Goodness of fit measures for discrete categorical data

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

1. Introduction

A goodness of fit χ^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 $r \times$ 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 χ^2 tests include simple difference (swing) and proportional difference ('percentage swing'). Simple swing can be defined as the difference in proportions:

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

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 $d \in [-1, 1]$. However, for $r \times 1$ 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 $r \times 1$ 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 particular relationship between present and past events and it is not *a priori* determined as to whether we would expect its use more commonly in texts which are more present- or past-referring. We may estimate the degree to which a text refers to the present by counting the frequency of present tensed verb phrases in it (and normalising as appropriate), ditto for the past.

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.

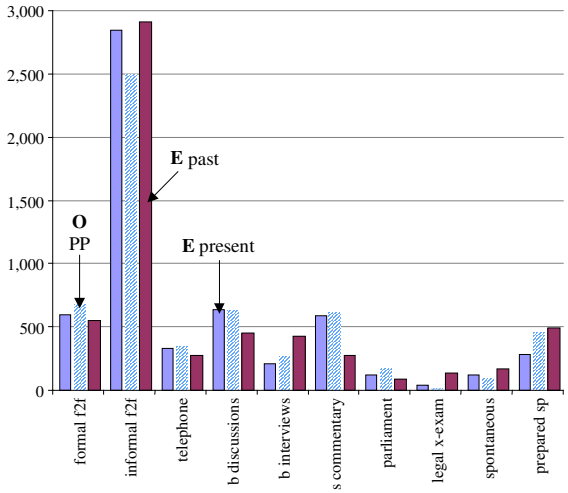


Figure 1. The distribution of the present perfect **O**, scaled distributions **E** for present and past, across text categories of DCPSE.

'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.

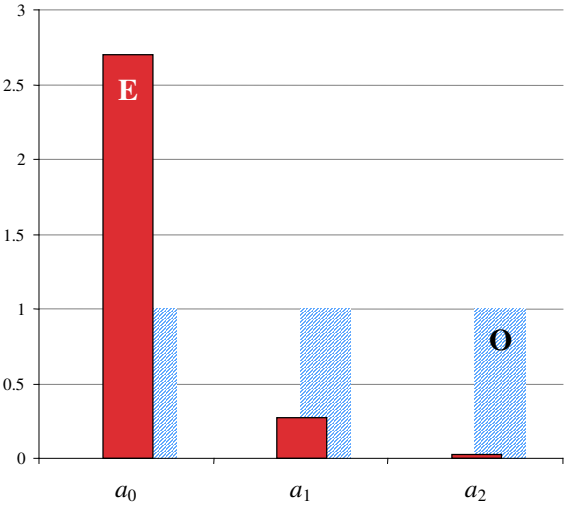


Figure 2. A test expected distribution **E** and an example observed distribution **O**.

1.2 Distributions for evaluation purposes

The present perfect example contains data in categories of radically different size, which a robust measure must accommodate. Where categories are guaranteed to be of the same or similar size we would expect the performance of measures to converge, as this variation is simply one less factor to take into account. We therefore employ a highly skewed idealised distribution for comparative purposes.

Consider the following discrete distribution, expressed over a three-valued variable $A = \{a_0, a_1, a_2\}$ (Table 2a). We employ an expected distribution that is highly skewed, indeed exponential: **D** = {100, 10, 1}. We can also express this as a prior probability distribution *p*.

χ^2 sums the square of differences between observed and expected distributions, scaled by the variance, taken to be the same as the expected distribution **E**. The effect of the skewed expected distribution can be clearly seen in Table 2b.

A	**D**	*p*
a_0	100	0.90
a_1	10	0.09
a_2	1	0.01
TOTAL	**111**	**1**

O	**E**	χ^2
1	2.70	1.07
1	0.27	1.97
1	0.03	35.03
N = **3**	**3.00**	**38.07**

Tables 2a and 2b. An example skewed three-valued prior distribution and sample χ^2 test.

In this paper we consider a wide range of methods utilising this simple test distribution.

2. Reduced χ^2

A common approach, employed in model-fitting, employs the so-called 'reduced chi-square',

$$\chi^2_{\text{red}} = \chi^2/\nu, \tag{2}$$

where ν represents the number of degrees of freedom in the table. The idea in model-fitting is to compute a chi-square against an expected distribution predicted by a function $f(a)$ and attempt to match that function to the observation.

If we substitute our expected distribution, $f(a) = \mathbf{E}(a)$, the number of degrees of freedom $\nu = k - 1 = 2$, where k is the number of categories. (In fitting, one would also subtract the number of parameters of the function.) For our data this obtains a reduced chi-square of 19.035. The interpretation of this result is simply that, as $\chi^2_{\text{red}} > 1$, the data 'does not match' the function, which a glance at Figure 2 reveals! But this conclusion does not allow us to compare the extent to which distributions match, merely to reject **E** as a good fit to our observation **O**.

Second, χ^2 increases in proportion to sample size N. We cannot easily employ this method to compare results with different sample sizes. This is less important when fitting different functions to a single observed distribution, because N is constant throughout and fitting against a model employs an information-theoretic argument (essentially: as N increases the model is permitted to become more complex). However if we wish to compare sizes of effect over samples, then we must scale measures proportionately.

Third, as Andrae *et al* (2010) note, reduced chi-square applies to *linear* models, and its behaviour is unreliable with nonlinear models, such as arbitrary categorical data.

3. Cramér's ϕ

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

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

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

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

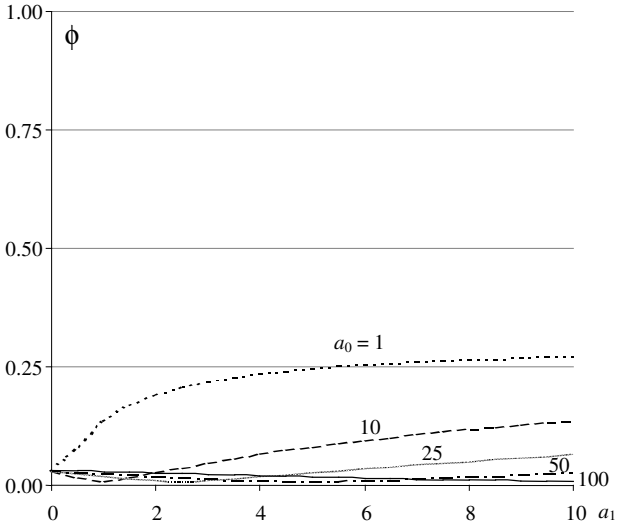


Figure 3. Cramér's ϕ as a_1 varies from 0 to 10, k = 1/min($p(a)$). The functions described in this paper reach a minimum at $a_1 = a_0/10$, corresponding to the expected ratio.

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 falls wholly at the **least** expected value (here a_2), i.e. **O** = {0, 0, N}. Substituting **O** = {0, 0, 1} obtains a χ^2 of 110.00 and ϕ of 7.42. The fact that χ^2 can exceed $N \times (k - 1)$ due to the independence between **O** and **E** means that we cannot limit ϕ to [0, 1].

The maximum value of a goodness of fit χ^2 can be shown to be

$$\text{limit}(\chi^2) = N \times (1/\min(p(a)) - 1), \qquad (5)$$

where min($p(a)$) represents the probability of the least probable element a. In our case min($p(a)$) = 0.0090 (to four decimal places). If we fix on this value as a maximum, with a skewed prior distribution, no amount of deviation from the expected distribution on the other values of A can obtain a ϕ of 1. This method is also sensitive to the minimum prior min($p(a)$). It is therefore difficult to recommend this method for comparing results with different prior distributions **E**.

To illustrate the performance of each function we will explore the effect of varying the middle value, a_1, from 0 to 10 with different lines representing different values of the most common value a_0. The least common value a_2 has zero items. Figure 3 plots ϕ where k = 1/min($p(a)$), i.e. 1/$p(a_2)$. In these circumstances ϕ cannot reach 1, and we can see that the line $a_0 = 1$ is tending to reach an asymptote as a_1 increases. In section 6 we will explore a different approach to constraining ϕ.

4. Normalised ϕ'

Bowie *et al.* used a normalised ϕ measure, ϕ', that reaches 1 where the observed distribution settles on any single value, irrespective of its prior frequency in **E**. Note that equation (5) above converges to (4) if the prior probability is even: $\min(p(a)) = p(a) = 1/k$, i.e. $1/\min(p(a)) = k$.

The method first flattens the expected distribution ($p(a) = 1/k$), and recalibrates the observed distribution in linear proportion. A new χ^2 computation (and therefore ϕ') sums these rescaled differences. We need two new normalised distributions as **E'** and **O'**. We require $\mathbf{E'}_i \equiv N/k$ and rescale the observed distribution **O'** in two steps.

Note that in applying the standardisation to **E** each term $\mathbf{E}_i$ has been adjusted by multiplying by **E'/E**, so first we repeat the scaling transformation for the observed distribution, thus:

$$\mathbf{O''}_i \equiv \mathbf{O}_i \times \mathbf{E'}_i/\mathbf{E}_i, \text{ and then} \qquad \text{(step 1)}$$

$$\mathbf{O'}_i \equiv \mathbf{O''}_i \times N/\Sigma\mathbf{O''}. \qquad \text{(step 2)}$$

We compute ϕ' from a goodness of fit $\chi^2(\mathbf{O'}, \mathbf{E'})$ using equation (3). If we perform the same computation as before, we now find that with $a_2 = 0$, whereas constrained ϕ could not exceed 0.5, this new ϕ' cannot fall below it! This seems rather counter-intuitive, but it is a result of the reweighting of the difference at a_2 in this data. Figure 4 also shows that this function reaches 1 when a_1 is zero. (Since $a_2 = 0$ already, this means that all observations are found at a_0.)

A	**O**	**E**	**E'**	**O''**	**O'**	χ^2
a_0	1	2.70	1	0.37	0.03	0.95
a_1	1	0.27	1	3.7	0.27	0.53
a_2	1	0.03	1	37	2.70	2.90
TOTAL	$N = \mathbf{3}$	**3.00**	**3.00**	**41.07**	**3.00**	**4.38**

Table 3. Recalibrating observed and expected distributions to obtain a normalised ϕ'.

5. Probabilistically-weighted ϕ_p

Suppose we weight χ^2 computations according to the prior $p(a_i)$, i.e.

$$\chi^2_p = \sum_{i=0}^{k-1} \frac{(\mathbf{O}_i - \mathbf{E}_i)^2}{\mathbf{E}_i} \times p(a_i) = \sum \frac{(\mathbf{O}_i - \mathbf{E}_i)^2}{\mathbf{E}_i} \times \frac{\mathbf{E}_i}{N} = \frac{\sum(\mathbf{O}_i - \mathbf{E}_i)^2}{N}. \qquad (6)$$

Probabilistically-weighted χ^2_p cannot exceed a limit of $2N$, and we may define ϕ_p accordingly.[1]

$$\phi_p = \sqrt{\frac{\chi^2_p}{2N}}. \qquad (7)$$

The performance of this function over our test data is illustrated by Figure 5. Weighted ϕ_p performs similarly to Cramér's ϕ, in that it reaches a maximum value on the least frequent value, and indeed Figure 5 obtains a similar pattern to Figure 2. However this function is not sensitive to the particular expected distribution **E**, as it does not rely on the minimum function. At saturation ϕ_p is constrained probabilistically, and will approach but not reach 1, i.e. $\phi_p \in [0, 1)$.

[1] Whereas we use the notation ϕ_p in this paper to identify relationships between ϕ measures, clearly this formula is simply the standardised *root mean square* (r.m.s.) of error terms.

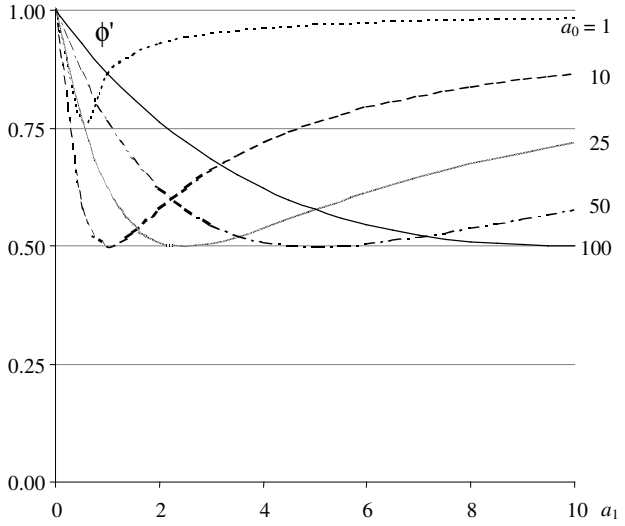


Figure 4. Plotting ϕ' over the same distributions (a_0 = 1, 10, 25 etc.), a_1 from 0 to 10.

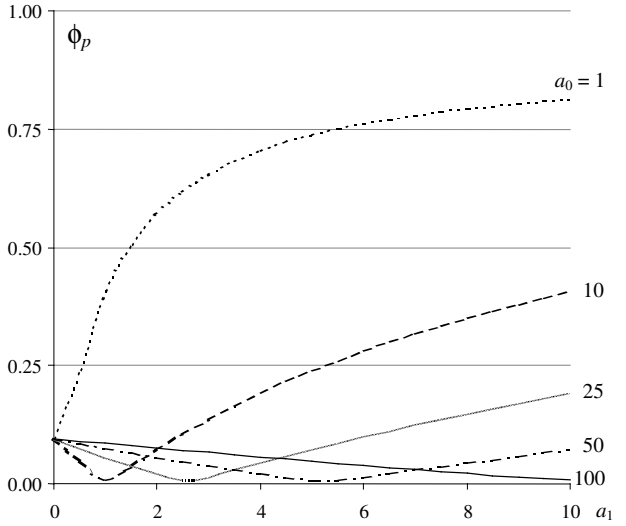


Figure 5. Plotting weighted ϕ_p over the same test distributions.

This formula has one further advantage. So far we have assumed that **O** is a true subset of **D**, so that **O** cannot take a value in any cell where $\mathbf{E}_i = 0$. Since equation (8) does not divide by $\mathbf{E}_i$, this requirement no longer applies and ϕ_p can be applied to any pair of sets with a limit of $2N$.

Consider the pair of distributions **O1** and **O2** in Table 4. To compute ϕ_p we simply rescale **O2** to match **O1** (obtaining **E1**) or vice-versa. We can see that the result is robust, does not require non-zero cell values, and is not dependent on choosing one distribution as a baseline. We can conclude that ϕ_p is a general measure of fit between any two baselines.

	O1	**O2**
a_0	1	0
a_1	0	1
a_2	0	1
a_3	0	0
a_4	1	1,000
a_5	100	1
TOTAL	102	1,003

E1	**(O1-E1)^2**
0	1
0.1017	0.0103
0.1017	0.0103
0	0
101.6949	10,139.4660
0.1017	9,979.6714
102	20,120.1580
χ^2_p	197.2565
ϕ_p	0.9833

E2	**(O2-E2)^2**
9.8333	96.6944
0	1
0	1
0	0
9.8333	980,430.0278
983.3333	964,978.7778
1003	1,945,507.5000
χ^2_p	1,939.6884
ϕ_p	0.9833

Table 4. Measuring the association of a pair of overlapping sets with paired ϕ_p.

6. Variance-weighted ϕ measures

The most general formula for chi-square is the sum of error squares, SS_{err} (see section 8):

$$\chi^2_v = \sum_{i=0}^{k-1} \frac{(\mathbf{O}_i - \mathbf{E}_i)^2}{\sigma_i^2}, \tag{8}$$

where σ_i^2 the expected Gaussian variance based on the prior for a_i, i.e.

$$\sigma_i^2 = p(a_i)(1 - p(a_i))N. \tag{9}$$

This formula is often overlooked. Pearson's χ^2 formula (whose form is very similar but the divisor is $\mathbf{E}_i$ in our notation) is most commonly cited and has been used thus far in computations of ϕ. The justification for Pearson's formula is that Binomial probabilities $p(a_i)$<0.5 are Poisson-distributed

where variance $\sigma_i^2 \approx \mathbf{E}_i$. However in skewed distributions some prior probabilities can be greater than 0.5. Employing the Gaussian equation (9) permits an alternative to standard χ^2.

We have already seen that division by $\mathbf{E}_i$ is the largest source of instability in determining ϕ values, and probabilistic ϕ_p replaced $\mathbf{E}_i$ by N. This removes this instability, but then each squared difference is equally weighted. The limit of χ_v^2 is the product of the sum of the inverse variance $\Sigma\sigma^{-2}$ and N^2.

$$\phi_v = \sqrt{\frac{\chi_v^2}{\sum \frac{1}{\sigma_i^2} N^2}} = \frac{\sqrt{\chi_v^2 / \sum \frac{1}{\sigma_i^2}}}{N}. \tag{10}$$

This formula has a similar 'shape' to other ϕ computations. Like ϕ_p it performs similarly to Cramér's ϕ limited by the minimum value, but it is not dependent on this limit and is therefore relatively robust.

We can also substitute $\sigma_i^2 \approx \mathbf{E}_i$ back into formula (8) and employ regular χ^2 to obtain a new version of Cramér's ϕ, which we will term ϕ_E (Figure 7). The resulting formula is not dependent on a single value of ϕ or scaled by a theoretical minimum $1/N$.

7. Bayesian mean dependent probability

A different approach is suggested by Bayes' Theorem. Here we compute the difference between the observed probability, $p(a_i \mid b)$, and the expected prior, $p(a_i)$. The absolute difference may be scaled as a proportion of the available range to obtain a relative difference, ranging from 0 to 1:

$$dp_R(a_i, b) \equiv \begin{cases} \dfrac{p(a_i \mid b) - p(a_i)}{1 - p(a_i)} & \text{if } p(a_i) < p(a_i \mid b) \\ \dfrac{p(a_i) - p(a_i \mid b)}{p(a_i)} & \textit{otherwise} \end{cases} \tag{11}$$

To combine these proportions we may employ the probabilistically weighted sum, i.e.

$$dp_R = \sum_{i=0}^{k-1} dp_R(a_i, b) \times p(a_i). \tag{12}$$

For any given value i, the value of variation at that point towards the total is proportional to *the*

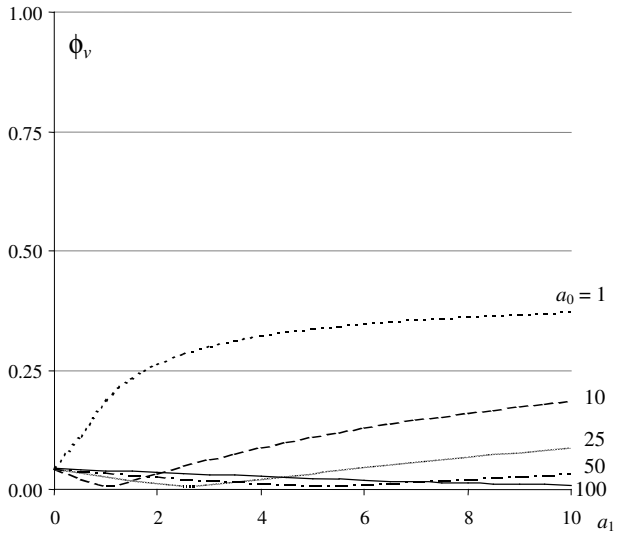


Figure 6. Variance-weighted ϕ_v for $a_0 = 1$, 10, 25 etc., a_1 from 0 to 10.

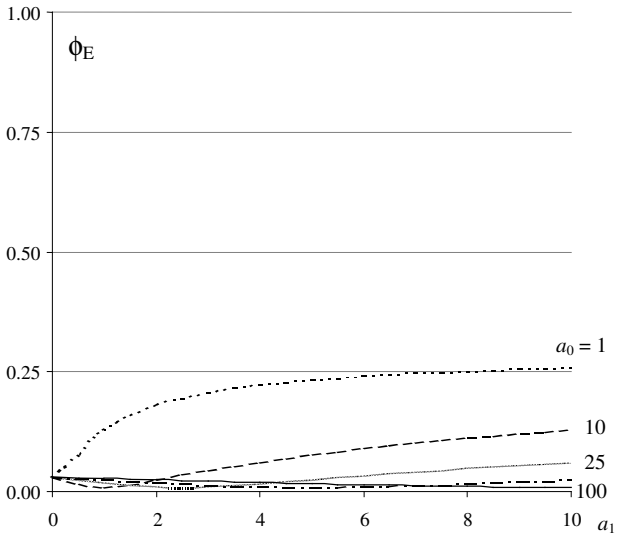


Figure 7. Variance-weighted ϕ_E.

prior probability of selecting it. Like ϕ' this function will also tend to 1 if any single value saturates.

Figure 8 demonstrates that dp_R covers most of the entire range [0, 1] over this computation. The 'floor' is small because the prior probability of selecting a_2 is very low. If, instead of employing the probabilistically weighted sum (13) we take the mean ($\Sigma dp_R(a, b)/k$), this 'floor' rises to 1/3.

Note that whereas chi-square approaches consist of a root mean square summation, this approach simply sums normalised probabilities. A worked example is given in Table 5.

A	**O**	**E**	$p(a \mid b)$	$p(a)$	dp_R	$dp_R \times p(a)$
a_0	1	2.70	0.33	0.90	0.63	0.57
a_1	1	0.27	0.33	0.09	0.27	0.02
a_2	1	0.03	0.33	0.01	0.33	0.00
TOTAL	$N = \mathbf{3}$	**3.00**	**1.00**	**1.00**		**0.59**

Table 5. Obtaining a weighted Bayes' dependent probability association measure.

8. Generalising R^2

We may also consider the coefficient of determination measure R^2, which is conventionally applied to continuous (Pearson r^2) or ranked (Spearman R^2) data. The obvious questions are whether it can be applied to discrete unranked data and if so, whether it improves on previous measures described. Note that unlike the contingency correlation ϕ, R^2 is typically defined such that a value of 1 is interpreted as an exact correlation and 0 represents no correlation:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}, \tag{13}$$

where SS_{err} and SS_{tot} represent the sum of the squares of the error and sample variance respectively. For purposes of comparison we will simply reverse this subtraction, and take the square root of the ratio obtaining a new ratio which we will call R^*:

$$R^* = \sqrt{\frac{SS_{err}}{SS_{tot}}}\,. \tag{14}$$

R^2 is conventionally applied to N observations, but in the case of a contingency table, i.e. where multiple observations are found in each cell, we need to factor in cell variance in these summations,

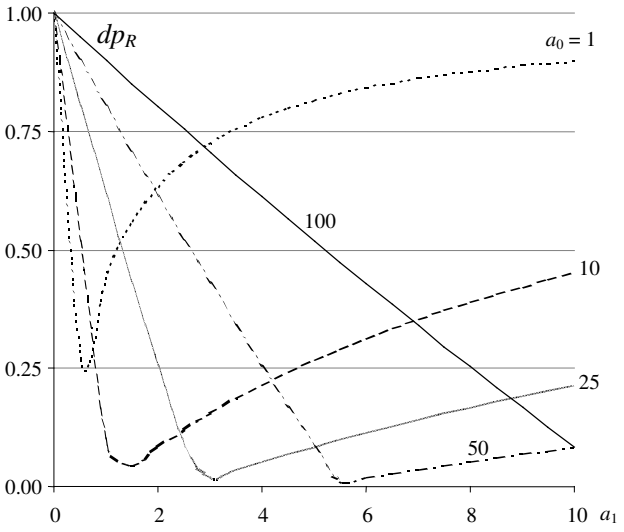


Figure 8. Plotting dependent probability *dpR*.

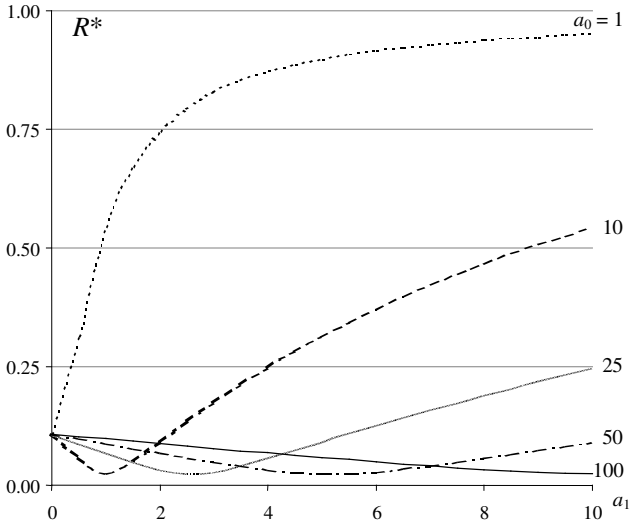


Figure 9. Plotting R^* over test distributions.

obtaining

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

where $\overline{O}$ represents the mean observation (either by simple division or by probabilistic summation) and σ_i^2 the expected Gaussian variance for a_i (equation 9).

Over our data this formula behaves similarly to ϕ_p as Figure 9 reveals. However it is not constrained probabilistically. Indeed SS_{tot} can tend to zero, obtaining an infinite R^*. R^* compares the deviation of the error (difference of observed to expected) to the overall deviation of the observed results from a constant mean $\overline{O}$. This is not a particularly useful comparison in categorical data! The fact that ϕ_p and R^* behave similarly over the same range and ϕ_p is better behaved, means that we can discount R^* as an alternative to ϕ-based measures. Nagelkerke's modification of R^2 constrains the measure to a limit similarly to the way we constrained ϕ to ϕ_E.

9. Numerical evaluation of extrema

We can draw out differences between measures by a tabular comparison of extreme sample points (Table 6, Figure 10). Recall that the expected distribution is highly skewed (**D** = {100, 10, 1}). The first line in Table 6 demonstrates that all functions obtain 0 when the expected and observed distributions match. Both reduced chi-square and R^* can obtain values in excess of 1.

The next three rows list measures when a single value is found in one cell and all others are zero. All functions with the exception of χ^2_{red} are scaled by N, and the value of the middle of the three, **O** = {0, 1, 0}, matches the fifth row, {0, 10, 0}, with this exception. We can also see that two measures, normalised ϕ' and dp_R, score 1 for all three maximally-skewed cases. Other measures order their scores so as to treat maximum saturation at the term with the lowest expected probability. Note that no other ϕ measure scores this value as 1. Saturation at the middle value a_1 obtains a ~90% score for ϕ_p but ϕ_v and ϕ_E have a lower score, much closer to Cramér's ϕ (against $1/\min(p(a))$). ϕ_p is less sensitive to variation at these lower-valued points.

The second set of three rows show what happens when data is distributed evenly between two out of three cells. Here we do not expect values to be equal. With the exception of ϕ', the highest-scoring patterns are in the second row {0, 1, 1}, which seems intuitively sound. Both probabilistically weighted estimates, dp_R and ϕ_p, score this approximately twice as much as the other two patterns, which include a 1 in the most probable cell a_0. ϕ_v rates it higher by ~7 percentage points, whereas Cramér's ϕ and ϕ_E rate it higher by about 2.

If we now see what happens if the number of cases in the least probable outcome, a_2, increases by 1,

O			
a_0	a_1	a_2	Σ
100	10	1	111
1	0	0	1
0	1	0	1
0	0	1	1
0	10	0	10
1	1	0	2
0	1	1	2
1	0	1	2
0	1	2	3
1	0	2	3

measures of fit							
χ^2_{red}	ϕ	ϕ_E	ϕ_v	ϕ_p	ϕ'	dp_R	R^*
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.055	0.0316	0.0299	0.0401	0.0949	1.0000	1.0000	0.0116
5.05	0.3030	0.2863	0.3766	0.9054	1.0000	1.0000	1.0046
55	1.0000	0.9449	0.9382	0.9492	1.0000	1.0000	2.2754
50.5	0.3030	0.2863	0.3766	0.9054	1.0000	1.0000	1.0046
2.0525	0.1366	0.1291	0.1688	0.4055	0.8672	0.4505	0.2947
29.525	0.5181	0.4895	0.5307	0.7813	0.8672	0.9459	8.1246
27.0275	0.4957	0.4684	0.4620	0.4527	0.9852	0.4955	6.0489
74.35	0.6713	0.6343	0.6559	0.8072	0.9295	0.9310	4.2555
72.685	0.6637	0.6271	0.6206	0.6176	0.9925	0.6636	3.7789

Table 6. Assessing measures of fit: sample points against **D** = {100, 10, 1}.

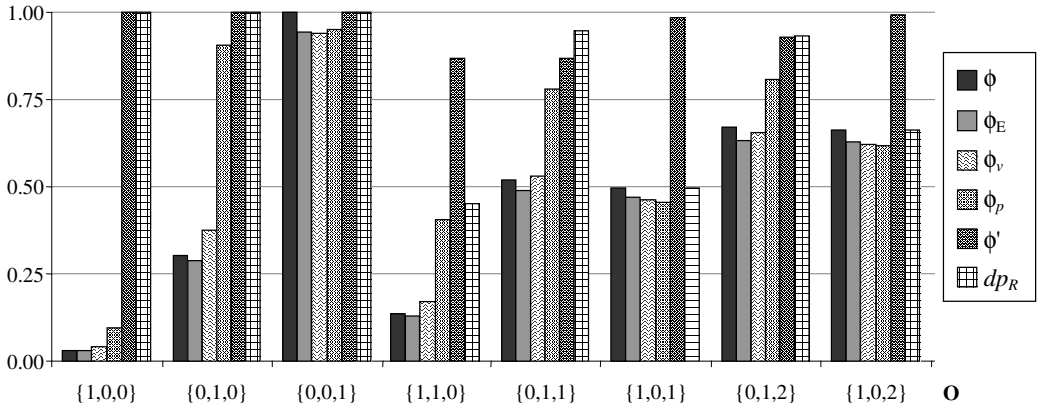


Figure 10. Visualising probabilistic measures of fit in Table 6. χ^2_{red} and R^* are not shown.

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 χ^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 χ^2 partials, or, to put it another way, it is proportional to the absolute sum of squares. Absolute variation at more probable terms (here, a_0) contributes a greater amount to the overall sum than would otherwise be the case. This fact is important to bear in mind when employing ϕ_p, and it is conceivable that a comparison could lead to a different ranked order than ϕ_E.

10. Correlating the present perfect

In the introduction we summarised the type of problem we wished to apply these measures to. It seems apposite to return to this example in conclusion. Selected correlation estimates are computed below. We have also included unconstrained Cramér's ϕ for comparison purposes. We model for the total distribution baseline to be inclusive of the observed (see Table 1). In this case we also examine ratios of measures in addition to the measures themselves. Variation of ratio tells us whether we can reliably employ a given measure to distinguish two baselines (which is of course the entire rationale for this exercise).

We will first evaluate DCPSE sampling categories to demonstrate the approach.

10.1 Time: LLC vs. ICE-GB

We apply our new measures to the data in Table 1 to obtain Table 7. The pair of 2×1 tables are approximately evenly distributed, and, as we suggested at the outset, with only two categories it is difficult to distinguish measures by performance. Results fall into two groups. Variance-weighted ϕ_v and ϕ_E perform similarly to probabilistically-weighted ϕ_p. ϕ and ϕ' are approximately double these scores. The past:present ratio between correlation measures is near constant, meaning that, irrespective of the chosen measure, present-perfect correlates closer to the present.

time	ϕ	ϕ'	ϕ_p	ϕ_v	ϕ_E
present	0.0227	0.0227	0.0114	0.0114	0.0114
past	0.0695	0.0694	0.0346	0.0346	0.0346
ratio	3.0587	3.0521	3.0408	3.0408	3.0408

Table 7. Comparing correlation measures for the present perfect against present and past tensed VP baselines, measured across LLC and ICE-GB subcorpora.

Note that the ICE-GB vs. LLC distinction is not evenly balanced by text size, with ICE-GB texts containing 2,000 words and LLC texts of 5,000 words. The total number of words is very similar and sampled for broadly equivalent text categories. Bear in mind that this evaluation has a single degree of freedom, and it should not be surprising therefore that different approaches to measuring correlation achieve the same result.

10.2 Genre

Next, we apply these measures to the ten sociolinguistic genre text categories of DCPSE to obtain Table 8. This confirms that, again, measured across text categories, present perfect constructions tend to correlate more closely with present tensed VPs than those marked for past tense. However the scores themselves are much more varied.

genre	ϕ	ϕ'	ϕ_p	ϕ_v	ϕ_E
present	0.0594	0.0871	0.0460	0.0104	0.0095
past	0.1049	0.1545	0.0642	0.0214	0.0207
ratio	1.7655	1.7733	1.3950	2.0596	2.1721

Table 8. Measuring goodness of fit across 10 DCPSE text categories.

In Table 7 we obtained very similar ratios between scores. However in Table 8 the ratio for ϕ metrics varies from between 1.4 and 2.2 times, with the greatest ratio applying to variance-based measures. This is also the only table where ϕ_v and ϕ_E differ by more than 1%. Text category obtains quite different results depending on the measure used.

We have already seen that text categories in DCPSE are extremely uneven in size (Figure 1), ranging from 126 to 3 texts per category. Moreover text categories may be influencing the baselines by grouping texts with more present- and past-referring VPs together. We return to this below.

10.3 Texts and subtexts

Finally, we perform the same set of calculations over every distinct text and subtext in DCPSE. As we note above, **texts** are approximately equal in length within LLC and ICE-GB. On the other hand **subtexts** differ in length from short phone-calls of a few hundred words to long monologues of 5,000+. Again, higher figures imply a lesser degree of correlation.

text	ϕ	ϕ'	ϕ_p	ϕ_v	ϕ_E
present	0.0247	0.0270	0.0184	0.0013	0.0013
past	0.0349	0.0401	0.0298	0.0017	0.0017
ratio	1.4141	1.4848	1.6204	1.2963	1.2952

subtext	ϕ	ϕ'	ϕ_p	ϕ_v	ϕ_E
present	0.0225	0.0333	0.0177	0.0005	0.0005
past	0.0303	0.0348	0.0280	0.0006	0.0006
ratio	1.3478	1.0436	1.5776	1.1259	1.1251

Table 9. Measuring goodness of fit against texts (upper) and subtexts (lower).

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.

Does the simple fact that texts are grouped into larger categories explain the difference in scores between Table 8 and 9? Are measures affected by **scale**, or by particular **distribution**?

10.4 Pseudo-genre

To answer this question we employ the following computational approach, and compare the results with Table 8 and 9.

- *Randomly assign each text into one of ten pseudo-text categories, with the same total number of texts per category as DCPSE's actual text categories. Calculate goodness of fit measures over these totals, and repeat 10,000 times to obtain a set of mean values.*

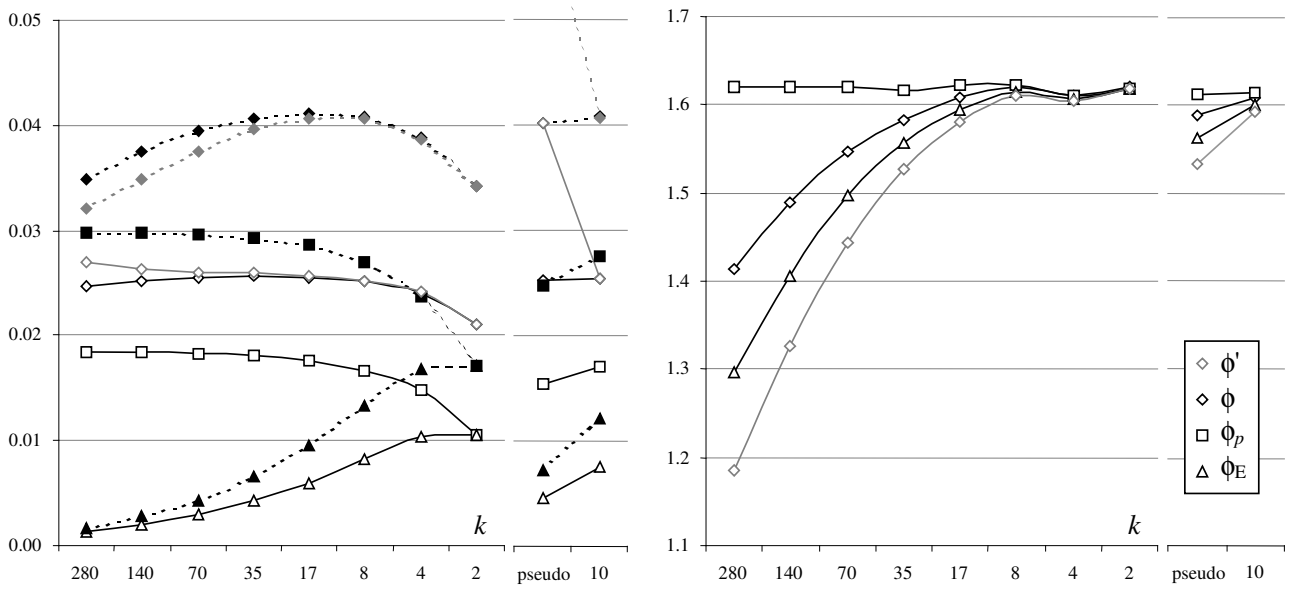
First, let us examine the effect of this resampling (cf. Table 9, upper). We can see that ϕ_p, ϕ_v and ϕ_E appear to be affected by the reduced number of categories, whereas unconstrained Cramér's ϕ seems more stable. Variance-weighted ϕ measures increase four-fold, whereas probabilistically-weighted ϕ_p has fallen by about a fifth. However the most useful comparison is to examine the *ratio* of measures, and here ϕ_p is stable across scales.

'genre'	ϕ	ϕ'	ϕ_p	ϕ_v	ϕ_E
present	0.0253	0.0401	0.0154	0.0048	0.0046
past	0.0402	0.0615	0.0248	0.0075	0.0072
ratio	1.5899	1.5333	1.6133	1.5655	1.5628

Table 10. Mean goodness of fit over 10,000 resamples of texts into pseudo 'text categories'.

The fact that Cramér's ϕ appears little affected by the grouping suggests that the rather higher scores in Table 8 are due to an interaction between text category and 'pastness'. This may also explain the difference between ϕ_v and ϕ_E. In other words, the real text categories tend to group texts according to whether they refer to the present or past, which this random allocation does not do.

We can test this hypothesis by carrying out a 10×2 χ^2 test and comparing χ^2 or Cramér's ϕ scores. In the genuine case we obtain a total $\chi^2 = 2{,}710$, and $\phi = 0.1664$. In the pseudo-category cases we obtain a mean $\chi^2 = 398$, and $\phi = 0.0621$. Note that this increase in ϕ (2.7 times) is of roughly the same order as increases in the ϕ column (present: 2.4, past 2.9 times). Examining χ^2 partials, we can see that broadcast interviews, spontaneous commentary, prepared speech and legal cross-examination categories all differentiate texts by 'pastness'. By way of contrast, the size of the effect of time on 'pastness' (cf. Table 7) is below this random allocation ($\chi^2 = 206$, $\phi = 0.0438$).



(a) ϕ measures (solid = present, dotted = past) (b) past:present ratios

Figure 11: Effect of number of categories k on Cramer's ϕ, ϕ', ϕ_E and ϕ_p.

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 pseudo-genre category, but this pseudo-genre was uneven in distribution. We would like to know how measures behave if we merely vary scale alone. We need another computational assessment.

The 280 texts are randomly reallocated to k categories 10,000 times and mean values of Cramer's ϕ, ϕ', ϕ_p and ϕ_E are computed for the present perfect against present- and past- VP baselines. Results are plotted in Figure 11. All values of ϕ are affected by the number of categories, but ϕ_p appears more stable than other measures, with the present:past ratio near-constant across all scales.

The final pair of columns in the plots in Figure 11 demonstrates the effect of different sizes of genre subcategories. This compares random sampling into uneven pseudo-categories (with between 126 and 3 texts per category) with random sampling into 10 categories of 28 texts each.

In conclusion, in evaluating measures we should consider a number of questions.

- **Stability of measures and stability of ratios.** Stability of ϕ measures means that a particular ϕ cited with k corpus subdivisions would predict ϕ with a different k. Stability of *ratios* means that the ratio between two measures is constant over k. Thus it seems that Cramér's ϕ is highly stable measuring against the present VP baseline from 280 to 17 (Figure 11(a), middle), but it increases steadily against the past over the same range, and therefore the ratio steadily increases. On the other hand ϕ_p appears to fall in an arc with increasing k in both cases, but the past:present ratio (Figure 11(b)) is extremely stable.

- **The impact of the number of categories**. As we have seen, all measures are affected by the number of categories, with ϕ_E tending to increase as k falls and ϕ_p decreasing (although not as dramatically). Different measures tend to converge ($\phi_E \leftrightarrow \phi_p$, $\phi \leftrightarrow \phi'$) as k approaches 2, a by-product of the Central Limit Theorem. However the impact of k is less dramatic than that found with conventional Cramér's ϕ. Comparing past and present VPs, $k \times 2$ ϕ falls with k (Figure 12).

- **The impact of uneven-sized categories**. The differences between irregular-sized pseudo-categories and categories with evenly allocated texts parallel the variation in scale, with the exception that ϕ' is particularly affected by uneven categories.

- **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.

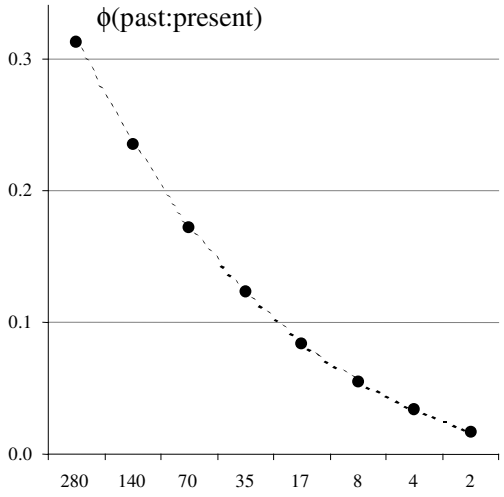


Figure 12: Effect of scale on $k \times 2$ ϕ.

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

$$\hat{\phi}_p(k) = \phi_p(K) + x/K - x/k.$$

These curves allow us to find $x = \phi_p(K)/y$, where $y \approx 1.2$, simplifying to

$$\hat{\phi}_p(k) = \phi_p(K)(1 + 1/yK - 1/yk). \qquad (16)$$

Further experimentation with DCPSE and ICE-GB finds optimum values of $y \approx 1.17$ and 1.25 respectively. This result appears robust with respect to other queries, but subsets obtain a scatter and may needed to be tested individually.

References

Andrae, R., Schulze-Hartung, T. and P. Melchior. 2010. *Dos and don'ts of reduced chi-squared*. Cornell University. eprint arXiv:1012.3754

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* χ^2. *Journal of Quantitative Linguistics*. www.ucl.ac.uk/english-usage/statspapers/z-squared.pdf

Wallis, S.A. 2011. *Comparing* χ^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