oodne o e e o d c e e c e o c d

Sean Wallis, Survey of English Usage, University College London
ePublished: January 30 2012

n od c on

A goodness of fit ^{2} test evaluates the degree to which an observed discrete distribution over one dimension differs from another. A typical application of this test is to consider whether a specialisation of a set, i.e. a subset, differs in its distribution from a starting point (Wallis forthcoming). Like the chi-square test for homogeneity (2 × 2 or generalised row ***r*** × column ***c*** test), the null hypothesis is that the observed distribution matches the expected distribution. The expected distribution is proportional to a given prior distribution we will term **D**, and the observed **O** distribution is typically a subset of **D**.

A measure of association, or correlation, between two distributions is a score that measures the degree of difference between the two distributions. Significance tests might compare this size of effect with a confidence interval to determine that the result was unlikely to occur by chance.

Common measures of the size of effect for two-celled goodness of fit ^{2} tests include simple difference (swing) and proportional difference ('percentage swing'). Simple swing can be defined as the difference in proportions:

$$\boldsymbol{d} = \frac{\mathbf{O}_1}{\mathbf{D}_1} - \frac{\mathbf{O}_0}{\mathbf{D}_0}. \qquad (1)$$

For 2 × 1 tests, simple swings can be compared to test for significant difference between pairs of test results. Provided that **O** is a subset of **D** then these are real fractions and ***d*** is constrained $\boldsymbol{d}$ [-1, 1]. However, for ***r*** × 1 tests, where ***r*** > 2, we will necessarily obtain an aggregate estimate of the size of effect. Secondly, simple swing cannot be used meaningfully where **O** is not a subset of **D**. In this paper we will consider a number of different methods to address this problem.

Correlation scores are a sample statistic. The fact that one is numerically larger than the other does not mean that the result is ***significantly greater***. To determine this we need to either

1. estimate confidence intervals around each measure and employ a ***z*** test for two proportions from independent populations to compare these intervals, or

2. perform an ***r*** × 1 separability test for two independent populations (Wallis 2011) to compare the distributions of differences of differences.

In cases where both tests have one degree of freedom, these procedures obtain the same result. With ***r*** > 2 however, there will be more than one way to obtain the same score. The distributions can have a significantly different pattern even when scores are identical.

1.1 A simple example: correlating the present perfect

Bowie, Wallis and Aarts (2013) discuss the **present perfect** construction. The present perfect expresses a pa.1525(s)-1.2312(c1)-fedcL.295585(r)-192(d)-0.26558( )-0.14779buti5585(e)3.74( )-0.144974p99

present	**LLC**	**ICE-GB**	Total	present perfect goodness of fit
present non-perfect	33,131	32,114	65,245	$d^{\%}$ = -4.45 ± 5.13%
present perfect	2,696	2,488	5,184	ϕ' = 0.0227
TOTAL	35,827	34,602	70,429	χ^2 = 2.68 ns
past				
other TPM VPs	18,201	14,293	32,494	$d^{\%}$ = +14.92 ± 5.47%
present perfect	2,696	2,488	5,184	ϕ' = 0.0694
TOTAL	20,897	16,781	37,678	χ^2 = **25.06 s**

Table 1. Comparing present perfect cases against (upper) tensed, present-marked VPs, (lower) tensed, past-marked VPs (after Bowie ***et al.*** 2013).

Bowie ***et al.*** limit their discussion to two 400,000 word text categories in the DCPSE corpus, divided by time, namely LLC (1960s) and ICE-GB (1990s) texts. Table 1 shows their analysis, employing percentage swing $d^{\%}$ and Wallis ϕ' (section 3). They found that the present perfect more closely associated with present tensed VPs. Note that in employing measures for this purpose, a higher value of χ^2, ϕ' or $d^{\%}$ implies a ***weaker correlation*** between the present perfect and the particular baseline being tested against it.

However with only two categories of text, this can only be a coarse-grained assessment. To test the hypothesis that the present perfect is more likely in **texts** with a greater preponderance of present-referring VPs than past-referring ones, we need to find a way to extend our evaluation to smaller units than 0.4M-word subcorpora, ideally to the level of individual texts.

Before we do this it seems sensible to consider a middle position. DCPSE is subdivided sociolinguistically into different **text genres** of different sizes. Figure 1 plots the observed distribution **O** and the distributions for the present referring and past referring VPs scaled by **O**, across these 10 text categories.

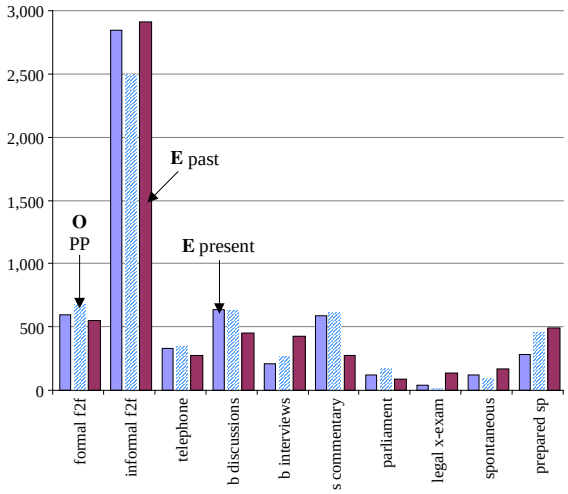

Figure 1. The distribution of the present perfect **O**, scaled distributions **E** for present and past, across text categories of DCPSE.

'Eyeballing' this data seems to suggest a close congruence between the distribution of the present perfect and the present in some categories (e.g. broadcast discussions, spontaneous commentary) and a closer relationship with the past in others (prepared speech). It appears intuitively that there is a closer relationship between present perfect and the present, ***but how might this be measured?***

Any measure of correlation between pairs of distributions needs to scale appropriately to permit populous categories, such as informal face-to-face conversation, and less populous ones, such as legal cross-examination, to add evidence to the metric appropriately.

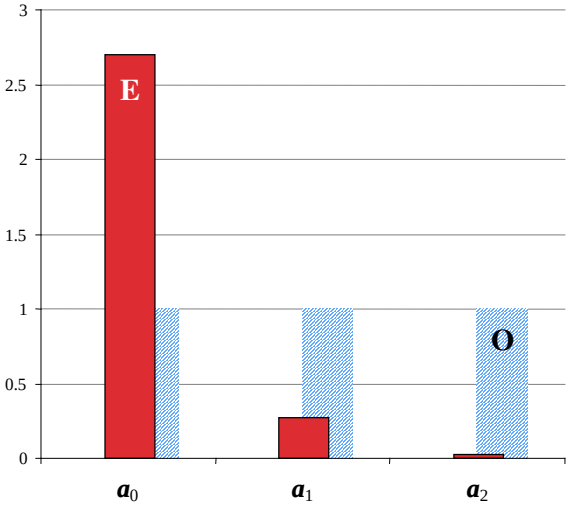

Figure 2. A test expected distribution **E** and an example observed distribution **O**.

C

For $r \times c$ tests of homogeneity (independence) a standard method employs Cramér's ϕ:

$$\phi \equiv \sqrt{\frac{\chi^2}{N \times (k-1)}} \tag{3}$$

where N is the total number of observed cases, k is the length of the diagonal, i.e. min(r, c) for a matrix of r rows and c columns. We guarantee that ϕ is constrained to the probability space [0, 1], where 0 represents an exact match and 1 a complete perturbation (Wallis 2012).

The corollary of (3) is that the maximum value of an $r \times c$ χ^2 computation can be identified as

$$\text{limit}(\chi^2) = N \times (k-1). \tag{4}$$

This formula may even be generalised to three dimensional chi-square tests, provided the limit is multiplied by 3. Indeed it can be shown that ϕ measures the linear perturbation from a flat matrix towards a diagonal. This is obtained irrespective of whether the expected distribution is skewed, and the maximum is achievable irrespective of prior distribution.

This formula cannot be applied as-is to a goodness of fit test, however, without hitting a major obstacle ***due to the distinction between the two tests***. Whereas the expected distribution in a test of homogeneity is determined exclusively from observed totals, employing the product of the row and column probabilities, the expected distribution in a goodness of fit test is **given**, and is independent from the observed distribution. As a result ϕ may exceed 1.

We can demonstrate the problem numerically. Suppose we calculate χ^2 in the normal manner (cf. Table 2b). The maximum value of the goodness of fit χ^2 is obtained when the observed distribution

This formula has one further advantage. So far we have assumed that **O** is a true subset of **D**, so that **O**

obtaining

$$\mathbf{SS_{tot}} = \sum_{i=0}^{k-1} \frac{(\mathbf{O}_i - \bar{\mathbf{O}})^2}{\mathbf{i}^2}, \text{ and} \qquad \mathbf{SS_{err}} = \sum_{i=0}^{k-1} \frac{(\mathbf{O}_i - \mathbf{E}_i)^2}{\mathbf{i}^2}. \tag{15}$$

where $\bar{\mathbf{O}}$ represents the mean observation (either by simple division or by probabilistic summation) and

we find that the dependent probability dp_R actually **falls** numerically in one of the patterns {0, 1, 2}, whereas ϕ-based measures increase. This leads us to discount dp_R. For a meta-comparison we order measures according to their similarity of performance. We find that among the ϕ-based measures we find that ϕ_E and ϕ_v is most similar to ϕ, and ϕ_p is closer to ϕ' (and relative dependence dp_R). Figure 10 shows this pattern in a striking manner.

We are left with four formulae based on ^{2} that behave in a reliable manner. Due to the scaling problem, Cramér's ϕ is not robustly applicable to different expected distributions, and can be replaced with ϕ_E. It is not clear what the Gaussian variance ϕ_v gains over ϕ_E, so ϕ_v can be eliminated. The most interesting cases are ϕ_p and ϕ_E, which are both robust fitness measures, ϕ_p is the most general and can be applied to partially-overlapping subsets, however we may prefer ϕ_E for true subsets because it appears to behave most like Cramér's ϕ.

For overlapping sets we can employ ϕ_p. Note that ϕ_p is the probabilistic sum of ^{2} partials, or, to put it another way, it is proportional to the absolute

We may summarise our initial observations on the basis of these results as follows.

- Probability-weighted ϕ_p factors out variance and has the smallest ratio between baselines in Table 8, indicating that present and past are distinguished the least from the perfect. However, this measure appears to be the most consistent across different scales.

- Variance-weighted ϕ_v ($\approx \phi_E$) seems to be less affected by noise, which we would expect, as each difference square is scaled by its variance. However this is at the cost of a tendency for ϕ_v to fall as the number of categories, k, increases. Table 9 has a large number of different categories (280 texts, 460 non-empty subtexts in the case of present tensed VPs).

- There is a relationship between Cramér's ϕ (first column) and ϕ_E (last column). ϕ_E is constrained to the range [0, 1] by scaling each to their limit (involving the sum of $^1/_E$ terms). If this limit is different for present and past cases, then the ratios for ϕ and ϕ_E will also differ.

10.5 The effect of category scale

W saw that measures were affected by the number of categories in a simple contrast of texts to the

- **The standard deviation of measures**. The standard deviation of each measure will increase as the number of categories k falls, because there are greater permutations of text to category (this may also be affected by different-sized DCPSE texts). However, considered as a proportion of the mean, the standard deviation of each measure is in the following order: $\sigma(\phi_p) < \sigma(\phi) \approx \sigma(\phi') < \sigma(\phi_E)$, meaning that ϕ_p is least affected by the particular allocation to category.

The result of our evaluation is that on a number of counts probabilistically-weighted ϕ_p (i.e. root mean square error) seems to be superior to other measures. It is the most stable with respect to variation of size and number of categories, and obtains a reliable ratio when comparing two different baselines. It is easily constrained to 1 and is one of the simplest measures to calculate. It also has the smallest standard deviation as a proportion of the measure. Finally it is robustly extensible to comparing non-overlapping sets.

10.6 Estimators for ϕ_p

In this paper we identified that text category impacted on the relationship between present perfect and present and past categories. To demonstrate this we calculated the mean for k = 10 from 10,000 repetitions of $\phi_p(k)$ for random subdivisions of the corpus.

Consider the following problem. Suppose we were to subdivide a corpus into two approximately equal halves and observe the value of ϕ_p for this subdivision. Depending on how the subdivision affects the dependent variable, the observed score will be above or below the expected value. What is the optimum expected value for ϕ_p, the **estimator**, written $\hat{\phi}_p(2)$? In short, how may we algebraically predict the expected value of ϕ_p for any given k from $\phi_p(K)$ where K is the number of texts, subtexts etc. (or some other categorically normative baseline)?

Note that we cannot apply a separability test (Wallis 2011) to compare results because the two experiments (K=280, k=2) have different degrees of freedom.

We need to find this optimum expected value. In this paper we relied on extensive computation to do this. Is there an algebraic solution?

Examining the curves for ϕ_p in Figure 11(a) reveals that the relationship can be closely predicted by the formula $\phi_p(k) + x/k = c$, so it follows that

$$\hat{\phi}_p(k) =$$

Bowie, J., Wallis, S.A. and Aarts, B. 2013. The perfect in spoken English. In Aarts, B., J. Close, G. Leech and S.A. Wallis (eds.). ***The English Verb Phrase: Corpus Methodology and Current Change***. Cambridge: CUP.

Wallis, S.A. forthcoming. z***-squared: The origin and use of*** ^{2}. ***Journal of Quantitative Linguistics***. www.ucl.ac.uk/english-usage/statspapers/z-squared.pdf

Wallis, S.A. 2011. ***Comparing*** ^{2} ***tests for separability***. London: Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/comparing-x2-tests.pdf

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

