Comparing χ^2 tables for separability of distribution and effect

Meta-tests for comparing homogeneity and goodness of fit contingency test outcomes

Sean Wallis, Survey of English Usage,
University College London, Gower Street, London, UK

email: s.wallis@ucl.ac.uk

Wallis, S. (2018). 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*. doi:10.1080/09296174.2018.1496537

Acknowledgments

I am grateful to Stefan Gries for his observations on an early version of this paper.

Abstract

This paper[1] describes a series of statistical meta-tests for comparing independent contingency tables for different types of significant difference. Recognising when an experiment obtains a significantly different result and when it does not is frequently overlooked in research publication. Papers are frequently published citing '*p* values' or test scores suggesting a 'stronger effect' substituting for sound statistical reasoning. This paper sets out a series of tests that together illustrate the correct approach to this question.

These meta-tests permit us to evaluate whether experiments have failed to replicate on new data; whether a particular data source or subcorpus obtains a significantly different result than another; or whether changing experimental parameters obtains a stronger effect.

The meta-tests are derived mathematically from the χ^2 test and the Wilson score interval, and consist of pairwise 'point' tests, 'homogeneity' tests and 'goodness of fit' tests. Meta-tests for comparing tests with one degree of freedom (e.g. '2×1' and '2×2' tests) are generalised to those of arbitrary size. Finally, we compare our approach with a competing approach offered by Zar (1999), which, while straightforward to calculate, turns out to be both less powerful and less robust.

Keywords: separability test, contingency test, χ^2 test, Wilson score interval, goodness of fit, homogeneity, heterogeneity, meta-analysis

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. By doing so they would be able to investigate the empirical question of the effect of modifying an experimental design on reported results – as distinct from a deductive argument concerning the optimum design.

One of the reasons for carrying out such a test concerns the question of replication. Significance tests and confidence intervals rely on an *a priori* Binomial model predicting the likely distribution of future runs of the same experiment. However, there is a growing concern that allegedly significant results published in eminent psychology journals have failed to replicate (see, e.g. Gelman and Loken 2013). The reasons may be due to variation of the sample, or problems with the experimental design (such as unstated assumptions or baseline conditions that vary over experimental runs). The methods described here permit us to define a 'failure to replicate' as occurring when subsequent repetitions of the same experiment obtain statistically separable results on more occasions than predicted by the error level, 'α', used for the test.

Let us begin with a real example. Consider Table 1, taken from Aarts, Close and Wallis (2013). The two tables summarise a pair of 2 × 2 contingency tests for two different sets of British English corpus data for the modal alternation *shall* vs. *will*. The spoken data is drawn from the *Diachronic Corpus of Present-day Spoken English*, which contains matching data from the *London-Lund Corpus* and the *British Component of the International Corpus of English* (ICE-GB). The written data is drawn from the *Lancaster-Oslo-Bergen* (LOB) corpus and the matching *Freiburg-Lancaster-Oslo-Bergen* (FLOB) corpus.

Both 2 × 2 subtests are individually significant (χ^2 = 36.58 and 35.65 respectively). The results (see the effect size measures ϕ and percentage difference $d^{\%}$) appear to be different.

How might we test if the tables are significantly different from each other?

We can plot Table 1 as two independent pairs of probability observations, as in Figure 1. We calculate the proportion $p = f/n$ in each case, and – in order to estimate the likely range of error introduced by the sampling procedure – compute Wilson score intervals at a 95% confidence level.

(spoken)	*shall*	*will*	**Total**	χ^2(*shall*)	χ^2(*will*)	
LLC (1960s)	124	501	625	**15.28**	2.49	$d^{\%}$ = -60.70% ±19.67%
ICE-GB (1990s)	46	544	590	**16.18**	2.63	ϕ = 0.17
TOTAL	170	1,045	1,215	**31.46**	**5.12**	χ^2 = **36.58**

(written)	*shall+*	*will+'ll*	**Total**	χ^2(*shall+*)	χ^2(*will+'ll*)	
LOB (1960s)	355	2,798	3,153	**15.58**	1.57	$d^{\%}$ = -39.23% ±12.88%
FLOB (1990s)	200	2,723	2,923	**16.81**	1.69	ϕ = 0.08
TOTAL	555	5,521	6,076	**32.40**	3.26	χ^2 = **35.65**

Table 1. A pair of 2 × 2 χ^2 tables for *shall*/*will* alternation, after Aarts *et al*. (2013): upper, spoken, lower: written, with other differences in the experimental design. Note that χ^2 values are near-identical but Cramér's ϕ and percentage swing $d^{\%}$ are different.

The intervals in Figure 1 are shown by 'I' shaped error bars: were the experiment to be re-run multiple times, in 95% of predicted repeated runs, observations at each point will fall within the interval. Where Wilson intervals do not overlap at all (e.g. LLC vs. LOB, marked '**A**') we can

identify the difference is significant without further testing; where they overlap such that one point is within the interval the difference is non-significant; otherwise a test must be applied.

In this paper we discuss two different analytical comparisons.

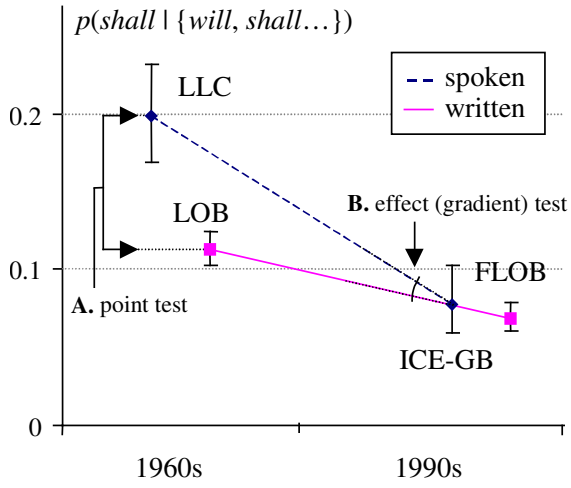

Figure 1: Example data in Table 1, plotted with 95% Wilson score intervals (Wallis 2013a). Points are separated horizontally for clarity.

A. 'Point tests' compare pairs of observations ('points') across the dependent variable (e.g. *shall/will*) and tables $t = \{1, 2\}$. To do this we compare the two points and their confidence intervals. We carry out a 2 × 2 χ^2 test for homogeneity or a Newcombe-Wilson test (Wallis 2013a) to compare each point. We can compare the initial 1960s data (LLC vs. LOB, indicated) in the same way as we might compare spoken 1960s and 1990s data (e.g. LLC vs. ICE-GB).

B. 'Gradient tests' compare differences in 'sizes of effect' (e.g. a change in the ratio *shall/will* over time) between tables *t*. We might ask, is the gradient significantly steeper for the spoken data than for the written data?

Note that these tests evaluate different things and have different outcomes. If plot-lines are parallel, the gradient test will be non-significant, but the point test could still be significant at every pair of points. The two tests are complementary analytical tools.

1.1 How not to compare test results

A common, but mistaken, approach to comparing experimental results involves simply citing the output of significance tests (Goldacre 2011). Researchers frequently make claims citing *t*, *F* or χ^2 scores, '*p* values' (error levels), etc, as evidence for the strength of results. However, this fundamentally misinterprets the meaning of these measures, and comparisons between them are not legitimate.

Consider the following pair of tables, $\mathbf{T_1}$ and $\mathbf{T_2}$ (Table 2).

A moment's glance reveals that $\mathbf{T_1}$ contains exactly 10 times the data of $\mathbf{T_2}$, but data is distributed identically, and the gradient is the same. Computing the 2 × 2 χ^2 test for homogeneity (Sheskin 1997: 209), we find that $\mathbf{T_1}$ is significant (*p* is very small), whereas $\mathbf{T_2}$, with exactly the same distribution and gradient, is non-significant. Despite the low '*p* value', it is also incorrect to refer to $\mathbf{T_1}$ as 'highly significant'. χ^2, *F*, *t* and *p* are estimates of the *reliability* of results (the likelihood that results would be found on experimental replication), rather than the *scale* of results.

$\mathbf{T_1}$	**x**	**¬x**	total
y	**290**	**110**	400
¬y	**220**	**200**	420
total	510	310	820

$\chi^2 = 35.27$ $p \approx 0.0000$

$\mathbf{T_2}$	**x**	**¬x**	total
y	**29**	**11**	40
¬y	**22**	**20**	42
total	51	31	82

$\chi^2 = 3.53$ $p = 0.0604$

Table 2. Some illustrative data with the results of χ^2 tests for homogeneity in both cases. '*p*' represents the error level, commonly referred to by 'α' to distinguish it from other probabilities.

The fact that one chi-square value or error level exceeds another merely means that reported indicators differ numerically, which can arise simply due to the fact that one experiment was able to draw from a larger experimental dataset than another. It does *not* mean that the results can be *statistically separated*, i.e. that tables of observed data can be said to be significantly different from each other at a given error level.

This type of erroneous reasoning is widespread, and is expressed in other ways than explicitly comparing '*p* values' or test scores. A survey of 513 published neuroscience papers in five top-ranking journals (Nieuwenhuis *et al.* 2011) found that, of those papers that compared results of experiments, almost exactly half asserted that two results were significantly different because one experiment detected a significant effect and the other did not. As Table 2 demonstrates, this argument is also false.

1.2 Comparing sizes of effect

'Effect size' statistics, such as probability difference, percentage swing, log odds, Cramér's ϕ, Cohen's *d*, etc. attempt to summarise observed distributions in terms of their absolute difference. They factor out differences due to the quantity of data observed and may legitimately be employed for comparison purposes.

Cramér's ϕ (Sheskin 1997: 244) is based on χ^2, but it is scaled by the quantity of data *N*. For a 2×2 table with cell frequencies represented as [[*a b*] [*c d*]], we can compute a signed score with equation (1) ranging from [-1, 1]. With larger tables of dimensions $r \times c$, the unsigned score (2) may be used, where *k* is the number of cells along the shorter side, i.e. min(*r*, *c*).

$$\phi \equiv \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{1}$$

$$|\phi| = \sqrt{\frac{\chi^2}{(k-1)N}} \tag{2}$$

In the case of the tables above, both tables obtain exactly the same score, $\phi = 0.2074$.

Effect size measures adjust for the volume of data and measure the pattern of change (the 'gradient') observed. However, effect size *comparisons* are discussed in the literature in surprisingly crude terms, e.g. 'strong', 'medium' and 'weak' effects (cf. Sheskin 1997: ibid.). This is unsatisfactory.

To claim a significant difference in experimental outcomes between experimental 'runs', one method would be to establish that effect sizes (e.g. 'gradients') significantly differ. In the case of larger tables of more than one degree of freedom, we extend this principle to one in which distributions are compared in an analogous manner.

As a shorthand we will refer to a difference in distribution as 'separability'. In this paper we attempt to address how the question of 'significant separability' may be evaluated.

1.3 Meta-tests

The tests we describe here represent the building blocks for 'meta-analysis': they provide a method for comparing experimental results.

Other researchers have proposed tests for comparing experimental runs. The McNemar test (Sheskin 1997: 315) translates cross-tabulated scores to a χ^2 test; Cochran's Q test (ibid. 469) generalises this to k sets of scores.

The most similar meta-test to the approach we discuss below is Zar's chi-square heterogeneity analysis (Zar 1999: 471, 500). Section 5 reviews these tests and compares them with our approach. The key difference is that Zar's method assumes that data in both samples have (approximately) the same prior distribution (i.e. the same starting point), whereas our tests do not. Zar's test does not distinguish between tests that compare points and those that compare gradients.

In this paper we discuss contingency tests. A comparable procedure for comparing multiple runs of t tests (or ANOVAs) is the test for interaction in a factorial analysis of variance (Sheskin 1997: 489) where one of the factors represents the repeated run.

This paper is laid out as follows. Following some preliminaries, in section 3 we introduce the 'point test' and 'multi-point test' for comparing the distribution of data across the dependent variable in homogeneity tables. Section 4 introduces 'gradient test' methods for comparing sizes of effect in homogeneity tables, commencing with intervals and tests with a single degree of freedom, a test comparing Cramér's ϕ effect sizes, and ending with formulae for generalising tests to compare larger tables ($r \times c$ homogeneity tables).

Section 5 introduces a similar range of 'gradient' meta-tests for comparing goodness of fit test results. In section 6 we compare our methods with Zar's alternative approach, and section 7 is the conclusion.

2. Some preliminaries

2.1 Test assumptions

In comparing experimental runs or designs, we assume 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. Table 1 contains data drawn from different data sources (corpora) and a different set of queries were employed in each case.[2]

The meta-test then compares these test results for separability. This tells us if the effect of changes in experimental design, or differences between samples, obtain a significantly different result.

Three broad classes of test are summarised in Figure 2: those that distinguish results of goodness of fit tests ('separability of fit'), point tests for homogeneity ('separability of observations') and those that compare the gradient of homogeneity ('separability of independence').

In this paper we will focus on 2×2 and 2×1 tests because they have one degree of freedom, so significant differences in size of effect may be explained by a single factor.

We will explain how these tests may be generalised to evaluating larger tables. However, it is good analytical practice (see e.g. Wallis 2013b) for such tables – which have many degrees of freedom and multiple potential axes of variation – to be analysed by subdivision into smaller tables in order to identify areas of significant difference each with a single degree of freedom. The simplest tests we describe here may therefore have the greatest utility.

2.2 χ^2, z and Wilson intervals

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 drawn from the *same population* are mathematically equivalent (Wallis 2013b).

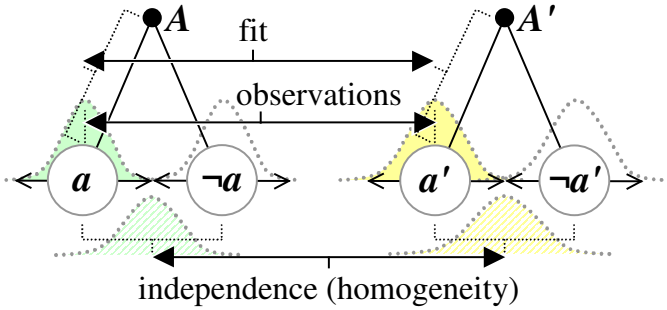

Figure 2: Visualising separability tests. From top: separability of fit compares fitness of a to A with a' to A'; separability of observations (the 'point test') compares a and a'; separability of independence compares homogeneity tests for a and $\neg a$, and a' and $\neg a'$.

These tests employ an approximation of the continuous and symmetric Gaussian (Normal) distribution to the discrete and asymmetric Binomial distribution. This model is robust except with small datasets and skewed data, and (as generations of student statisticians have discovered), leads to advice regarding low frequency cells, employing Yates' continuity correction, etc. For an analysis of the performance of these tests, see Wallis (2013a).[3]

The z test is performed by comparing the difference between two proportions, $d = p_1 - p_2$, 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. The difference interval is calculated by combining the confidence intervals for each single proportion separately. The test has one degree of freedom, and the difference can range from [-1 to +1].

A related z test compares two independent proportions from *independent populations*. In this case we do not assume that the population probability is the same in both samples. The test is more appropriate for a 'between subjects' experimental design, i.e. one in which the independent variable partitions data by speaker. This type of test is therefore clearly more theoretically appealing for the type of evaluation discussed in this paper: we wish to compare different 'runs' of the same experiment, so we do not assume that the populations are identical.

Both types of z test can be expressed in terms of testing a difference against an error threshold e_d. In the case of Gaussian methods, we first compute a standard deviation of the difference, s_d, using the sum of variances rule (also known as the Bienyamé rule).

$$\textit{standard deviation } s_d \equiv \sqrt{s_1^2 + s_2^2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}, \qquad (3)$$

where P_1 and P_2 are the population probabilities for each observation. We then take $e_d = z_{\alpha/2}.s_d$ where $z_{\alpha/2}$ represents the critical value of the Normal distribution, z, with a two-tailed error of α to obtain the interval $(-e_d, e_d)$. The test is not significant if $-e_d < p_1 - p_2 < e_d$.

For the same-population z test (equivalent to the 2 × 2 χ^2 test) we may substitute the pooled probability estimate $\hat{p}$ for both P_1 and P_2. We have $\hat{p} = F/N$, and (3) becomes

$$\textit{standard deviation } s_d \equiv \sqrt{\hat{p}(1-\hat{p})\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)}. \qquad (4)$$

In the case of independent populations, population means P_1 and P_2 could be different. Sheskin (1997: 229) simply proposes that we substitute the observed probabilities p_1 and p_2 into (3), employing the sum of variances rule.

However, this substitution is an example of an oft-repeated mathematical mistake: the 'Wald' interval (Wilson 1927).[4] The Wald interval (often cited as 'standard error') assumes that the expected probability distribution of the population mean P around an observation p is Normal with variance $p(1 - p) / n$.

In fact, the correct approximation is in the opposite direction – the Normal approximation to the Binomial means that we expect to find p distributed Normally about P. To compute confidence intervals on p, Wilson derives the following asymmetric score interval:

$$\textit{Wilson score interval } (w^-, w^+) \equiv \left(p + \frac{{z_{\alpha/2}}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n} + \frac{{z_{\alpha/2}}^2}{4n^2}} \right) \Bigg/ \left(1 + \frac{{z_{\alpha/2}}^2}{n} \right). \quad (5)$$

In cases 'at the edge of significance', where $P = w^-$ or w^+, Wilson's interval for p obtains the exact same interval width as the equivalent Normal interval for P (Wallis 2013a). This means that in a goodness of fit *test*, it does not matter whether one computes the Normal interval about P or the Wilson interval about p – the result of the test is identical. But the best measure of the error *interval* for an observed p is obtained with equation (5) (or a continuity-corrected version[5] of it).

Newcombe (1998) uses Wilson's score interval to obtain an accurate method for estimating a confidence interval on the difference between two independent observed proportions $d = p_1 - p_2$.[6] The Newcombe-Wilson difference interval (w_d^-, w_d^+) is obtained as follows:

$$w_d^- \equiv -z_{\alpha/2}\sqrt{\frac{w_1^+(1-w_1^+)}{n_1} + \frac{w_2^-(1-w_2^-)}{n_2}}, \; w_d^+ \equiv z_{\alpha/2}\sqrt{\frac{w_1^-(1-w_1^-)}{n_1} + \frac{w_2^+(1-w_2^+)}{n_2}}, \quad (6)$$

where $p_i = f_i / n_i$, w_i^- and w_i^+ are the upper and lower bounds of the Wilson interval for p_i, and w_d^- and w_d^+ are the bounds of the new interval.[7] The related test is not significant if the difference is within the interval, i.e. where $w_d^- < p_1 - p_2 < w_d^+$.

In practical terms this computation means that if $p_1 < p_2$, w_d^- is considered; if $p_1 > p_2$, w_d^+ is relevant. Although it is common to compute both inner and outer sides with Newcombe's formula, for testing purposes, only the inner side need be considered. In Figure 3, $p_1 > p_2$, d is positive, and we compute w_d^+ from the inner intervals indicated.

2.3 Example data and notation

We will use the data in Table 1 to exemplify what follows, but first we need to introduce some notation. Each table contains a **dependent variable** A (columns: modal *shall* vs. *will*) and an **independent variable** B (rows: time period). In Table 3 we simply represent this data and relevant terms.

The Binomial proportion (probability) of selecting the first value in column i in subtest t is $p_{t,i} \equiv f_{t,i}/n_{t,i}$ (thus $p_{1,1} = 124/625 = 0.1984$). We will use 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.

$f_{1,1}$ = **124**	$f_{1,2}$ = **46**	F_1 = 170	$f_{2,1}$ = **355**	$f_{1,2}$ = **200**	F_2 = 555
501	**544**	1,045	**2,798**	**2,723**	5,521
$n_{1,1}$ = 625	$n_{1,2}$ = 590	N_1 = 1,215	$n_{2,1}$ = 31,53	$n_{2,2}$ = 2,923	N_2 = 6,076
$p_{1,1}$ = 0.1984	$p_{1,2}$ = 0.0780	$\hat{p}_1$ = 0.1399	$p_{2,1}$ = 0.1126	$p_{2,2}$ = 0.0684	$\hat{p}_2$ = 0.0913

Table 3. Table 1, revisited, with notation and probabilities calculated as a proportion of column totals.

Difference measures can be easily visualised, as in Figure 4. The question we are concerned about is determining the appropriate confidence interval for D, and thus a significance test.

These tests compare a pair of χ^2 tables, each table with data drawn from different populations. The data within each separate contingency table may come from the same population (the optimum test being Yates' continuity-corrected χ^2) or from different populations, e.g. different time periods. In the calculations that follow we will use an error level $\alpha = 0.05$. [8]

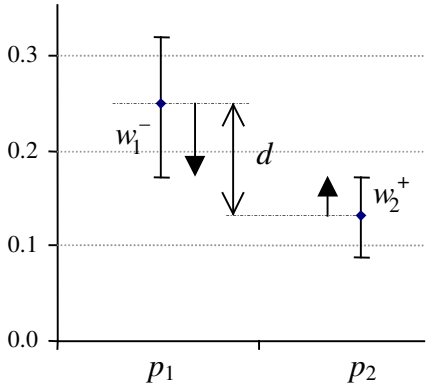

Figure 3: Identifying the inner interval (arrows) for a difference d.

3. Point and multi-point tests for homogeneity tables

3.1 The Newcombe-Wilson point test

If data is plotted with Wilson score intervals, the most natural next step is simply to carry out a Newcombe-Wilson difference test. This tests whether the observed difference between two points, $d = p_{1,i} - p_{2,i}$, is greater than a difference interval (w_d^-, w_d^+) computed from the two Wilson intervals with equation (6).

Consider the 1960s data in Table 1 (Figure 1). We draw data from row 1 in both tables ({*shall*, *will*} = {124, 501} and {355, 2798}), to obtain probabilities $p_{1,1} = 0.1984$ and $p_{2,1} = 0.1126$. The difference in probabilities, $d = 0.0858$.

At $\alpha = 0.05$, we compute a Newcombe-Wilson difference interval of (-0.0347, 0.0312) from the inner Wilson score intervals (see Figure 3), an interval that d exceeds. The difference is therefore significant.

3.2 The Gaussian point test

We can also employ the z test or χ^2 test to perform the same comparison. These tests assume that both observations are Binomially distributed about the pooled probability estimate, and we compute a standard deviation from that figure using equation (4).[9]

In our data we obtain a pooled probability estimate $\hat{p} = 0.1268$ and standard deviation $s_d = 0.0146$. The Gaussian error $e_d = z_{\alpha/2}.s_d = 0.0285$, which obtains an interval that $d = 0.0858$ exceeds.

An alternative computation that achieves the same result employs the standard homogeneity χ^2 formula, $\chi^2 = \Sigma(f_{t,j} - E_{t,j})^2/E_{t,j}$, where $E_{t,j} = n_t \times n_j / N$. The χ^2 computation for the point test is then carried out by arranging data to create a new 2×2 contingency table, DV $\times$ table t, [[124, 501] [355, 2798]]. This obtains $\chi^2 = 34.69$, which exceeds the critical value for one degree of freedom.

Using χ^2 has one advantage: it can be extended to variables with multiple values, and applied to multiple experimental runs. For dependent variables with c values, over t runs, we may employ a homogeneity test for t x c tables.

3.3 The multi-point test for **r × c** *homogeneity tests*

We can generalise the 'point test' to a 'multi-point test' by simply applying the following formula.

$$\chi_d^2 \equiv \sum_{i=1}^{r} \chi^2(i)\,, \qquad (7)$$

where $\chi^2(i)$ represents the χ^2 score for homogeneity for each set of data at position i in the distribution. This has $r \times df(i)$ degrees of freedom, where $df(i)$ is the degrees of freedom for each χ^2 point test.

Note that whereas χ^2 is generally associative (bi-directional), equation (7) is not. The multi-point test factors out variation between tests over the independent variable – so if there is a lot more data in one table in one particular time period, it does not skew the results – but does not factor out variation over the dependent variable. (After all, this is precisely what we wish to examine.)

In brief, the calculation is applied over the dependent variable, e.g. {*shall*, *will*}; {*shall*, *will*, *BE going to*}, etc. and table t, and summed over the independent variable.

We will discuss the Binomial case with $c = 2$ below and use it to examine the data in Table 1.

In 2 × 2 homogeneity tables the z score is the square root of the χ^2 score (Wallis 2013b), so we can also apply the following substitution:

$$\chi^2_d \equiv \sum_{i=1}^{r} \frac{\left(p_{1,i} - p_{2,i}\right)^2}{s_i^2} \,. \tag{8}$$

where

- *variance* $s_i^2 \equiv \hat{p}_i(1 - \hat{p}_i)\,(1/n_{1,i} + 1/n_{2,i})$
- *expected (pooled) probability* $\hat{p}_i \equiv (f_{1,i} + f_{2,i}) \,/\, (n_{1,i} + n_{2,i})$
- *observed probability* $p_{t,i} \equiv f_{t,i} \,/\, n_{t,i}$,

and $f_{t,i}$ represents the observed cell frequency in the first column of test t, and $n_{t,i}$ the row sum for that column. The formula sums r point tests, so it has r degrees of freedom.

Table 4 shows the computation for the data in Table 3.

	shall	*will*	total
1960s	**124**	**501**	625
1990s	**46**	**544**	590
total	170	1045	1215

	shall	*will*	total
1960s	**355**	**2798**	3153
1990s	**200**	**2723**	2923
total	555	5521	6076

p_1	p_2	$\hat{p}$	s^2	χ^2
0.1984	0.1126	0.1268	0.0002	34.6906
0.0780	0.0684	0.0700	0.0001	0.6865
				35.3772

Table 4. Table 1 revisited, with the generalised point test calculation (equation (8)). $\chi^2 = 35.38$ is significant with 2 degrees of freedom and $\alpha = 0.05$.

Since the computation sums independently-calculated χ^2 scores, each score may be individually considered for significant difference (with $df(i)$ degrees of freedom). Hence the large score for the 1960s data (individually significant) and the small score for 1990s (individually non-significant). [10]

4. Gradient tests for homogeneity tables

For each table we calculate the pooled probability estimate and standard deviation using equation (4). The differences in proportions in this case are $d_1 = 0.2500$ (0.7924 – 0.4794) and $d_2 = 0.1328$ (0.6396 – 0.5068). By subtraction we obtain the difference in differences (Figure 4), $D = d_1 - d_2 = 0.1171$.

4.1 The 2 × 2 χ^2 test for homogeneity

We compute standard deviations with equation (4) and error intervals for d_1 and d_2 with $\alpha = 0.05$. Employing the equivalence $e \equiv z_{\alpha/2}.s$, the symmetric error interval about zero may be computed as

$$e_D = \sqrt{e_{d_1}^2 + e_{d_2}^2} \qquad (9)$$
$$= \sqrt{0.0810^2 + 0.0436^2}$$
$$= 0.0920.$$

In our case, difference $D = 0.1171$ exceeds the interval (-0.0920, 0.0920), and therefore the test is deemed significant at error level α. The gradients are different.

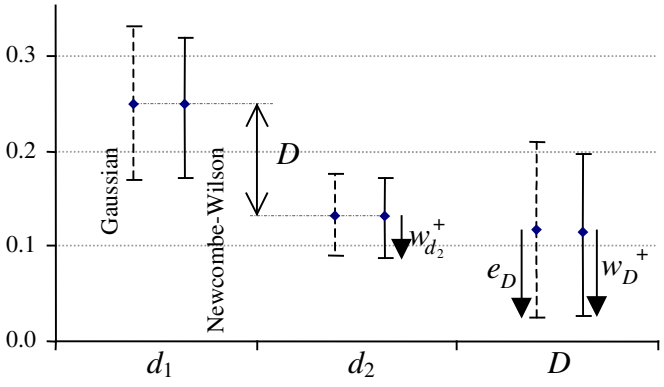

Figure 4: Gaussian (dashed) and Newcombe-Wilson difference intervals about zero for d_1, d_2 and D, centred on the observations. Where an interval does not include zero ($e_D < D$, $w_D^+ < D$), the difference is significant.[11]

4.2 The 2 × 2 Newcombe-Wilson test

The optimum 2 × 2 test for samples obtained from independent populations is the Newcombe-Wilson test with continuity correction. It may be used in preference to Yates' test where values of the independent variable ('samples') are obtained from distinct data sources (Wallis 2013a). Exactly the same data is applied, and D is computed as before, but the confidence interval we obtain is calculated differently.

In this evaluation, we use Wilson's (1927) asymmetric score interval for p (equation 5) with $\alpha = 0.05$, to derive Wilson score intervals (see Figure 1). The difference interval is computed with equation (6). In our *shall/will* data, this obtains the Newcombe-Wilson difference intervals for d_1 and d_2. We can write these as follows:

$d_1 = 0.2500 - (-0.0682, 0.0775) > 0$, and
$d_2 = 0.1328 - (-0.0410, 0.0429) > 0$.

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

$$w_D^+ = \sqrt{(w_{d_1}^+)^2 + (w_{d_2}^-)^2} = \sqrt{0.0775^2 + 0.0410^2} = 0.0877. \qquad (10)$$

Similarly the lower bound $w_D^- = 0.0805$, obtaining an interval of $(-w_D^-, w_D^+) = (-0.0805, 0.0877)$. Again, $D = 0.1171$ exceeds this range, and the difference of differences is significant.

4.3 Cramér's ϕ interval and test

The methods we have seen thus far in this section estimate intervals for differences in differences (gradients). We will briefly consider an alternative approach that uses an alternative, widely-used measure of effect size.

Cramér's ϕ measures the degree of perturbation of the 2 × 2 matrix in a well-defined linear manner (Wallis 2012: Appendix 1). It combines gradient and starting point in a single statistic with one degree of freedom. We have already seen the signed 2 × 2 measure of association, ϕ, in equation (1), which ranges from –1 to +1.

$$\phi \equiv \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{f_1(n_2-f_2)-f_2(n_1-f_1)}{\sqrt{n_1 n_2 F(N-F)}}. \qquad (1')$$

The second formula adopts the notation in Table 3, discarding subtest indices (*t*) for clarity. Using our contingency tables this formula obtains $\phi_1 = 0.1735$ and $\phi_2 = 0.0766$. The difference $\phi_1 - \phi_2 = 0.0969$. (Differences range from –2 to +2.)

An alternative method for comparing 2 × 2 test outcomes is to compare this difference with a difference interval for ϕ. The standard deviation of ϕ is given by Bishop, Fienberg and Holland (1975: 386) as:

$$s(\phi) \approx \frac{1}{2\phi N}\left\{4\sum_{ij}\frac{p_{ij}^3}{p_{i+}^2 p_{+j}^2} - 3\sum_i \frac{1}{p_{i+}}\left(\sum_j \frac{p_{ij}^2}{p_{i+}p_{+j}}\right)^2 - 3\sum_j \frac{1}{p_{+j}}\left(\sum_i \frac{p_{ij}^2}{p_{i+}p_{+j}}\right)^2 + 2\sum_{ij}\left[\frac{p_{ij}}{p_{i+}p_{+j}}\left(\sum_k \frac{p_{kj}^2}{p_{k+}p_{+j}}\right)\left(\sum_l \frac{p_{il}^2}{p_{i+}p_{+l}}\right)\right]\right\}, \text{ for } \phi \neq 0 \qquad (11)$$

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

Applying this formula to our data obtains standard deviations of $s(\phi_1) = 0.0265$ and $s(\phi_2) = 0.0125$. In each case ϕ exceeds the respective error interval ($e(\phi_1) = 0.0519$; $e(\phi_2) = 0.0244$), confirming that each 2 × 2 table is individually significant.

Second, we create an interval for comparing values of ϕ using the Bienyamé equation (9). This assumes that the error around ϕ is Normally distributed.

$$e_d = \sqrt{e(\varphi_1)^2 + e(\varphi_2)^2} = \sqrt{0.0519^2 + 0.0244^2} = 0.0573.$$

Since $\phi_1 - \phi_2 = 0.0969$ exceeds this interval, the difference is significant.

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

Unfortunately, even for 2 × 2 tests, the ϕ method has three further disadvantages over difference methods. For $\phi = 0$ it fails. It employs a conservative Gaussian (Normal) approximation. Thirdly, ϕ is 'associative', so directional information is factored out. In Wallis (forthcoming) we show how Newcombe-Wilson differences (section 4.2) may be compared across both variables in a 2 × 2 table using a separability test to decide if one variable obtains a greater difference than the other. This test is simply not possible with ϕ, because the difference between ϕ scores in either direction is always zero.

Considering alternative approaches, in comparing 2 × 2 contingency tests for gradient, the Newcombe-Wilson method is to be preferred over the Gaussian or ϕ. It makes no assumptions about the permissibility of pooling probability, and it does not ignore directional information.

The tests we have discussed so far in this section 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 have the same effect size.

4.4 r × *2 homogeneity tests*

In this section we briefly discuss how gradient tests can be extended to larger tables with multiple degrees of freedom. Here, this notion of differences in *size* of effect (gradient) might be better defined as differences in '*patterns* of effect', i.e. the distribution of individual differences between the two test evaluations. We will demonstrate Gaussian solutions to this problem using χ^2.

For $r \times 2$ tables we can use equation (12). The difference $D_i = d_{1,i} - d_{2,i}$ is squared and divided by the sum of the two variances; and each term is summed over the table.

$$\chi^2_D \equiv \tfrac{1}{2} \sum_{i=1}^{r} \frac{\left(d_{1,i} - d_{2,i}\right)^2}{s^2_{1,i} + s^2_{2,i}}, \tag{12}$$

where

- *difference* $d_{t,i} \equiv p_{t,i,1} - p_{t,i,2}$,
- *variance* (equation (4)) $s^2_{t,i} \equiv \hat{p}_{t,i}(1 - \hat{p}_{t,i})\,(1/n_{t,i,1} + 1/n_{t,i,2})$
- *pooled probability* $\hat{p}_{t,i} \equiv F_{t,i} / N_{t,i}$.

This test has a particular application. In analysing large $r \times c$ tables it is often helpful to plot data in a column (or row) as a series of probabilities with confidence intervals.

It is common to plot the probability of selecting a value of the dependent variable (e.g. the chance of selecting *shall* out of the set {*shall*, *will*}). With two values, probabilities are mutually exclusive (p(*shall*) = 1 – p(*will*)). But if there are three or more values of the dependent variable (e.g. {*shall*, *will*, BE *going to*}), then it is useful to plot each line separately.

Consider the example data in Table 5. We will label each row, which represents values of an independent variable, A, B and C for clarity. We plot the data with Wilson score intervals in Figure 5.

A	$f_{1,1,A}$ = **20**	35	$F_{1,A}$ = **55**
B	**40**	40	**80**
C	**1**	2	**3**
total	$n_{1,1}$ = 61	77	N_1 = 138

A	$f_{2,1,A}$ = **20**	35	$F_{2,A}$ = **55**
B	**2**	10	**12**
C	**3**	23	**26**
total	$n_{2,1}$ = 25	68	N_2 = 93

Table 5. A pair of 3 × 2 tables with example data.

Although our test statistic is calculated using row totals ($n_{1,1}$, etc), this fact does not matter. For gradients the χ^2 homogeneity test is 'associative', i.e. it obtains the same result if it is summed in either direction.[12]

Our meta-test compares the two tables, or plot-lines, against each other, averaging error intervals. In this instance we obtain $\chi^2 = 7.6110$, which is significant for $\alpha = 0.05$ and two degrees of freedom. This means we can state that the pattern of effect illustrated by the two lines are significantly different, without necessarily specifying where differences lie.

If we examine Figure 5, we can then see that most of the difference between the lines is attributable to row B. Row A observes identical probabilities. In row C, the probability $p_{2,C}$ falls within the confidence interval for $p_{1,C}$, so it cannot represent a significant difference at this point.

4.5 r × c *homogeneity tests*

For arbitrary-size $r \times c$ tables, we sum over both r and c, for $(r-1)(c-1)$ degrees of freedom. This gives us the formula:

$$\chi^2_D \equiv \tfrac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(d_{1,i,j}-d_{2,i,j}\right)^2}{s^2_{1,i,j}+s^2_{2,i,j}}, \qquad (13)$$

where

- *difference* $d_{t,i,j} \equiv p_{t,i,j} - \hat{p}_{t,i}$,
- *variance* $s^2_{t,i,j} \equiv \hat{p}_{t,i,j}(1-\hat{p}_{t,i,j}) / n_{t,j}$
- *pooled probability* $\hat{p}_{t,i,j} \equiv F_{t,i,j} / N_{t,i}$.

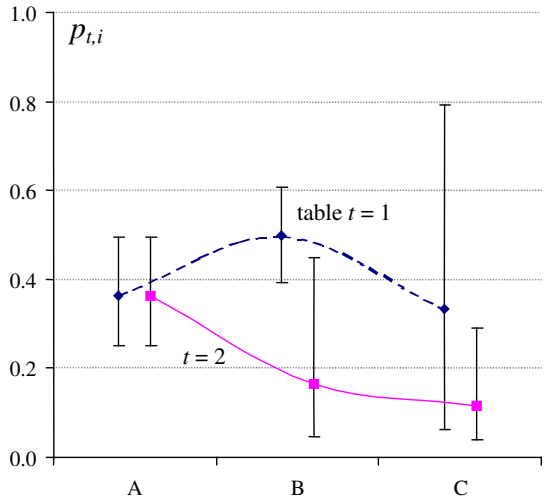

Figure 5: Plotting data from Table 5 with 95% Wilson score intervals. Each line represents column data from tables t = 1,2 plotted against column totals, i.e. $p_{t,\mathrm{A}} = f_{t,1,\mathrm{A}}/F_{t,\mathrm{A}}$, etc.

4.6 Interpreting gradient meta-tests for large tables

This illustration bears on a more general point. As tables increase in numbers of cells and degrees of freedom, the meaningful interpretation of a single significant result calculated across many data points becomes more difficult. A single number has to express the total 'difference' between two test runs – a difference that could be due to variation at any point.

Large tests are thus more usefully analysed as a series of independent tests. First, we can divide values of the dependent variable into plot lines and employ the $r \times 2$ test (section 4.4) to consider each value in turn. Second, we may carry out the multi-point test, comprising a series of point-by-point comparisons, each with a single degree of freedom, to explore where differences lie. Plotting data with Wilson score intervals leads us naturally to the Newcombe-Wilson point test (section 3.1) and, if required, the Newcombe-Wilson meta-homogeneity test (4.2).

Note that meta-tests require both tables to possess an identical structure. They cannot be meaningfully applied if tables are structured differently. If one or other table has been reduced in dimension due to the application of Cochran's rule for low frequency cells,[13] the same procedure must be applied to the other table before the test is applied.

5. Gradient tests for goodness of fit tables

A 2×1 'goodness of fit' chi-square test evaluates whether the distribution of a subset is consistent with (it 'fits') an overall expected distribution of a superset. It can be computed using the chi-square statistic where expected values are simply the scaled total, or rewritten as a single-sample z test for population proportions (Sheskin 1997: 118; Wallis 2013b).

Let us apply this test to the proposition that modal *shall* is more frequent in the 1960s than the superset *shall+will* would otherwise indicate.

$f_{1,1}$ = **124**	501	F_1 = **625**	$f_{2,1}$ = **355**	2,798	F_2 = **3,153**
46	544	**590**	**200**	2,723	**2,923**
$n_{1,1}$ = 170	1,045	N_1 = 1,215	$n_{2,1}$ = 555	5,521	N_2 = 6,076
$p_{1,1}$ = 0.7924	0.4794	$\hat{p}_1$ = 0.5144	$p_{2,1}$ = 0.6396	0.5068	$\hat{p}_2$ = 0.5189

Table 6. For a goodness-of-fit test, we compare an observed column with the expected total column.

Expressed as a z test, we must determine whether an observed proportion in column 1, $p_{1,1} = f_{1,1}/n_{1,1}$, differs from the same proportion for the entire set, $\hat{p}_1 = F_1/N_1$. In Table 6, we have $p_{1,1} = 0.7294$ and $\hat{p}_1 = 0.5144$; $p_{2,1} = 0.6396$ and $\hat{p}_2 = 0.5189$.

We obtain the following differences between probabilities drawn from column 1 and the 'total'.

$$d_1 = p_{1,1} - \hat{p}_1 = 0.2150, \text{ and}$$
$$d_2 = p_{2,1} - \hat{p}_2 = 0.1207.$$

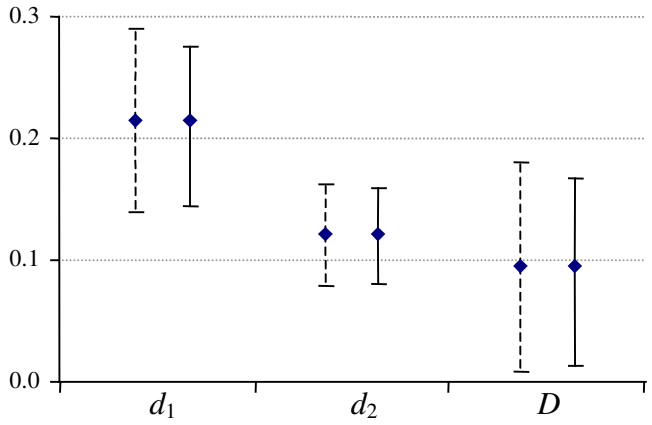

Figure 6: Gaussian (dashed) and Wilson 95% confidence intervals for 2 × 1 goodness of fit tests, with the difference intervals derived from them.

In this test condition we compare differences between d_1 and d_2. This is still a 'difference of differences test', not a point test (comparing $p_{1,1}$ and $p_{2,1}$).

However, like the point test, variation exists only at observed probabilities, e.g. $p_{2,1}$. The intervals for d_1 and d_2 are computed assuming the expected prior is 'true'.

5.1 The 2 ×1 goodness of fit χ^2 test

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 $e_1 = z_{\alpha/2}.s_1 = 0.0751$ and $e_2 = 0.0416$ at the 0.05 error level. These errors are smaller than the equivalent differences, and each individual test is significant (Figure 6).

These samples are drawn from independent populations, so to evaluate the difference of these differences we employ the sum of variances rule (6) to combine e_1 and e_2. This gives us a confidence interval of $e_D = (-0.0859, +0.0859)$. The difference of differences, $D = d_1 - d_2 = -0.1207$, is outside this interval and is therefore significant.

However, just as the Newcombe-Wilson test is to be preferred over the standard z test for independent population proportions, a Wilson-based approach is liable to be more accurate and informative (data should be plotted with Wilson score intervals). We turn to this model next.

5.2 The 2 ×1 Wilson interval test

We compute Wilson intervals for each observed probability $p_{1,1}$ and $p_{2,1}$ (recall that the second subscript '1' refers to the first column in each table in Table 3). The Wilson score intervals for each single difference $d_1 = p_{1,1} - \hat{p}_1$ (etc.) are then

$$(w_1^-, w_1^+) = (0.6581, 0.7906); \text{ and } (w_2^-, w_2^+) = (0.6785, 0.5989).$$

Again, we may combine intervals using equation (6). As the difference is positive, the inner side of the interval is simply based on the lower bound of $p_{1,1}$ and the upper bound of $p_{2,1}$. We may substitute $e_1^- = p_{1,1} - w_{d_1}^-$ and $e_2^+ = w_{d_2}^+ - p_{2,1}$ into Zou and Donner's equation:

$$w_D^+ = \sqrt{(e_1^-)^2 + (e_2^+)^2} = \sqrt{0.0713^2 + 0.0389^2} = 0.0812.$$

Similarly the lower bound $w_D^- = \sqrt{0.0612^2 + 0.0408^2} = 0.0735$.

We obtain an interval of $(-w_D^-, w_D^+) = (-0.0735, 0.0812)$. $D = 0.1207$ is outside this range, and the difference is significant.

5.1 r × *1 goodness of fit tests*

We can generalise the paired 2×1 goodness of fit test to a paired $r \times 1$ goodness of fit χ^2 test with $r - 1$ degrees of freedom. The Gaussian formula for the goodness of fit test uses the Normal approximation to the Binomial about cells in the 'total' column.

The 2×1 test can be rewritten to give us a z statistic. In the following t and j represent the test and column indices respectively.

$$z_D(j) \equiv \frac{\left(d_{1,j} - d_{2,j}\right)}{\sqrt{s_{1,j}^2 + s_{2,j}^2}},$$

where $d_{t,j}$ represents the difference between the probability in column j in test t and the probability in the equivalent 'total' column (the 'pooled probability estimate', $\hat{p}_{t,}$). See equation (4). The standard deviation, $s_{t,j}$ is computed with the Gaussian $\sqrt{\hat{p}_t(1-\hat{p}_t)/n_{t,j}}$.

To extend this test to an arbitrary number of rows, r, we will need to evaluate the sum of these scores. If we use i to represent a row, (i.e. $i = 1 \ldots r$), the equation above becomes

$$z_D(j) \equiv \frac{\left(d_{1,i,j} - d_{2,i,j}\right)}{\sqrt{s_{1,i,j}^2 + s_{2,i,j}^2}},$$

where

- *difference* (observed – expected) $d_{t,i,j} \equiv p_{t,i,j} - \hat{p}_{t,i}$,
- *variance* $s_{t,i,j}^2 \equiv \hat{p}_{t,i}(1 - \hat{p}_{t,i}) / n_{t,i,j}$,
- *expected (pooled) probability* $\hat{p}_{t,i} \equiv F_{t,i} / N_{t,i}$.

The summation is similar to that applied to the $r \times 2$ test for homogeneity. Pearson's χ^2 is the square of the z distribution extended over any number of degrees of freedom (Wallis 2013b). We may convert a sum of r squared z scores to a chi-square score. We obtain equation (12), a formula whose form we have already seen, although in this case the terms refer to those above.

$$\chi_D^2 \equiv \tfrac{1}{2}\sum_{i=1}^{r} \frac{\left(d_{1,i} - d_{2,i}\right)^2}{s_{1,i}^2 + s_{2,i}^2}. \qquad (12')$$

Observant readers may note that this formula differs from Pearson's χ^2 by computing variance with the Gaussian $s^2 = p(1 - p) / n$, and dividing by 2. We divide by two because the summation 'sums the z test twice'. The test is evaluated against the critical value of χ^2 with $r - 1$ degrees of freedom.

6. Heterogeneity χ^2 tests

The tests we have performed thus far do not assume that data from each subtest is drawn from the same population. On the contrary, we do not know whether data derives from the same source, or is sampled in the same way, or that variables that we use to refer to linguistic concepts are operationalised in the same manner (Aarts *et al*. 2013).

An alternative approach to the gradient tests described here is Zar's heterogeneity analysis (1999: 471, 500) method.[14] We therefore discuss this test as a potential alternative candidate to ours.

Zar's formula is very simple. It can be summarised as

$$\chi^2_{\text{het}} \equiv \chi^2_{\text{sum}} - \chi^2_{\text{pool}} \tag{14}$$

where χ^2_{sum} is simply the sum of individual χ^2 tests and χ^2_{pool} the result of a χ^2 test 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 is generalisable to $t>2$ test runs (the degrees of freedom being multiplied by $t - 1$).

First lets us consider how Zar's method performs with the paired 2×2 homogeneity test (section 3). The pooled table sums cells across the two tables (the first cell is $f_{1,1}+f_{2,1}$ etc.). Using our example data (Table 3) we obtain $\chi^2_{\text{pool}} = 65.49$, $\chi^2_{\text{sum}} = 36.58+35.65$. To four decimal places the heterogeneity chi-square $\chi^2_{\text{het}} = 6.7394$, which is significant at the 0.05 error level.

Our Gaussian gradient test (see section 3.1) can be converted to arrive at a χ^2 statistic with one degree of freedom for comparative purposes using equation (14).

$$\chi^2_D = \left(D \cdot z_{\alpha/2}/e_D\right)^2 \tag{15}$$

This obtains $\chi^2 = 6.2273$. How do we explain this difference?

Zar's method does not reliably measure significant difference in cases where tables have different prior probabilities $\hat{p}$ (or distribution of row totals). It assumes that data is drawn from the same population, and then tests whether this is the case! Our tests do not make this assumption.

The easiest way to see this is to examine the simplest test of all, the 2×1 goodness of fit test, a test that is particularly sensitive to the prior distribution.

Consider two 2×1 tables defined so that the total number of cases $N_1 = N_2 = 100$. These tables have one degree of freedom, whether we are considering tests of goodness of fit, as here, or homogeneity. We will consider two further parameters for each table: skew π and prior $\hat{p}$.

We can define a general contingency table $\mathbf{T}(N, \pi, \hat{p})$ as:

$N\hat{p}\,\alpha$	$N\hat{p}\,(1-\alpha)$	$N\hat{p}$
$N(1-\hat{p})(1-\alpha)$	$N(1-\hat{p})\alpha$	$N(1-\hat{p})$
$N(1-\hat{p}-\alpha+2\alpha\hat{p})$	$N(\hat{p}+\alpha-2\alpha\hat{p})$	N

where $\alpha = (\pi+1)/2$ and the last row and column are totals. By creating two tables $\mathbf{T_1}$ and $\mathbf{T_2}$ we can then compare the effect of different values of N, π and $\hat{p}$ to create testing conditions for comparing the performance of equations (14) and (15). Both equations obtain identical results if the prior is identical, i.e.

$$\chi^2_{\text{het}} = \chi^2_D \qquad \text{if } \hat{p}_{1,1} = \hat{p}_{2,1}.$$

Zar's test assumes that the priors are identical, i.e., *as if they come from the same population*. It does not accurately measure separability when priors are different.

Changing N or π in either case does not affect this equality. However, if we consider the substitutions $\mathbf{T_3}(100, 0.5, 0.5)$ and $\mathbf{T_4}(100, 0.5, 0.1)$, we can clearly see the problem. The second table has a visibly different prior distribution ({90,10}) than the first.

	37.5	12.5	**50**		**67.5**	22.5	**90**
$\mathbf{T_3}$ =	**12.5**	37.5	**50**	$\mathbf{T_4}$ =	**2.5**	7.5	**10**
	50	50	100		70	30	100

Zar's test obtains $\chi^2_{sum} = 15.7143$, $\chi^2_{pool} = 17.5$, and therefore $\chi^2_{het} = -1.7857$.

So, according to Zar, these two tests are not significantly separable. Indeed the negative sign implies that the question is not even applicable!

However, if we apply our gradient methods we obtain a different result. Equation (13) obtains $\chi^2_D = 5.4870$, significant at $\alpha = 0.05$. We obtain significantly different results with both Gaussian (equation 6) and Wilson formulae (7), and can plot graphs with confidence intervals, etc.[15]

This difference in outcome is not due to the absence of a Yates' correction being applied to the χ^2 (note that the number of cases, *N*, plays no role). It is entirely due to differences in prior probabilities.

7. Conclusions

Researchers often wish to make statements about their results relative to others. However, as Goldacre (2011) notes, experimental science papers in highly prestigious journals frequently fail to perform this analysis correctly. As we have discussed, this is because a difference in effect sizes, or 'difference of differences', is a stochastic property that can only be properly evaluated by a statistical test.

It is far preferable to cite standardised effect sizes such as Cramér's ϕ over values of the χ^2 test statistic (or '*p* values' computed from them), because the former normalises the statistic to a probabilistic scale, making comparison more straightforward. However, whereas ϕ may be cited, an observed difference between values of ϕ is not necessarily a significant one, even if both tests are found to be independently significant.

The common practice of citing χ^2 scores or error levels combines two distinct concepts: the size of the effect (e.g. ϕ) and the size of the data (*N*). The correct approach is to construct significance tests for the comparisons you wish to make.

One approach, applicable to homogeneity (independence) tests, is to employ a 'point test', or 'multi-point test', to compare points or datasets. This allows us to claim that the data in one table (sample) is distributed significantly differently than in another table. A point test is an independent homogeneity test expressed across the dependent variable and table. The multi-point test simply sums these tests along an independent variable.

Why then might we wish to employ a 'gradient test'? There are two circumstances. First, because we have already expressed experimental claims in terms of change: "the dependent variable has increased/decreased in our data". We wish to compare this claim with other similar claims in the literature. Second, in order to compare goodness of fit test results, the only option is a gradient test.

With one degree of freedom, effect sizes differ in one dimension: there is a single difference of differences. If this difference is greater than a certain limit then we can say that 'the difference is significant', i.e. the difference of differences is non-zero, within a given probability of error.

We derived a method for comparing values of Cramér's ϕ. This method requires a complex Gaussian approximation that is vulnerable to inaccuracy in small skewed datasets. Moreover, ϕ averages change across the diagonal and disposes of directional information. As there is only one

degree of freedom in comparing 2 × 2 tables, the ϕ test can be replaced by either of the other two tests plus the point test. With the fewest assumptions, the Wilson-based test retains the most information and is likely to be more accurate than the Gaussian test.

We also demonstrated how the same approach can be applied to 2 × 1 goodness of fit tests (where a gradient test is mandatory): here the Wilson method is to be strongly preferred. The methods described here are included in the online spreadsheet (see note 1) and users are encouraged to experiment with these.

For all tests except ϕ, we proposed generalisations by employing χ^2 to generalise over multiple differences of differences.

In reviewing the performance of our methods, we discussed heterogeneity χ^2 analysis (Zar 1999). This test is conceived in terms of the legitimacy of pooling samples (hence the interest in scalability to *m* samples), whereas we wish to determine whether one experimental outcome is significantly different from another. Zar obtains the same results as ours if the prior distribution for each table is identical, but in all other cases the results differ.

This might be a positive attribute – in our case we have to consider point tests as well as gradient tests – if it were not for the fact that Zar's method, whilst simple, is not robust. The method obtains negative (meaningless) χ^2 results in some cases and substantially higher estimates than ours in others. As such, this method can only be tentatively recommended for the more limited purpose Zar suggests: of deciding on the legitimacy of pooling results from multiple sources. To reliably test for significant differences between experimental runs we need to employ the methods derived in this paper.

We end with a caveat. In this paper we compared solutions to the problem of comparing the results of two contingency tests. As we noted at the outset, a statistical method is an adjunct to a process of experimental refinement based on underlying theoretical principles. Significance tests cannot fix problems of poor experimental design, although they may draw attention to them. They cannot guarantee that the phenomenon measured has real theoretical meaning, that baselines for comparison are meaningful, or that observations are free to vary.

Notes

1. A spreadsheet including all the tests discussed in this paper is at www.ucl.ac.uk/english-usage/statspapers/2x2-x2-separability.xls. As the explanation in this paper is quite involved we would recommend downloading it.
2. Aarts *et al.* (2013) break this comparison into a series of different intermediate experimental design changes to identify where precisely the significant difference in results arises.
3. Some statisticians have correctly pointed out that the Binomial model, and thus χ^2, *z*, etc., rely on the assumption that data is drawn from a random sample where each datum is independent, rather than (as is common in corpus linguistics) a random sample of contiguous texts. This may be addressed by adapting variance for random-text sampling using the method outlined in http://corplingstats.wordpress.com/2015/09/22/adapting-variance. This adjustment, like the continuity correction, can be combined with the methods outlined in this paper.
4. This extremely common mistake introduces substantial errors in practice (Newcombe 1998; Wallis 2013a). It also produces absurd results, such as projecting possible values of *P* outside of the range [0, 1], which leads to further conservative injunctions to experimentalists to avoid highly-skewed values.
5. 'Continuity-corrected' versions of these intervals are given in Wallis (2013a). This correction is usually recommended with small samples, and it increases the interval width to compensate for the

fact that the continuous Normal distribution is being approximated to the discrete Binomial. It therefore makes the test slightly more conservative. We do not present continuity-corrected formulae in this paper, primarily for reasons of presentation, but also because it is not common to use them in meta-testing. However substitution of the Gaussian and Wilson errors with the continuity corrected formulae is simple.

6. Like Yates' χ^2, this can also employ a continuity correction, and, so corrected, outperforms other methods (except, arguably, computationally intensive 'exact' methods).

7. Compare equations (3) and (6). Newcombe's reasoning is that at the lower bound of the difference, the best estimate of P_1 and P_2 is at w_1^+ and w_2^- respectively (and vice versa for the upper bound).

8. When applying multiple tests, researchers should divide the error level α by the number of independent tests to be carried out. We do not do this here for reasons of simplicity.

9. Since data is drawn from two independent populations, strictly speaking the Wilson method is to be preferred, but it does not make a substantive difference to the result.

10. What is being attempted is *not* a three dimensional chi-square test, e.g. a 2×2×2 test (Zar 1999: 506). That test has three degrees of freedom, one for each dimension. It aggregates differences within each table with differences between tables. Here we focus only on differences *between* tables.

11. In Figure 4 we have centred these intervals on the observation, rather than the zero axis. The upper bound of D, w_D^+, is therefore on the lower side of D, i.e. instead of testing $D > w_D^+$, we test $D - w_D^+ > 0$.

12. Summing over the independent variable (rows) may make residuals or 'chi-square contributions' misleading, but the sum will be the same.

13. This is the rule that says that where an expected cell in a table has a frequency of less than 5 it should be collapsed.

14. I am grateful to Stefan Gries for referring me to this test.

15. Difference in differences $D = 0.1857$, Gaussian interval = (-0.1554, 0.1554), Wilson interval (-0.1624, 0.1429).

References

Aarts, B., Close, J, and Wallis, S.A. 2013. 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*. 14-43. Cambridge: CUP.

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

Gelman, A. and Loken, E. 2013. *The garden of forking paths*. Columbia University. Published at: www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

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

Nieuwenhuis, S., Forstmann, B.U. and Wagenmakers, E.-J. 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience* 14: 1105–1107

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, S.A. 2012. Measures of association for contingency tables. London: Survey of English Usage, UCL. Available at: http://www.ucl.ac.uk/english-usage/statspapers/phimeasures.pdf

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

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

Wallis, S.A. forthcoming. Detecting direction in interaction evidence. London: Survey of English Usage. Available at: www.ucl.ac.uk/english-usage/statspapers/detecting-direction.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.