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 *re y*. Stating that a result is significant at a certain level of error (0.01, for example) is another way of stating that, were we to repeat the experiment many times, the likelihood of obtaining a result other than that reported will be below this error level.

The origin of

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". Generalisations from a corpus of English speech and writing, such as ICE-GB (Nelson *e* 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 rationalising beyond this population to a wider population is possible – by arguing why this 'operationalising' population is, in the respect under consideration, *repre en e* of this wider population – but it is not given by the statistical method.

ndo ne nd ndependence

The first assumption we need to make is that the sample is *r ndo p e* 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 *n* we need to be aware of this and consider the effect of clustering on their independence (see Nelson *e* 2002: 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.

e p e ery re e o e pop on

The second assumption is that the population is muc

will take *A* as our *ndependen r e*, and *B* as our *dependen r e*. This means that we try to see if *A*

to our current experiment. We can now use the Normal 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

z- *q red e or g n nd pp c on of* χ^2

e *goodne* *of f* *e*

An alternative calculation employs the ^{2} formula:

$$^2 = \sum \frac{(o \quad e)^2}{e} , \qquad (2)$$

where the observed distribution o = {20, 10} and expected distribution is based on t

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

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

We may obtain a similar interval for the second ‘writing’ column, $p(\ |$

z

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 or , for every case sampled. We have seen that tests can

3.4 Analysing larger tables

In some experiments we may start with a larger table than 2 2: a multi-valued *r* *c* contingency table (see, e.g., Nelson *e*

2) **Where does *B* impact on *A*?** Examine the ^{2} contribution for each row in the r c table. This gives us an indication as to the values of the *dependen* variable contributing to the overall result. We can see that $_3$ 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 ^{2}**. Cells are reclassified as to whether they are in the same row or column as a given cell, obtaining a 2 2 table $\{\{o, r - o\}, \{c - o, N - c - r + o\}\}$. 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 $r \times c$

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 ^{2} test means that the absolute difference between distributions $\mathbf{O}_1$ and $\mathbf{O}_2$ exceeds the confidence interval, i.e. $|p_1 - p_2| > z_{/2}$ '. Saying that two ^{2} results are individually *gn f c n* does not imply that they are jointly *ep r e*, 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

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 . This may be calculated very simply as

$$\sqrt{\frac{\chi^2}{N\ (\ \ 1)}} \tag{8}$$

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

Cramér's has two important properties which makes it superior to other competing measures of association. First, is probabilistic, i.e. [0, 1]. Second (and this is a point rarely noted), measures *e e e of per r on* from a flat matrix **F** to the identity matrix **I**. For any intermediate matrix for a point, p [0, 1] between **F** and **I**, = p (Figure 6). 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?).

We'll denote the previously seen swing for Table 1 as $d_1(\ )$ and the

WALLIS, S.A. 2012. *oodne off e re for d cre e c egor c d* . London: Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/gofmeasures.pdf
WALLIS, S.A. 2013. Binomial confidence intervals and contingency tests. *o rn of Q n e L ng c* **20**:3: 178-208
WALLIS, S.A. forthcoming. *e ed pro e of c o ce*. London: Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/vexedchoice.pdf
WILSON, E. B. 1927. Probable inference, the law of succession, and statistical inference. *o rn of e A er c n c A oc on* **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