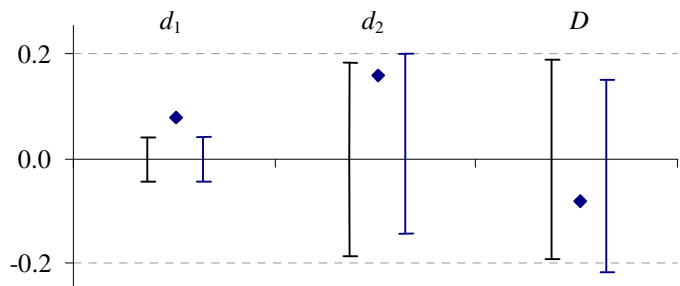


Figure 4: Gaussian (left) and Wilson (right) 95% confidence intervals for 2×1 goodness of fit tests.

of $e_D = (-0.1905, +0.1905)$. The difference of differences, $D = d_1 - d_2 = -0.0818$, is within this interval and is therefore non-significant.

3.2 Wilson

We compute Wilson intervals for each population probability p_1 and p_2 , and combine these intervals with Newcombe's method to obtain an appropriate interval for the difference of differences. The intervals for each single difference are then

$$\text{Wilson score interval } (w_1^-, w_1^+) = (0.4468, 0.5328) \text{ and } (w_2^-, w_2^+) = (0.5053, 0.8540).$$

Again, we combine intervals using the sum of squares (4). However note now that the source intervals are based at p_1 and p_2 , so the inner side of the interval when D is negative ($p_2 > p_1$) requires no inversion (cf. Figure 3). It is simply based on the upper bound of p_1 and the lower bound of p_2 . We substitute $e_1^+ = w_{d_1}^+ - p_1$ and $e_2^- = p_2 - w_{d_2}^-$ into Zou and Donner's equation:

$$w_D^- = \sqrt{(e_1^+)^2 + (e_2^-)^2} = \sqrt{0.0430^2 + 0.2043^2} = 0.2088.$$

Similarly the upper bound $w_D^+ = \sqrt{0.0429^2 + 0.1443^2} = 0.1506$, obtaining an interval of $(-w_D^-, w_D^+) = (-0.2088, 0.1506)$. Again, $D = -0.0818$ is within this range and the difference is not significant.

4. Generalisation to $r \times c$ and $r \times 1 \chi^2$ test pairs

4.1 $r \times c$ homogeneity tables

Note that whereas ϕ will generalise to $r \times c$ cases (§2.3), confidence intervals on ϕ collapse variation to a single degree of freedom. Consequently, where tables have more than one degree of freedom, two very similar values of ϕ can be obtained from very differently distributed tables – simply because there are multiple ways of obtaining the same score! If you wish to test for separability in multinomial conditions, therefore, we need to perform a different approach.

For $r \times c$ tables Zar's (1999: 500) heterogeneity test is recommended. This employs the additive property of χ^2 in an interesting way. Zar's formula is

$$\chi^2_{\text{het}} \equiv \chi^2_{\text{sum}} - \chi^2_{\text{pool}} \tag{8}$$

where χ^2_{sum} is simply the sum of individual χ^2 tests and χ^2_{pool} the result after summing paired cells. The number of degrees of freedom is the same as each single table, i.e. $(r - 1)(c - 1)$. This method has the advantage of being generalisable to $m > 2$ test runs (the degrees of freedom being multiplied by $m - 1$).

The pooled table simply sums cells over the two original tables. The first cell is $f_{1,1} + f_{2,1}$ and so forth. Using our 2×2 example data we obtain $\chi^2_{\text{pool}} = 39.82$ and $\chi^2_{\text{sum}} = 33.34 + 11.06$. The heterogeneity $\chi^2_{\text{het}} = 4.58$, which is significant at the 0.05 error level.

As we noted in the introduction, the weakness of large contingency tables is that a significant result may be obtained from multiple potential areas of variation (see also Wallis 2010). It is possible to apply the method of standardised residuals or their squares (' χ^2 partials') by applying (8) separately to each cell. We can then perform individual difference tests as required.

4.2 $r \times 1$ goodness of fit tables

Zar (1999: 471) also proposes his method for goodness of fit comparisons. However we have found that if one table is skewed, equation (8) underestimates the significant difference. Overall the results are different from those obtained using our derivation from first principles. An easy way to see this

is to adjust one cell in the second example contingency table until we obtain a borderline-significant result using the 2×1 test.

$f_{1,1} = 293$	113	$F_1 = 406$
223	200	423
516	313	829

Contingency Table 1

$f_{2,1} = 20$	2	$F_2 = 22$
3	12.57	15.57
23	14.57	37.57

Contingency Table 2'

This obtains a difference $D = -0.2059$, and a Gaussian interval of ± 0.2059 ($p_{err} = 0.05$). However if we employ Zar's method we obtain $\chi^2_{het} = 3.9855$ ($p_{err} = 0.0459$). This is a large difference.

However, we can generalise the paired 2×1 z test (§2.1) to a paired $r \times 1$ goodness of fit χ^2 test by employing a modified χ^2 test with $r - 1$ degrees of freedom (one for each pair of cells). Note that a standard chi-square test can be rewritten in terms of probabilities by dividing by n_i^2 .

$$\chi^2 \equiv \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^r \frac{(p_i - \hat{p}_i)^2}{\hat{p}_i / n_i},$$

where \hat{p}_i represents the expected probability, p_i the observed probability and n_i the column total. This formula can then be extended to comparing differences of differences:

$$\chi_D^2 \equiv \sum_{i=1}^r \frac{(d_{1,i} - d_{2,i})^2}{s_{1,i}^2 + s_{2,i}^2} = \sum_{i=1}^r \frac{((p_{1,i} - \hat{p}_{1,i}) - (p_{2,i} - \hat{p}_{2,i}))^2}{\hat{p}_{1,i} / n_{1,i} + \hat{p}_{2,i} / n_{2,i}}, \quad (9)$$

where, for each subset $t \in \{0, 1\}$,

- *difference* (observed – expected) $d_{t,i} \equiv p_{t,i} - \hat{p}_{t,i}$,
- *variance* $s_{t,i}^2 \equiv \hat{p}_{t,i} / n_{t,i}$ and
- *expected probability* $\hat{p}_{t,i} \equiv E_{t,i} / n_{t,i}$.

Performing this test on the same data obtains $\chi_D^2 = 3.8361$ ($p_{err} = 0.0502$), a slight underestimate due to rounding.

Finally, both $r \times 1$ and $r \times c$ tests rely on table pairs having the same structure. They cannot be meaningfully applied in cases where tables are structured differently, either because they refer to different variables, or because one or other table has been reduced in dimension due to the application of Cochran's rule for low frequency cells. If one table is reduced by combining cells, then the same procedure must be applied to the other table before the test is applied.

5. Conclusions

Researchers often wish to make statements about their results relative to others. However, as Goldacre (2011) notes, experimental science papers even in the most prestigious journals frequently fail to do this correctly. As we noted, this is because the difference of differences is a stochastic property that can only be properly evaluated by a statistical test. With one degree of freedom, there is a single difference of differences. If this is greater than a certain limit then we can say that the result is significant, i.e. the difference of differences is non-zero, within a given probability of error.

The 2×2 χ^2 test for homogeneity and the 2×1 test for goodness of fit are highly ubiquitous. The ability to properly compare the outcomes of such tests is therefore extremely useful.

It is preferable to cite standardised measures of association such as Cramér's ϕ over values of the χ^2 test statistic, because the former normalises the statistic to a probabilistic scale, making comparison more straightforward. However although ϕ may be cited, an observed difference between values of ϕ is not necessarily a significant one (cf. Table 1), even if both tests are found to be so.

As we have seen, both contingency tables individually obtain significant 2×2 results, and we may obtain a numerical difference between values of d and ϕ computed from these tables and test if this difference is significant. Indeed, in some circumstances, a pair of 2×2 distributions may be significantly different *from each other* even when the individual tests are **not** deemed significant (i.e. their swings are not significantly different from zero). Consider a situation where d_1 is negative and d_2 positive, for example. Significant difference *between* tables is not the same as significant difference *within* tables.

We derived Gaussian and Wilson intervals for comparing values of d and a method for comparing values of ϕ . The method for comparing ϕ values requires a complex calculation derived from χ^2 , which, as it is also based on the Gaussian, is vulnerable to inaccuracy in small skewed datasets. As there is only one degree of freedom in comparing 2×2 tables, the ϕ test can be more accurately replaced by a test employing Wilson's score interval (§2.2).

We also demonstrated how the same approach can be applied to 2×1 goodness of fit tests. With our test data this obtained difference evaluations which were within significant bounds, and therefore could not be deemed statistically separable. The methods described here are included in the online spreadsheet (see note 1) and users are encouraged to experiment with these.

Finally we demonstrated how we can extend these z -based separability tests to multi-valued conditions ($r \times 1$ goodness of fit and $r \times c$ homogeneity). Here we return to Zar's (1999) method noted in the introduction. Zar discusses his test in terms of the legitimacy of pooling samples (hence the interest in scalability to m samples), whereas we are employing this type of test to determine differences in experimental outcome. We are therefore particularly concerned with accuracy and are less concerned about more than two samples.

By comparing border-line cases, we found that Zar's heterogeneity test tends to underestimate significant difference when comparing goodness of fit tests if at least one is skewed. We therefore propose an alternative generalisation of the 2×1 Gaussian test, which employs χ^2 to generalise over multiple differences of differences. On the other hand, Zar's method performs sufficiently well for comparing outcomes of paired homogeneity tests.

References

- Aarts, B., Close, J, and Wallis, S.A. forthcoming. Choices over time: methodological issues in investigating current change. Chapter 2 in Aarts, B., Close, J, Leech, G. and Wallis, S.A. (eds.) *The Verb Phrase in English*. Cambridge: CUP.
- Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- Goldacre, B. 2011. The statistical error that just keeps on coming. *Guardian*, 9 September 2011. www.guardian.co.uk/commentisfree/2011/sep/09/bad-science-research-error
- Newcombe, R.G. 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* **17**: 873-890.
- Sheskin, D.J. 1997. *Handbook of Parametric and Nonparametric Statistical Procedures*. 1st Edition. Boca Raton, FL: CRC Press.
- Wallis 2009. *Binomial distributions, probability and Wilson's confidence interval*. London: Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/binomialpoisson.pdf
- Wallis, S.A. 2010. *z-squared: The origin and use of χ^2* . London: Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/z-squared.pdf

- Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**: 209-212.
- Zar, J. H. 1999. *Biostatistical analysis*. 4th Edition. Upper Saddle River, NJ: Prentice Hall.
- Zou G.Y. and Donner A. 2008. Construction of confidence limits about effect measures: A general approach. *Statistics in Medicine* **27**: 1693-1702.