Journal of Quantitative Linguistics, 2013
Vol. 20, No. 4, 350–378, http://dx.doi.org/10.1080/09296174.2013.830554

Routledge
Taylor & Francis Group

z-squared: The Origin and Application of χ^2*

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

ABSTRACT

A set of statistical tests termed *contingency tests*, of which χ^2 is the most well-known example, are commonly employed in linguistics research. Contingency tests compare discrete distributions, that is, data divided into two or more alternative categories, such as alternative linguistic choices of a speaker or different experimental conditions. These tests are highly ubiquitous, and are part of every linguistics researcher's arsenal. However, the mathematical underpinnings of these tests are rarely discussed in the literature in an approachable way, with the result that many researchers may apply tests inappropriately, fail to see the possibility of testing particular questions, or draw unsound conclusions. Contingency tests are also closely related to the construction of *confidence intervals*, which are highly useful and revealing methods for plotting the certainty of experimental observations. This paper is organized in the following way. The foundations of the simplest type of χ^2 test, the 2×1 goodness of fit test, is introduced and related to the z test for a single observed proportion p and the Wilson score confidence interval about p. We then show how the 2×2 test for independence (homogeneity) is derived from two observations p_1 and p_2 and explain when each test should be used. We also briefly introduce the Newcombe-Wilson test, which ideally should be used in preference to the χ test for observations drawn from two independent populations (such as two sub-corpora). We then turn to tests for larger tables, generally termed $r \times c$ tests, which have multiple degrees of freedom and therefore may encompass multiple trends, and discuss strategies for their analysis. Finally, we turn briefly to the question of differentiating test results. We introduce the concept of *effect size* (also termed "measures of association") and finally explain how we may perform statistical *separability tests* to distinguish between two sets of results.

*Address correspondence to: Sean Wallis, Survey of English Usage, Department of English, University College London, Gower Street WC1E 6BT, London, England. Email: s.wallis@ucl.ac.uk

© 2013 The Author(s). Published by Taylor & Francis.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/3.0/, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The moral rights of the named author(s) have been asserted.

1. INTRODUCTION

Karl Pearson's famous chi-square contingency test is derived from another statistic, called the *z* statistic, based on the Normal distribution. The simplest versions of χ^2 can be shown to be *mathematically identical* to equivalent *z* tests. The tests produce the same result in all circumstances.[1] For all intents and purposes "chi-squared" could be called "*z*-squared". The critical values of χ^2 for one degree of freedom are the square of the corresponding critical values of *z*.

The standard 2×2 χ^2 test is another way of calculating the *z* test for two independent proportions taken from the same population (Sheskin, 1997, p. 226).

This test is based on an even simpler test. The 2×1 (or 1×2) "goodness of fit" (g.o.f.) χ^2 test is an implementation of one of the simplest tests in statistics, called the Binomial test, or population *z* test (Sheskin, 1997, p. 118). This test compares a sample observation against a predicted value which is assumed to be binomially distributed.

If this is the case, why might we need chi-square? Pearson's innovation in developing chi-square was to permit a test of a larger array with *multiple values greater than 2*, i.e. to extend the 2×2 test to a more general test with *r* rows and *c* columns. Similarly the *z* test can be extended to an $r \times 1$ χ^2 test in order to evaluate an arbitrary number of rows. Such a procedure permits us to detect significant variation across multiple values, rather than rely on two-way comparisons. However, further analysis is then needed, in the form of 2×2 or 2×1 g.o.f. χ^2 tests, to identify *which* values are undergoing significant variation (see Section 3).

The fundamental assumption of these tests can be stated in simple terms as follows. An observed sample represents a limited selection from a much larger population. Were we to obtain multiple samples we might get slightly different results. In reporting results, therefore, we need a measure of their *reliability*. Stating that a result is significant at a certain level of error ($\alpha = 0.01$, for example) is another way of stating that, were we to repeat

[1]The known limitation of χ^2, which states that results cannot be relied upon if an expected cell frequency is less than 5, has its interval equivalent. It also has the same solution, namely to replace 'Wald' confidence intervals with the more accurate Wilson score confidence interval (Wallis, 2013). We return to this issue in Section 6.

the experiment many times, the likelihood of obtaining a result other than that reported will be below this error level.

2. THE ORIGIN OF χ^2

2.1. Sampling Assumptions

In order to estimate this "reliability" we need to make some mathematical assumptions about data in the population and our sample.

The concept of the "population" is an ideal construct. An example population for corpus research might be "all texts sampled in the same way as the corpus". In a lab experiment it might be "all participants given the same task under the same experimental conditions". Generalizations from a corpus of English speech and writing, such as ICE-GB (Nelson et al., 2002), would apply to "all similarly sampled texts in the same proportion of speech and writing" – not "all English sentences from the same period" and so forth. Deductively rationalizing beyond this population to a wider population is possible – by arguing why this "operationalising" population is, in the respect under consideration, *representative* of this wider population – but it is not given by the statistical method.

2.1.1. *Randomness and Independence*

The first assumption we need to make is that the sample is *a random sample* from the population, that is, each observation is taken from the population at random, and the selection of each member of the sample is independent from the next. A classical analogy is taking a fixed number of mixed single-colour billiard balls (say, red or white) from a large bag of many balls.

Where we are compelled to break this independence assumption by taking several cases from the same text (common in corpus linguistics), at *minimum* we need to be aware of this and consider the effect of clustering on their independence (see Nelson et al., 2002, p. 273 and Section 3.3). Ideally we should be able to measure and factor out any such clustering effects. Currently methods for such "case interaction" estimation are work in progress.

2.1.2. *The Sample is Very Small Relative to the Population*

The second assumption is that the population is much larger than the sample, potentially infinite. If the sample was, say, half the size of a finite

Table 1. An example 2 × 2 contingency table.

	a	$\neg a$	Σ
b	20	5	25
$\neg b$	10	10	20
Σ	30	15	45

population "in the bag", we would know that half the population had the observed distribution of our sample, and therefore we should have a *greater* confidence in our estimate of the distribution of the entire population than otherwise. In such circumstances, using a z or χ^2 test would tend to underestimate the reliability of our results. In linguistics this assumption is only broken when generalising from a large subset of the population – such as treating Shakespeare's First Folio as a subset of his published plays.[2]

2.1.3. *Repeated Sampling Obtains a Binomial Distribution*

The third assumption of these tests is perhaps the most complicated to explain. This is that repeated sampling from the population of a frequency count will build up into a Binomial frequency distribution centred on a particular point, and this distribution may be approximated by the Normal distribution.

Suppose we carry out a simple experiment as follows. We sample 45 cases over two Boolean variables, $A = \{a, \neg a\}$ and $B = \{b, \neg b\}$, and obtain the values $\{\{20, 5\}, \{10, 10\}\}$ (Table 1). We will take A as our *independent variable*, and B as our *dependent variable*. This means that we try to see if A affects the value of B, schematically, $A \rightarrow B$.

This kind of table might summarise the results of an experiment measuring a speaker's tendency to employ, say, modal *shall* rather than *will* in first person singular cases (so b stands for *shall* and $\neg b$ for *will*), in a spoken rather than written English sample (a = spoken, $\neg a$ = written). For this discussion we will use invented data to keep the arithmetic simple.

Next, imagine that we repeated our experiment, say, 1000 times, to obtain a "sample of samples". The more repetitions we undergo, the greater will be our confidence that the average result will be close to the "correct" average (if we could measure it) in the population. Sheskin (1997, p. 37)

[2]A standard adjustment is Singleton et al. (1988) correction, multiplying the standard deviation by $v = \sqrt{1 - n/N_P}$, where n/N_P is the ratio of the sample to population size. If N_P is infinite, $v = 1$.

explains that "the standard error of the population mean represents a standard deviation of a sampling distribution of means." This "standard error of the population mean" is also a theoretical value. The tests we discuss here estimate this value from the standard deviation calculated from a single sample.

The Binomial model states that the result for any single cell in our table will likely be distributed in a particular pattern derived from combinatorial mathematics, called the Binomial distribution, centred on the population mean. This pattern is represented by the columns in Figure 1.[3] The frequency axis, F, represents the number of times a value is predicted to have a particular outcome on the x axis, assuming that each sample is randomly drawn from the population.

The Binomial distribution is a *discrete* distribution, that is, it can have particular integer values, hence the columns in Figure 1. Cell frequencies must be whole numbers. According to the Central Limit Theorem this may be approximated to a *continuous* distribution: the Normal or "Gaussian" distribution, depicted by the curve in Figure 1.[4] Note that in inferring this distribution we are not assuming that the linguistic sample is Normally distributed, but that very many samples are taken from the population, randomly and independently from each other.

A Normal distribution can be specified by two parameters. The observed distribution O[$\bar{x}$, s] has a *mean*, $\bar{x}$ (the centre point), and *standard deviation*, s (the degree of spread). The Normal distribution is *symmetric*, with the mean, median and mode coinciding.

These distributions are *sampling models*, i.e. mathematical models of how future samples are likely to be distributed, based on a single initial sample. The heart of *inferential* statistics is attempting to predicting how future experiments will behave, and our confidence that they will behave similarly to our current experiment. We can now use the Normal

[3]Suppose we toss a coin twice and count the number of heads, h. There are four possible outcomes: *HH*, *HT*, *TH*, *TT*. With an unbiased coin, the most likely value of h will be 1, because there are two ways to achieve this result. On the other hand the chance of h being 0 or 2 will each be ¼. We can summarise the expected distribution after four repetitions as $F(h) = \{1, 2, 1\}$. We say that the variable h forms a Binomial distribution centred on the mean.

[4]It *is* possible to calculate significance using only the Binomial distribution (Sheskin, 1997, p. 114) or Fisher's 2 × 2 test (ibid. 221; see also Wallis, 2013) – but these tests require combinatorial calculations which are onerous without a computer. The Binomial approximation to the Normal has therefore traditionally proved attractive.

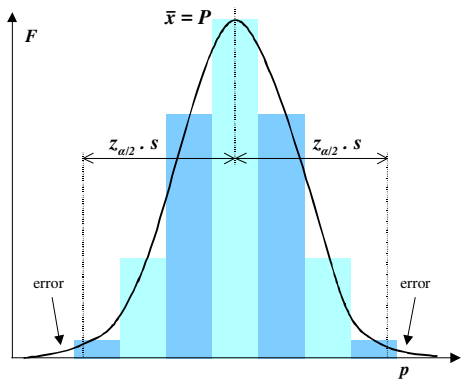

Fig. 1. Binomial approximation to a Normal frequency distribution plotted over a probabilistic range $p \in [0, 1]$.

distribution model to "peek into the future" and estimate the reliability of our single sample.

2.2. The "Wald" Confidence interval

We are going to approach our discussion of χ^2 from the perspective of defining *confidence intervals*. Approaching the problem this way makes it easier for us to visualise why χ^2 is defined in the way that it is – as an alternative calculation for estimating confidence and certainty – and therefore what statistically significant results may mean.

A confidence interval is the range of values that an observation is likely to hold given a particular probability of error (say, $\alpha = 0.05$). This can also be expressed as a degree of confidence: "95%". What we are interested in is how far an observation must deviate from the expected value to be deemed to be statistically significantly different from it at a given error level (exactly as we would with a χ^2 test). Consider Figure 1. The area under the curve adds up to 100%. The two tail areas under the curve marked "error" represent extreme values. Suppose we find the tail areas that each cover 2.5% of the total area. We then have a range between them inside which 95%, or 1 in 20 experimental runs, would fall. If we insist on a smaller error ($\alpha = 0.01$) then these tails must be smaller (0.005 or 0.5%) and the interval must be larger.

Suppose that we performed an experiment and obtained the results in Table 1, {{20, 5}, {10, 10}}. Consider the first column, a = {20, 10},

Table 2. Dividing by column totals rewrites Table 1 in terms of probabilities.

	a	$\neg a$	Σ
b	$^2/_3$	$^1/_3$	$^5/_9$
$\neg b$	$^1/_3$	$^2/_3$	$^4/_9$

which in our experiment represents spoken data. Out of $n = 30$ observations, 20 are of type b (*shall*) and 10, $\neg b$ (*will*). The probability of picking type b at random from this set is equal to the proportion of cases of type b, so $p = {}^2/_3$. The probability of choosing type $\neg b$ given a is the remaining probability, $q = 1 - p = {}^1/_3$ (Table 2).

The first method we will discuss for creating a confidence interval about p is the 'Wald' method, which is almost universally recommended, but is based on a misapplication of the Central Limit Theorem. It assumes that the population mean probability, which we will denote with P, is Normally distributed around the observed probability p. This is the *inverse relationship* from that shown in Figure 1, which is centred around P.

We may calculate a 'normalised' Wald interval (one standardised to the probabilistic range) by the following steps. We will employ an error level, α, of 0.05.

$$\begin{aligned}
&\textit{probabilistic mean} && \bar{x} \equiv p && = 0.667,\\
&\textit{standard deviation} && s \equiv \sqrt{pq/n} && = 0.086,\\
&\textit{error}(\alpha = 0.05) && e \equiv z_{\alpha/2} s && = 0.169,\\
&\textit{Wald interval} && (\bar{x} - e, \bar{x} + e) && = (0.498, 0.835). \qquad (1)
\end{aligned}$$

where the term $z_{\alpha/2}$ is the critical value of the Normal distribution for the given error level. Common values are $z_{\alpha/2} = 1.95996$ ($\alpha = 0.05$) and 2.57583 ($\alpha = 0.01$). Note that as n increases, the standard deviation falls. More data means a narrower interval, and increased confidence, as we would expect. The predicted range of the population mean P is between 0.498 and 0.835.

It is now straightforward to scale this standardized interval by the data, i.e. multiply the above by n, to obtain the standard deviation for the first cell, b. The values for the second cell, $\neg b$, have the *same* standard deviation: $\bar{x} = 10$ and $s = 2.58$. Swapping p for q does not affect the formula for s. The Wald interval (scaled) = (14.94, 25.06).

Were we to repeat the same sampling exercise many times, we would expect that only in $^1/_{20}$ of cases would the value in the upper left cell in Table 1, $F(a, b)$, fall outside the range (14.94, 25.06). (Of course samples will be represented in terms of integer frequency values; these decimal fractions derive from the continuous nature of the Normal distribution.)

However, this approach, whilst extremely common, is incorrect. The error lies in the assumption that the population mean, P, is Binomially (and approximately Normally) distributed around the observation, p.

Wallis (2013) argues that the correct way to think about the confidence interval on p is by considering what the observation p tells us about likely values of the population probability, P. To do this we need to estimate the Binomial distribution about P for n observations. If p is at the upper bound of P, then P must be at the lower bound of p, and vice versa.

In a goodness of fit test we compare two values, one of which we take as the predicted value, P and the other our observation p. We can calculate the Wald interval about P and then test that p is far enough away to be significantly different. However to calculate a confidence interval about an observation p we should not use this Wald formula. A better, but less well-known, formula for computing confidence intervals about p is discussed in Section 2.4 below.

2.3. Single-Sample Population z tests and Goodness of Fit χ^2

The single-sample z test for a sample probability (Sheskin, 1997, p. 34) simply compares a given value, p, with a Normal (Wald) interval about P. Values of p inside the interval do not overturn the null hypothesis that there is no difference between the predicted result P and the sample. Those outside the range are significant at the chosen error level.

The z test can be applied to any value of P, and is mathematically equivalent to the 2×1 χ^2 goodness of fit test, in conditions where P is simply the mean across all values in a dataset, $P = p(b) = {}^5/_9$. The term "goodness of fit" (g.o.f.) refers to the fact that we compare one distribution for its consistency with a given expected distribution. In the simplest case the distribution has two cells, but may be extended to multinomial variables.

We will demonstrate how the test is computed using both methods.

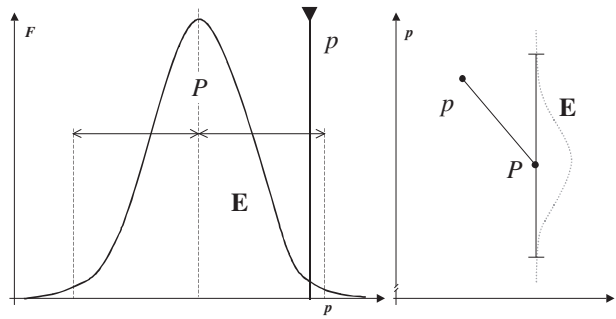

Fig. 2. The single-sample population z test: left, comparing an observed $p = 0.667$ with the population mean confidence interval (0.378, 0.733). The confidence interval (dotted) may be plotted as an I-shaped error bar in an equivalent graph (right).

2.3.1. *The single sample z test*

The null hypothesis of the z test is that an observation p is consistent with the expected value P. The χ^2 test uses the distribution over the total column (Table 1, column marked Σ) as its expected distribution, scaled to the same number of cases as the observed column (in our case, a). This distribution is {25, 20}, or, normalised, $\{{}^5/_9, {}^4/_9\} \cong \{0.556, 0.444\}$.

We calculate the interval at P, so $s = \sqrt{(0.556 \times 0.444/30)} = 0.091$. We then compare our expected interval $P \pm z.s = 0.556 \pm 0.178$ (all data) with the observation given a (spoken) – so $p = p(b \mid a) = 0.667$. Note that the sample size n is the observation sample size (30) for a. After all, this is the data supporting the observation. The test is significant if p falls outside the confidence interval of the expected distribution $\mathrm{E}[\bar{x}, s]$. In our case the result is not significant.

2.3.2. *The 2 × 1 Goodness of fit χ^2 test*

An alternative calculation employs the χ^2 formula:

$$\chi^2 = \sum \frac{(o-e)^2}{e}, \tag{2}$$

where the observed distribution $o = \{20, 10\}$ and expected distribution is based on the total column, $e = \{25, 20\} \times 30/45 = \{{}^5/_9, {}^4/_9\} \times 30$. This test result obtains $\chi^2 = 0.667 + 0.833 = 1.5$, which is below the critical value of χ^2 with one degree of freedom, and is not significant.

In the χ^2 test we sum terms calculated from both cells. In effect, we also compare $p(\neg b \mid a)$ with $x = p(\neg b) = 0.444$. We don't need to do this in

a z test because these values depend entirely on the other cell values ($p(\neg b) \equiv 1 - p(b)$, etc). *The significance of the result obtained for the lower cell is entirely dependent on the significance test for the upper.* This is what is meant by saying "the test has one degree of freedom". It is only necessary to carry out a single comparison because the second comparison is mathematically determined by the first.[5]

Both tests assume that the population probability is correctly estimated by the average probability $p(b)$. Only the subset column is free to vary, so the confidence interval is calculated using the number of cases n in that column. In section 3 we discuss when this test is appropriate.

2.4. The Wilson score interval

Calculating the confidence interval for p accurately requires a different approach than the "Wald" approach, which although common, is based on a misconception. The nub of the problem is, quite simply, that the confidence interval around a given observation taken from a Binomially distributed population *is not itself Binomially distributed*. With small volumes of data, or where cell values and hence probability are skewed, *the interval does not approximate to the Normal distribution*.

As a result, if p is very close to zero or 1, a Wald interval can exceed the probabilistic range [0, 1]. This cannot make sense: how can a probability be negative – or greater than 1?

To attempt to address this, researchers are told to not use the interval for highly skewed-values, leading to rules like "the normalised mean be more than 3 standard deviations from 0 or 1".[6] However, highly skewed distributions are common in linguistics: we may well be interested in the behaviour of low frequency terms (particularly in lexical studies), and data is often hard to come by. Clearly, simply ceasing an analysis simply due to the presence of unevenly distributed data would be very restrictive.

The correct approach uses the Wilson score interval about an observation p (Wilson, 1927).

[5]Note that comparing $p(b \mid \neg a)$ with $p(b)$ is a different test and obtains a different result. This considers how closely the distribution for$\neg a$ (writing) matches that for the total column.

[6]This error is often conflated with the less severe question of *continuity correction*, which we end this section with. This leads to the 'Cochran rule' of avoiding using χ^2 with small expected cell values or arguments for using log-likelihood in preference to χ^2. Wallis (2013) compares the performance of a range of tests against comparable exact Binomial and Fisher tests and finds that this rule is overcautious and log-likelihood does not perform well.

$$\textit{Wilson score interval}(w^-, w^+) \equiv \left(p + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2} \sqrt{\frac{pq}{n} + \frac{z^2_{\alpha/2}}{4n^2}} \right) / \left(1 + \frac{z^2_{\alpha/2}}{n} \right). \quad (3)$$

We can use this interval to test to see if a point x falls outside the expected range (w^-, w^+). Wilson's interval about p inverts the Gaussian distribution about P (Wallis, 2013) and so obtains the same result as the single sample z test and 2 × 1 goodness of fit χ^2 test.

This method should be used for plotting confidence intervals on graphs, where we may plot values of p and error bars extending to w^- and w^+.

With p = 0.667, n = 30 and $z_{\alpha/2}$ = 1.95996, we obtain a score interval for the probability of using *shall* in speech, $p(b \mid a)$, of (0.488, 0.808). Figure 3 (left datapoint) has a narrower interval than the equivalent Wald interval (0.498, 0.835), which is slightly skewed toward the centre.

We may obtain a similar interval for the second 'writing' column, $p(b \mid \neg a)$, Figure 3 (right datapoint). Any number of points may be plotted with confidence intervals, and in time series data, a best-fit line may be plotted between them.

Wilson's score confidence interval is based on an asymmetric distribution, which is always restricted to the probability range [0, 1]. For p = 0.5, Wilson's interval is similar to the Wald confidence interval, as one might expect.

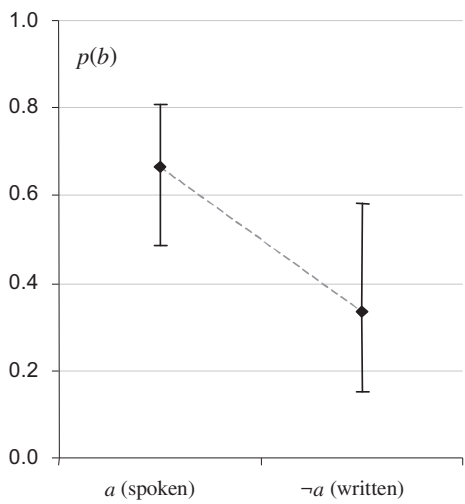

Fig. 3. Plotting Wilson score intervals on the probability of *b*.

2.5. 2 × 2 χ^2 and z test for two independent proportions

The z test for two independent proportions (Sheskin, 1997, p. 226) compares two observations to see if they are alike. This is equivalent to comparing columns a and $\neg a$ in Table 1 using a 2 × 2 χ^2 test. In this case both samples are free to vary (Figure 4). Suppose we use O_1 to represent the Normal distribution for p_1 in the first column of Table 2 (i.e. $p_1 = p(b \mid a)$ = probability of *shall* in spoken data). Similarly O_2 will stand for the distribution for written data, $p_2 = p(b \mid \neg a)$.

This z test combines distributions O_1 and O_2 into a single *difference* distribution D[0, s'] centred on 0. D represents the sampling distribution of the difference, d, between observed probabilities.

To carry out this test we must calculate the standard deviation of the difference, s', which depends on *the pooled probability estimate*, $\hat{p}$. In our contingency table this works out simply as the row total over the grand total $N = n_1 + n_2$, i.e. the overall probability of b, $p(b)$.

$$\textit{probability estimate} \quad \hat{p} \equiv (n_1 p_1 + n_2 p_2)/N,$$

and

$$\textit{standard deviation} \quad s' \equiv \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}. \tag{4}$$

The equivalent confidence interval is simply $\pm z_{\alpha/2}.s'$. Once s' has been estimated, carrying out the test becomes extremely simple. Since the distribution is symmetric and centred at zero we can dispense with signs. We simply test if $|d| > z.s'$ where $d = p_1 - p_2 = p(b \mid a) - p(b \mid \neg a)$.

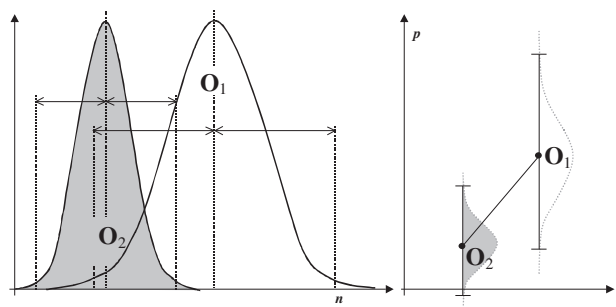

Fig. 4. The z test/2 × 2 χ^2 test assumes uncertainty in both observations $O_1[\bar{x}_1, s_1]$ and $O_2[\bar{x}_2, s_2]$.

Our data in Table 1 gives us the following:

$$\textit{probability estimate} \quad \hat{p} = (20 + 5)/45 = 0.556,$$

$$\textit{standard deviation} \quad s' = 0.157.$$

This gives us a confidence interval of ± 0.308 at the $p < 0.05$ level.

The difference in the observed probabilities $d = p_1 - p_2 = {}^2/_3 - {}^1/_3 = 0.333$. Since this exceeds 0.308, this result is significant. As before, it is possible to demonstrate that the result one obtains from this test is equivalent to that obtained from the 2 × 2 χ^2 test where the expected distribution is computed in the usual way (Section 3.4).

The 2 × 2 χ^2 test has one further useful property. It can be calculated by simply adding together the results of the two g.o.f. χ^2 values for a and $\neg a$. The degrees of freedom and critical value of χ^2 do not change. Consequently, a 2 × 1 g.o.f. χ^2 test is stricter than its corresponding 2 × 2 χ^2, which will be significant if *either* goodness of fit test is significant.

2.6. The z test for two independent proportions from independent populations

In the 2 × 2 χ^2 test it is generally assumed that a and $\neg a$ are drawn from the same population. For example, if a and $\neg a$ represented two different grammatical choices uttered by speakers/writers, then the same participant could genuinely choose to use one or other construction.

However in some cases it is more correct to characterise the z test in terms of sampling *different* populations. In our example we gave the example of speech versus writing: here participants and the texts they contribute to are categorised by the independent variable A group (also known as a "between subjects" design).

In this situation the two populations are considered to vary independently, so the standard deviation is taken to be simply the Pythagorean sum of the independent standard deviations $s' = \sqrt{s_1^2 + s_2^2}$ (Sheskin, 1997, p. 229) – rather than being based on the pooled probability.

If the independent variable divides the corpus by texts (a sociolinguistic variable or subcorpus) this test should be used in place of 2 × 2 χ^2. However, Sheskin's formula performs poorly, because it is based on the incorrect "Wald" standard deviation at p_1 and p_2. Newcombe (1998) demonstrated that a better approach uses the Wilson score interval for p_1 and p_2.

The idea is to combine both inner interval widths in Figure 3 using the Pythagorean formula. Since $p_2 < p_1$, the inner interval is composed of the lower bound of p_1 (w_1^-) and the upper bound of p_2 (w_2^+). If $p_2 > p_1$ then we would need to combine the other pair. This obtains the following "Newcombe-Wilson" interval against which we can test the difference $d = p_1 - p_2$.

$$NW\ interval\ (W^-, W^+) = \left(-\sqrt{(p_1 - w_1^-)^2 + (w_2^+ - p_2)^2}, \sqrt{(w_1^+ - p_1)^2 + (p_2 - w_2^-)^2}\right). \quad (5)$$

Using the data in Table 1 this obtains an interval of (−0.230, 0.307), which is slightly less conservative than the z test for samples from the same population. Recall that in this type of experiment we are assuming that our data is constrained into two independent samples where participants fall into one or other group, so distributions are more constrained.

2.7. Yates' correction, Log-likelihood and other methods

When we employed the Normal approximation to the Binomial distribution we approximated a discrete distribution (a set of possible values) into a continuous curve. By rounding this curve we introduced a small error into the calculation.

Yates' correction to the chi-square statistic performs a *continuity correction* by subtracting 0.5 from the absolute difference $|o - e|$ in the chi-square formula, and is recommended for 2×1 and 2×2 tests. A similar correction may also be applied to the Wilson score interval (Wallis, 2013), slightly increasing the width of the confidence interval (and can be applied to the Newcombe-Wilson interval). This correction is slightly conservative, meaning that in some borderline cases it may lead to rejecting hypotheses which would otherwise be accepted, but this is preferable to the alternative.

Another school of thought (e.g. Dunning, 1993) has advanced the case for contingency tests based on the likelihood ratio statistic, citing some demonstrable accuracy in selected cases. However, recent exhaustive evaluation using computation (by among others, Newcombe, 1998 and Wallis, 2013) has found that in fact log-likelihood performs less reliably than comparable χ^2 tests, and if caution is required a continuity corrected chi-square or Wilson interval is to be preferred.

Wallis (2013) demonstrates that Yates' "continuity corrected" χ^2 obtains a closer approximation to the Binomial distribution than standard χ^2, and

notes that Newcombe's continuity corrected interval also mirrors this adjusted Gaussian.

Finally note that for small samples, we should fall back on "exact" tests. These are the Binomial test for 2×1 goodness of fit tests and Fisher's exact test for 2×2 tests. For more information, see Wallis (2013), which also employs exact paired Binomial intervals for independent sample conditions.

3. THE APPROPRIATE USE OF χ^2

3.1. Selecting Tests

Figures 2 and 4 pictorially demonstrate the mathematical difference between the two types of χ^2 test we have discussed. Figure 6 summarises their different purpose.

(1) The goodness of fit test (Figure 6, left) can be used to examine variation in the distribution of a single value in a typological hierarchy, as a proportion of a super-ordinate value.

Conclusion: Use the g.o.f. test when you want to examine if a single value, a, has a different distribution than a superset $A = \{a, \neg a$ (e.g. modal *must* compared to all modal verbs). If a is affected by the independent variable B differently to A, then it may be worth reporting.

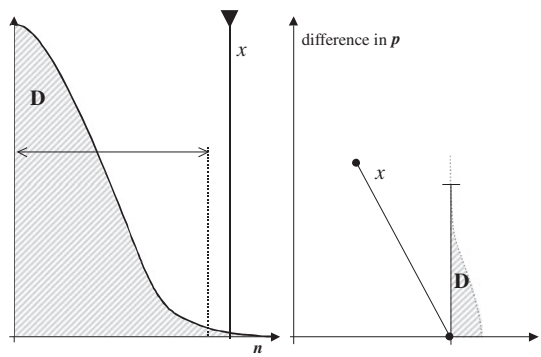

Fig. 5. The new *difference* confidence interval D[0, s'] centred on the origin, and the difference $x = |\bar{x}_1 - \bar{x}_2|$.

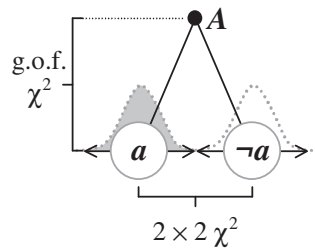

Fig. 6. Employing χ^2 tests for different purposes. The superset A is considered a fixed origin; a and $\neg a$ are free to vary; a and $\neg a$ are free to vary.

(2) The $2 \times 2\ \chi^2$ test (Figure 6, lower) examines variation within the set $\{a, \neg a\}$. It assumes that the expected distribution is averaged between two observed distributions. The standard deviation of the corresponding difference distribution is also a kind of average of the two distributions. The test takes variance over both columns, a and $\neg a$, into account.

Conclusion: Use the $2 \times 2\ \chi^2$ test when you want to examine if there is variation *within* the set $A = \{a, \neg a\}$ (e.g. modal *shall* vs. *will*). The test tells us whether the values a and $\neg a$ behave differently from each other with respect to B.

If these columns divide the dataset into independent samples, such as spoken vs. written, different sub-corpora, or represent two independent runs of the same experiment under different conditions, then you should ideally use a Newcombe-Wilson test. However, the difference in performance between the Newcombe-Wilson test and the standard chi-square is small.

3.2. The Problem of Linguistic Choice

The correct application of these tests relies on the assumption that speakers or writers are free to choose either construction, such as *shall* or *will*, for every case sampled. We have seen that tests can be employed, with modifications, for cases where the independent variable is not free to vary – speakers assigned to one category or another cannot "choose" to contribute to the other data set. However all tests assume that the dependent variable is free to vary, i.e. p can range from 0 to 1.

Note that we specifically limited our example of *shall* and *will* to first person singular cases, where contemporary alternation in English is possi-

ble. Speakers can, with a few formulaic exceptions, freely choose to say *I shall go to the pictures* or *I will go to the pictures* without violence to the semantic context (cf. Lavendera, 1978).

However, this requirement of the test is not always upheld. This problem is quite prevalent in corpus linguistics where it is not possible to experimentally condition choices in advance. In a highly influential paradigm, corpus linguists have frequently cited rates per million words, and used log-likelihood tests (Rayson, 2003)[7] to evaluate whether change in such rates are significant.

It should be apparent from the foregoing discussion that this type of approach undermines the mathematics of the test – whether carried out using a χ^2, log-likelihood or other formula. It is not plausible to allow that every word in a sentence could be *shall* or *will*, and so these other cases should be excluded from the experiment.

Wallis (2012a) demonstrates that the introduction of invariant terms (cases that can be neither *shall* nor *will*) first, introduces unwanted noise into the experiment (although these words may not replace *shall* or *will* they may change in frequency), and second, causes the test to overestimate confidence intervals and lose power. The conclusion, quite simply, is to eliminate these invariant terms as far as possible and try to focus on alternates.[8]

3.3. Case Interaction in Corpora

One of the assumptions we made was that the sample was randomly drawn from a much larger population. However, in corpus linguistics in particular it is common to derive samples from a database consisting of substantive texts. Participants may produce the sampled choice of constructions (*shall* or *will* in our example) multiple times in the same text. If we maximise the use of our data, and include every case found, then we are not strictly engaging in random sampling – even if the texts themselves are effectively randomly obtained. There may be numerous reasons why a speaker may prefer one construction over another, tend to reuse a construction, prime others' uses, or, in writing at least, editorially avoid repetition.

[7]See also http://ucrel.lancs.ac.uk/llwizard.html

[8]A slightly better option for the Rayson school would be to employ goodness of fit tests to test the distribution of the word against the distribution of numbers of words, but in fact, as this website notes, they employ 2×2 homogeneity tests. Again, the validity of conclusions drawn depends on the experimental design and test employed.

This issue affects all standard statistical tests, not just contingency tests. The principal effect is to overestimate the significance of results, because we have assumed in our mathematical model that all n cases are independent from each other. If all cases were dependent on the first then n should be 1! So one correction that could be applied is to try to estimate the degree to which case interaction applies in the data, and then reduce n accordingly. Addressing this problem is beyond the scope of this paper, but the author has published some guidelines on the web.[9]

The key point is to exercise due diligence over your data, and check for clustering and priming effects. If a small number of texts are the source for a large proportion of the data, there is a bigger problem than if data is evenly distributed. As a rule of thumb, if you were to divide your frequencies by 2 and still obtain a significant result, you are likely to be in the clear.

3.4. Analysing Larger Tables

In some experiments we may start with a larger table than 2 × 2: a multi-valued $r \times c$ contingency table (see, e.g., Nelson et al., 2002, p. 276). In such cases you can carry out an initial $r \times c$ χ^2 test to determine if variation is being observed at all. However, a general $r \times c$ test merely tells us that there appears to be significant variation somewhere! The data may be more revealing if it is analysed more closely. It is necessary to "drill down" using more focused tests.

These tests have problems with small expected cell values (usually $e_{ij} < 5$). In this case either rows or columns should be added together and the test performed again.

Let us assume that the independent variable $A = \{a_1, a_2, a_3\}$ is distributed across the columns as before. The dependent variable B is distributed across the rows. Table 3(a)–(c) illustrates a simple 3×3 χ^2 test for homogeneity. Table 3(a) shows an observed distribution $O = \{o_{ij}\}$ with row and column totals r_j and c_i respectively, and grand total N. Table 3(b) then has the expected distribution $E = \{e_{ij}\}$ obtained by the χ^2 homogeneity formula $e_{ij} = (r_j \times c_i)/N$. Finally, Table 3(c) contains a table of χ^2 partial values calculated by the formula $\chi^2(i, j) = (o_{ij} - e_{ij})^2/e_{ij}$.

[9]See also http://corplingstats.wordpress.com/2012/04/15/case-interaction

Table 3. (a)–(c) Analysing a 3 × 3 contingency table.

(a) observed O	a_1	a_2	a_3	$\sum$
b_1	20	5	30	55
b_2	10	10	25	45
b_3	20	10	10	50
$\sum$	50	25	65	140

(b) expected E	a_1	a_2	a_3
b_1	19.64	9.82	25.54
b_2	16.07	8.04	20.89
b_3	14.29	7.14	18.57
$\sum$			

(c) χ^2	a_1	a_2	a_3	Σ
b_1	0.01	2.37	0.78	3.15
b_2	2.29	0.48	0.81	3.58
b_3	2.29	1.14	3.96	7.38
$\sum$	4.59	3.99	5.54	14.12

An important property of χ^2 is that it is *additive*, that is the total value of χ^2 is the sum of its component parts $\chi^2(i, j)$, all of which are positive.[10] The final column and row in Table 3(c) display the "χ^2 contribution" for each row and column respectively. The bottom right cell contains the overall sum. This is significant at four degrees of freedom ($\chi^2_\alpha = 9.488$ for $\alpha = 0.05$).

The first steps are simply to look at some of the totals used to calculate the $r \times c$ chi-square (Table 3(c), bottom right).

(1) Where does A vary? Examine the χ^2 contribution (the partial sum used in the calculation), $\chi^2(i)$, for each column in the $r \times c$ table. As the χ^2 test is additive this tells us which values of the *independent* variable are contributing the greatest amount to the overall result. Columns with smaller values (e.g. a_2) will be more similar to the overall distribution across the column.

[10]Sheskin (1997, p. 240) offers an alternative method involving comparing standardized residuals $R_{ij} = (o_{ij} - e_{ij})/e_{ij}$. This is essentially the same idea, but it reveals the direction (sign) of the variation from the expected value.

These χ^2 contribution values are equivalent to $r \times 1$ goodness of fit tests for each column. All we need do is compare $\chi^2(i)$ with the critical value of χ^2 for $r - 1$ degrees of freedom (5.991 at $\alpha = 0.05$). As we have seen this tells us whether any particular value of the dependent variable is distinct from the overall distribution.[11] In this case no single column distribution ($A = a_i$) over B is significantly different from the overall distribution.

You can also compute percentage *swing* and Cramér's ϕ values for the size of the effect for each column, and compare column swings. See the following section.

(2) Where does B impact on A? Examine the χ^2 contribution for each row in the $r \times c$ table. This gives us an indication as to the values of the *dependent* variable contributing to the overall result. We can see that b_3 is having a greater impact on the result than any other value of B.

Any large contingency table may be simplified by collapsing columns and rows to obtain one of a large set of possible 2×2 tables. The question is then how to proceed. There are two approaches:

(a) Compare every cell against the others to produce the "x against the world" 2×2 χ^2. Cells are reclassified as to whether they are in the same row or column as a given cell, obtaining a 2×2 table $\{\{o_{ij}, r_j - o_{ij}\}, \{c_i - o_{ij}, N - c_i - r_j + o_{ij}\}\}$. Thus to examine the upper left cell o_{11} we would collapse rows 2 and 3, and columns 2 and 3, to obtain the array {{20, 35}, {30, 55}}. The problem is that there will be $r \times c$ cells each requiring a separate 2×2 χ^2 test. While this is possible to compute, it seems like overkill.

(b) A more selective and meaningful approach would likely be premised on linguistic theoretical assumptions. The idea is to simplify the experiment on the basis of a linguistic argument (e.g. that sub-types of transitivity are more meaningfully related than non-transitive and copular types). You might group say, b_1 and b_2 together, and then

[11] You can also collapse the other column values and carry out an $r \times 2$ chi-square where the other column represents the remaining values of the dependent variable. However this is less powerful than the g.o.f. test and would obtain a significant result if the rest of the table was varying but the current column did not.

compare b_3 with the group, so B becomes hierarchically structured: $B = \{\{b_1, b_2\}, b_3\}$. The experiment is thereby divided into two or more sub-experiments at different levels of granularity (e.g. all types of transitivity vs. other classes of verb, followed by sub-types of transitivity).

With care there is no need to dispense with your original result. Once you have obtained χ^2 contribution values for rows and columns it should be possible to identify which sub-values are behaving differently than others. Thus it is straightforward to see that b_3 is contributing 7.38 toward the overall χ^2 sum (this g.o.f. test is significant with 2 degrees of freedom at $\alpha = 0.05$).

4. COMPARING THE RESULTS OF EXPERIMENTS

A frequent question asked by researchers is whether they can argue that the result of one experiment is in some sense "better" or "stronger" than another. It is a common fallacy to state that if one χ^2 test is significant at $\alpha = 0.01$ and another at $\alpha = 0.05$ that the first is a "better" result than the second. There are three objections to this line of argument:

(1) Standard deviation, and thus the threshold for a difference to be significant, falls with increased sample size. Different experiments will typically be based on different sample sizes. Quoting χ^2 and z values is problematic for the same reason: measures must be scaled by $1/\sqrt{N}$. Tests such as χ^2 or z do two things: estimate the size of effect (ϕ, d) and test this value against a threshold (a critical value or confidence interval width).

(2) Reducing the chance of one type of error increases the chance of another. As the error level α falls we increase our confidence that, were the experiment to be repeated many times, we would reach the same conclusion. In other words our results are more *robust*. But it also means that we will tend to be *conservative*, and prematurely eliminate promising results (so-called Type II errors). Researchers should select different α values to control this trade-off, and not conflate this question with the size of the result.

(3) Correlations are not causes. Numerical assessment is secondary to experimental design. As we saw in Section 2, a test is crucially premised on how data is sampled, which may vary for different datasets or corpora.[12] This is another way of pointing out that statistical tests cannot distinguish between correlation and causality. A "better" experiment is one that is framed sufficiently precisely to eliminate alternate hypotheses. Accurately identifying linguistic events and restricting the experiment to genuine choices (semantic alternates, Section 3.2) is more important than the error level reached.

We have seen how a significant result for a 2×2 χ^2 test means that the absolute difference between distributions O_1 and O_2 exceeds the confidence interval, i.e. $| p_1 - p_2 | > z_{\alpha/2}\, s'$. Saying that two χ^2 results are individually *significant* does not imply that they are jointly *separable*, i.e. that one result is significantly greater than the other. However, this question may be correctly determined using a test based on the methods we have already discussed. We end this section by introducing the topic of separability tests.

4.1. Measuring Swing on a Single Dependent Value

An easy way of expressing the impact of an independent variable on a dependent variable in a 2×2 table involves considering how the probability of selecting a particular dependent value $B = b$ changes over the independent variable A. This approach is compatible with goodness of fit tests.

If we return to Table 1 we can see that the probability of selecting b given a (spoken) is $p(b \mid a) = 20/30$ (0.667), whereas the probability of selecting b given $\neg a$ (writing), $p(b \mid \neg a)$, is 5/15 (0.333).

The difference between these p values, $d = p_2 - p_1$, represents the absolute change in probability of selecting b between values of a:

$$swing\ d = p(b|\neg a) - p(b|a) = -0.333. \tag{6}$$

[12]Thus, in claiming that a change detected between 1960s and 1990s data taken from the Lancaster-Oslo-Bergen corpus set {LOB, FLOB} is attributable to a change in *time*, one must eliminate all other variables. We tend to assume that corpora are sampled equally randomly for all other variables, such as *level of formality*, but this is seldom possible to guarantee in practice.

We can also divide this difference by the starting point to obtain a swing relative to the starting point, $p(b \mid a)$:

$$\textit{percentage swing } d^{\%} = p(b|\neg a) - p(b|a)/p(b|a) = -50\%. \tag{7}$$

The advantage of citing percentage swing is that it minimizes the impact of normalization. It is possible to use any normalised frequency to perform this calculation, e.g. per million words, when the absolute swing will tend to be very small. Percentage change is in common usage, to the extent that it is rarely expressed formulaically.

Thus far we have quoted simple frequencies. We can use the z test for two independent proportions (Section 2.4) or Newcombe-Wilson interval (Section 2.5) to compute a confidence interval on d (and, by dividing by $p(a \mid b)$, on $d^{\%}$). Assuming that samples are drawn from a within subjects design, $\hat{p} = 25/45$ and $s' = 0.157$. The Gaussian interval for d at 0.05 level, $z_{\alpha/2}\, s' = 0.308$, thus: *swing* $d = -0.333 \pm 0.308$.

Alternatively, as a percentage of $p(a \mid b)$, *percentage swing* $d^{\%} = -50\% \pm 46.20\%$. The confidence interval will not span zero if the test is significant (cf. Figure 7).

4.2. Measuring Effect Size Over All Dependent Values

There are two main problems with swing-based measures. The first is numerical: percentage swing is unequal about zero – i.e. a percentage *fall* of 50% is not the same as a percentage *rise* of 50%, and any change from zero is infinite! Secondly, measures of change should be constrained to the same range, but $d^{\%}$ can range over any value. Ideally, measures of effect size should share a probabilistic range ([0, 1], or potentially, if sign is important, $[-1, 1]$).[13]

A second set of methods use the χ^2 calculation to measure the size of an effect. The optimum standard measure is called Cramér's phi or ϕ. This may be calculated very simply as

$$\phi \equiv \sqrt{\frac{\chi^2}{N \times (k-1)}} \tag{8}$$

[13]A signed ϕ for 2×2 tables can be calculated with an alternative formula (Sheskin, 1997, p. 244).

Table 4. ϕ measures the degree by which a flat matrix F is perturbed towards the identity matrix I.

$\phi = 0$	F	a	$\neg a$	$\phi = p$	Φ	a	$\neg a$	$\phi = 1$	I	a	$\neg a$
	b	$^1/_2$	$^1/_2$		b	$^{(p+1)}/_2$	$^{(1-p)}/_2$		b	1	0
	$\neg b$	$^1/_2$	$^1/_2$		$\neg b$	$^{(1-p)}/_2$	$^{(p+1)}/_2$		$\neg b$	0	1

where k is the smaller of the number of rows and columns, i.e. $k = \min(r, c)$. For a 2×2 table, $k = 1$ and the formula simplifies to the root of χ^2 over the total n. For Table 1, $\phi = 0.32$.

Cramér's ϕ has two important properties which makes it superior to other competing measures of association. First, ϕ is probabilistic i.e. $\phi \in [0, 1]$. Second (and this is a point rarely noted), ϕ measures *the level of perturbation* from a flat matrix F to the identity matrix I. For any intermediate matrix for a point, $p \in [0, 1]$ between F and I, $\phi = p$ (Table 4). This is conceptually appealing and similar to the idea of information flow from A to B (to what degree does the value of A determine the value of B?).

It is possible to calculate confidence intervals on ϕ but this is outside the scope of this current paper. Readers are referred to Wallis (2011) for more information.

4.3. Using ϕ to Measure Effect Size on a Single Dependent Value

The formula above measures perturbation for a 2×2 (or $r \times c$) χ^2 test of homogeneity.

It is also possible to calculate an equivalent ϕ for a goodness of fit test. However the g.o.f. χ^2 cannot just be substituted into the formula above. In order to limit ϕ to a probabilistic range (and render results easily comparable) we need to modify it.

Wallis (2012b) considers a wide range of alternative approaches to measuring the size of goodness of fit before concluding that a simple formula, ϕ_p, is the most reliable. First, we calculate a revised chi-square calculation multiplying each contribution by its prior probability $p(b_i) = e_i/n$.

$$\chi^2_p = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \times p(b_i) = \frac{\sum (o_i - e_i)^2}{n} \tag{9}$$

and

$$\phi_p = \sqrt{\chi^2_p / 2n}. \tag{10}$$

Second, this probabilistically-weighted ${\chi^2}_p$ cannot exceed a limit of $2n$, so we define $\phi_p \in [0, 1]$ accordingly. This formula is simply the standardised *root mean square* (r.m.s.) of differences.

4.4. Testing Swings for Statistical Separability

In some circumstances it can be valuable to compare results to decide whether or not one change is greater than the other. Consider Table 5, which represents the results of a second run of the same experiment in Table 1. The swing is clearly a larger value numerically. We might ask whether the swing observed here is significantly greater than that seen in the first experiment.

We will denote the previously seen swing for Table 1 as $d_1(a)$ and the new swing Table 4 as $d_2(a)$. We obtain $d_1(a) = -0.333 \pm 0.308$ and $d_2(a) = -0.833 \pm 0.245$ (Figure 7). Both results are independently statistically significant, that is they express a non-zero change over values.

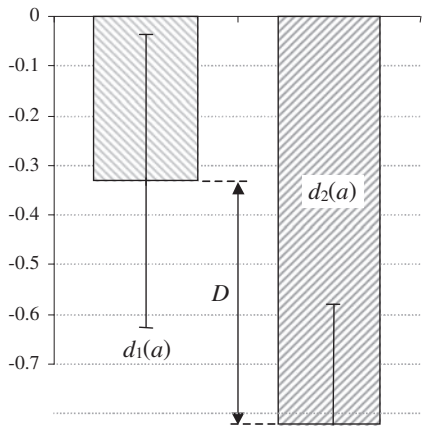

Fig. 7. Comparing two experiments by contrasting each swing.

Table 5. The results of a second "experiment".

	a	$\neg a$	Σ
b	50	0	50
$\neg b$	10	20	30
Σ	60	20	80

In absolute terms, the second result is greater than the first. The difference between $d_1(a)$ and $d_2(a)$, $D = -0.5$, which is large. However, this does not mean that the second swing is sufficiently greater than the first such that *their difference D* would be considered significant. Imagine if we repeated both experiments many times: *Will the swing for the second experiment exceed that of the first, 95% or more of the time?*

To test this hypothesis, Wallis (2011) uses an independent-population z-test for 2 independent samples (see Section 2.5) to compare $|d_1(a) - d_2(a)| > z_{\alpha/2} \cdot s'$. We calculate the standard deviation using this formula to reflect the fact that the experiments are conducted on different datasets.

We use the sum of variances rule $s' = \sqrt{s_1^2 + s_2^2}$. A shortcut exploits the equality $e \equiv z_{\alpha/2} \cdot s$, so we can reuse confidence intervals $e' = \sqrt{e_1^2 + e_2^2}$ to calculate the new interval. We obtain $e' = 0.397$, which is less than D. The experimental runs are therefore significantly different or statistically separable.

This method can be used to compare the results of different experiments (as here) or different swings observed within a multi-valued experiment. Wallis (2011) provides separability tests for repeated runs of all the tests we have discussed, i.e. goodness of fit, 2×2 homogeneity and independent population tests, and multinomial $r \times 1$ and $r \times c$ tests.

This type of test has the same number of degrees of freedom as each individual test. If individual tests have one degree of freedom (and hence interpretation), so does the separability test.[14] Therefore, not only can we report that the two swings significantly differ, we can also say that *the tests* differ in their result. 2×2 signed ϕ values (see Section 4.2) must also significantly differ. The effect of A on B is significantly greater on the second experimental run.[15]

5. CONCLUSIONS

Inferential statistics is a field of mathematics that models the likely outcome of repeated runs of the same experiment, based on an underpinning mathematical model, from the observation of a single experiment. For types of

[14]If the swings significantly differ for one column, say a, then the same is true for $\neg a$. The differences are equal but inverted, i.e. $d_1(a) - d_2(a) = d_2(\neg a) - d_1(\neg a)$, with the same confidence interval.

[15]A spreadsheet for separability tests is available at www.ucl.ac.uk/english-usage/statspapers/2x2-x2-separability.xls

linguistic experiments where the dependent variable represents a single Boolean alternative, the Binomial model is the appropriate mathematical model for carrying out this type of prediction. A related model, termed the multinomial model, applies for more than two outcomes. This paper concentrates on the simplest versions of tests because only one degree of freedom implies only one potential conclusion.

Since the Binomial distribution is arduous to compute from first principles, it is common to employ a Normal approximation to the Binomial, and on this basis, carry out χ^2 tests and compute confidence intervals. Although this is an approximation, the errors introduced in this step are small, and may be compensated for by employing Yates' continuity correction.

A rather more serious and common error is the use of this approximation in 'Wald' confidence intervals, which leads to the perplexing situation that we may observe p at zero, but the confidence interval for p appears to allow values to be less than zero! Since confidence intervals increase in size with small n, this problem also affects less skewed observations supported by small amounts of data. The correct approach is to use Wilson's score interval in place of the Wald interval. This assumes a Poisson skewed distribution which can never exceed the range [0, 1].

We discussed the difference between three experimental designs: the goodness of fit test, the test for independence (also known as the homogeneity test) and the independent samples test. The first compares an observed distribution against a specific given distribution; the second and third compare two sample distributions against each other, the difference between them lying in whether we assume that the samples are drawn from the same underlying population of participants (sometimes termed the within-subject/between-subject distinction).

Section 3 concerns the correct application of these tests. We discussed when different tests should be used and note common problems, particularly prevalent in *ex post facto* corpus linguistics research, of attempting to employ contingency tests on data where dependent variables are not free to vary and where cases are not independently sampled. These problems are not easily addressed by modifying the statistical test, but may be minimised by refining the experimental design and taking care when sampling potentially interacting cases.

Methods for analysing larger tables were also covered in Section 3. The purpose of re-analysing larger tables is to examine experimental data drawing linguistically meaningful distinctions and simplifying tables to do so. Here the emphasis on simplicity of test can pay off.

Finally we turned to the question of the strength of different results. We introduced some basic effect size measures which are applicable to a variety of conditions, including goodness of fit tests.

We noted that the common practice of citing χ^2 scores or error levels is highly misleading and should be avoided. As an alternative, we introduced and demonstrated a separability test, a statistical test for evaluating whether the size or pattern of effect found in one set of results is significantly different from that found in another. This type of test is based on the same underlying mathematics as the simpler χ^2 tests, but allows researchers to make statistically sound claims about multiple runs of the same test.

ACKNOWLEDGEMENTS

This paper is the result of carrying out numerous corpus studies and discussions with linguists over several years, and was initially self-published on the author's website and blog, corp.ling.stats (http://corplingstats.wordpress.com), generating a fair amount of discussion. Special thanks go to Bas Aarts, Jill Bowie, Joanne Close, Gunther Kaltenböck, Geoff Leech and Seth Mehl, and an anonymous reviewer for *JQL* whose comments helped refocus the paper for publication.

REFERENCES

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*, 61–74.

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

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine, 17*, 873–890.

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison.* PhD thesis, Lancaster University.

Sheskin, D. J. (1997). *Handbook of Parametric and Nonparametric Statistical Procedures*, 2nd Edition. Boca Raton, FL: CRC Press.

Singleton, R. Jr., Straits, B. C., Straits, M. M., & McAllister, R. J. (1988). *Approaches to Social Research*. New York/Oxford: OUP.

Wallis, S. A. (2012a). That vexed problem of choice. London: Survey of English Usage, UCL. Retrieved 14 August 2013, from www.ucl.ac.uk/english-usage/statspapers/vexed-choice.pdf

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

Wallis, S. A. (2011). Comparing χ^2 tests. London: Survey of English Usage, UCL. Retrieved 14 August 2013, from www.ucl.ac.uk/english-usage/statspapers/comparing-x2-tests.pdf

Wallis, S. A. (2012b). Goodness of fit measures for discrete categorical data. London: Survey of English Usage, UCL. Retrieved 14 August 2013, from www.ucl.ac.uk/english-usage/statspapers/gofmeasures.pdf

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

SPREADSHEETS

2 × 2 chi-squares and associated tests: www.ucl.ac.uk/english-usage/statspapers/2x2chisq.xls

Separability tests for paired tables: www.ucl.ac.uk/english-usage/statspapers/2x2-x2-separability.xls