Adapting random-instance sampling variance estimates and Binomial models for random-text sampling

Sean Wallis, Survey of English Usage, University College London
Originally published online 2015, updated 2020

1. Introduction

Conventional stochastic methods based on the Binomial distribution rely on a standard model of random sampling whereby freely-varying instances of a phenomenon under study can be said to be drawn randomly and independently from an infinite population of instances.

These methods include confidence intervals and contingency tests (including multinomial tests), whether computed by Fisher's exact method or variants of log-likelihood, χ^2, or the Wilson score interval (Wallis 2013). These methods are also at the core of others. The Normal approximation to the Binomial allows us to compute a notion of the variance of the distribution, and is to be found in line-fitting and other generalisations.

In many empirical disciplines, samples are rarely drawn "randomly" from the population in a literal sense. Medical research tends to sample available volunteers rather than names compulsorily called up from electoral or medical records. However, provided that researchers are aware that their random sample is limited by the sampling method, and draw conclusions accordingly, such limitations are generally considered acceptable. Obtaining consent is occasionally a problematic experimental bias; actually recruiting relevant individuals is a more common problem.

However, in a number of disciplines, including *corpus linguistics*, samples are not drawn randomly from a population of independent instances, but instead consist of randomly-obtained contiguous subsamples. In corpus linguistics, these subsamples are drawn from coherent passages or transcribed recordings, generically termed 'texts'. In this sampling regime, whereas any pair of instances in independent subsamples satisfy the independent-sampling requirement, pairs of instances in the same subsample are likely to be co-dependent to some degree.

To take a corpus linguistics example, a pair of grammatical clauses in the same text passage are more likely to share characteristics than a pair of clauses in two entirely independent passages. Similarly, epidemiological research often involves 'cluster-based sampling', whereby each subsample cluster is drawn from a particular location, family nexus, etc. Again, it is more likely that neighbours or family members share a characteristic under study than random individuals.

If the random-sampling assumption is undermined, a number of questions arise.

- Are statistical methods employing this random-sample assumption simply *invalid* on data of this type, or do they gracefully degrade?
- Do we have to employ *very different tests*, as some researchers have suggested, or can existing tests be modified in some way?
- Can we measure the *degree* to which instances drawn from the same subsample are interdependent? This would help us determine both the scale of the problem and arrive at a potential solution to take this interdependence into account.
- Would revised methods only affect the *degree of certainty* of an observed score (variance, confidence intervals, etc.), or might they also affect *the best estimate of the observation* itself (proportions or probability scores)?

2. Previous research

2.1 Employing rank tests

One approach, suggested by Vaclav Brezina and others (see e.g. Brezina and Meyerhoff 2014), is to dispense with Binomial models entirely, on the basis that they employ an assumption that is unsafe.

Brezina suggests using statistics premised on rank order correlations. In examining the frequency of particular lexical items (words) per text, he successfully demonstrates that standard Binomial models are unreliable for this reason, and proposes the Mann-Whitney test in place of χ^2.

This approach has three fundamental drawbacks.

- It dispenses with a great deal of information from the sample (actual scores are replaced by rankings and instances by whole texts), so the method is necessarily conservative. (It is axiomatic in statistics that where one sound model is premised on less information than another, the lower-information model will need to err on the side of caution and be more conservative.)

- By dispensing with the assumption that data is Binomial, the method is of limited generality. The Mann-Whitney test compares two grouped subsamples, but not multiple groups on multiple dimensions (cf. χ^2). Although it might conceivably be feasible to extend the test in this way, this is not a commonplace.

- Results are also of limited utility. We lose the ability to cite, with confidence, the true rate of an observed proportion or probability. This is because the test does not compare two observed rates of occurrence, but tests instead if the scores in one group of texts tend to have a significantly higher rank than those in another group. We may be able to *infer* that this observed ranking difference is probably the result of a higher true rate in the population from which samples are drawn. But we cannot cite any estimate of this true rate.

2.2 Case interaction models

At the other end of the information spectrum lies a method that attempts to mine as much information as possible from the sample. A corpus linguistics approach we have experimented with, and which remains a work in progress at the time of writing, attempts to estimate the interdependence between instances within subsamples by closely examining their source texts.

This is a theory-led approach. In the case of grammatical clauses, which we use as an exemplar below, linguistically we can anticipate that clauses are to be found

- embedded within other clauses to varying degrees of depth,
- found in sequences at varying distances, and
- in a co-ordinated list (for non-linguists: in a sequence with ‘and’, ‘or’, ‘but’, etc.).

In a parsed corpus, such as ICE-GB (Nelson, Wallis and Aarts 2002), not only can we obtain reliable counts of clauses but we can also obtain measures of grammatical proximity along these varying axes.

This ‘case interaction’ approach (Wallis and Aarts 2007) aggregates data from all instances to construct a model, and then estimates a prior probability score for every instance. Methods based on counting frequencies are adjusted to simply sum prior probabilities.

Compared to the method of this paper, the case interaction approach has an important potential benefit. In building this model, we first gather detailed information regarding the degree to which *specific* instances within a text (subsample) interact with others. This is potentially a valuable research method in its own right. In linguistics, studying the interaction between lexical and grammatical decisions (i.e. where the likelihood of a subsequent choice by a speaker being of a particular type is affected by an earlier choice) offers a route into psycholinguistic models of priming and spreading activation.

Note that not all sources of interdependence in a passage will be psycholinguistic. A range of contextual influences on co-occurring clauses, from topics of discussion to semantic and social biases, will tend to affect them simultaneously.

A final potential advantage of the case interaction approach is that with sufficient data it should be possible to control for these alternative sources of interaction (i.e. focus on psycholinguistic causes alone). From the purely stochastic point of view of this paper, however, such an ideal psycholinguistic model, by excluding these other sources of interaction, would still tend to overestimate the independence of instances and the significance of results. Therefore we see this research programme as complementary for what follows.

3. Adjusting the Binomial model

We will employ a method related to ANOVA and F-tests, applying the approach to a probabilistic rather than linear scale. This step is not taken lightly but as we shall see in Section 6, it can be justified.

Consider an observation p drawn from a number of texts, t', based on n total instances. These must be *non-empty texts*, i.e. for all $i = 1..t'$ texts, $n_i > 0$. Were the sample drawn randomly from an infinite population, we could employ the Normal approximation to the Binomial distribution

$$\begin{aligned} &\textit{standard deviation } S \equiv \sqrt{P(1-P)/n}\,, \\ &\textit{variance } S^2 \equiv P(1-P)/n\,, \end{aligned} \tag{1}$$

where P is the mean proportion in the population and n the sample size. Inverting the above (so that $P = w^-$ when $p = P + z_{\alpha/2}.S$), we have the Wilson score interval, which may be directly computed by

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

where $z_{\alpha/2}$ is the critical value of the Normal distribution for a given error level α (see Wallis 2013 for a detailed discussion). Other derivations from (1) include χ^2 and log-likelihood tests, least-square line-fitting, and so on.

The model assumes that all n instances are randomly drawn from an infinite (or very large) population. However, we suspect that our subsamples are not equivalent to random samples, and that this sampling method will affect the result.

To investigate this question, our approach involves two stages. First, we measure the variance of scores between text subsamples according to two different models, one that simply assumes that each subsample is a random sample, and another calculated from the distribution of subsample scores. Consider the frequency distribution of probability scores, p_i, across all t' non-empty texts. This is centred on the mean probability score of the subsamples, $\bar{p}$. An example distribution of this type is illustrated by Figure 1.

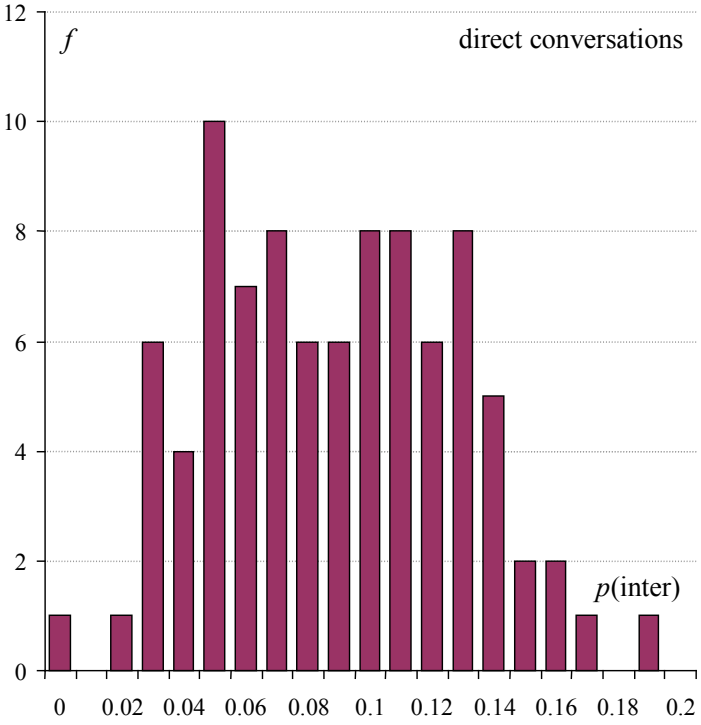

Figure 1: A frequency distribution of mean p per text, quantised to two decimal places, tends to the Binomial distribution. See Section 4.

$$\textit{subsample mean } \bar{p} = \frac{\sum p_i}{t'}.$$

If subsamples were randomly drawn from the population, it would follow from (1) that the variance could be **predicted** by

$$\textit{predicted between-subsample variance } S_{ss}^2 = \frac{\bar{p}(1-\bar{p})}{t'}. \tag{3}$$

To measure the **actual** variance of the distribution we employ a method derived from Sheskin (1997: 7). The formula for the *unbiased estimate of the population variance* may be obtained by

$$\textit{observed between-subsample variance } s_{ss}^2 = \frac{\sum (p_i - \bar{p})^2}{t'-1} \tag{4}$$

where $t' - 1$ is the number of degrees of freedom of our set of observations.

Second, we adjust the weight of evidence for the sample estimate according to the degree to which these two variances (Equations (3) and (4)) disagree. If the observed and predicted variance estimates coincide, then the total set of subsamples is, to all intents and purposes, a random sample from the population, and no adjustment is needed to sample variances, standard deviations, confidence intervals, tests, etc.

We can expect, however, that in many cases the actual distribution has greater spread than that predicted by the randomness assumption. In such cases, we employ the **ratio of variances**, F_{ss}, as a scale factor for the number of random independent cases, n.

Gaussian variances with the same probability p are inversely proportion to the number of cases supporting them, n, i.e. $s^2 \equiv p(1-p)/n$ (cf. Equation (1)). We might estimate a corrected independent sample size n' by multiplying n by F_{ss} and scale the weight of evidence accordingly.

$F_{ss} = S_{ss}^2 / s_{ss}^2$, and
adjusted sample size $n' = n \times F_{ss}$.

To put it another way, the ratio n':n is the same as S_{ss}^2:s_{ss}^2. This adjustment substitutes the observed variance (Equation (4)) for the predicted Gaussian variance. However, we already know that t' cases must be independent, so we scale the excess, $n - t'$.

$$\textit{adjusted sample size } n' = (n - t') \times F_{ss} + t'. \tag{5}$$

Thus if $n = t'$, n' is also equal to t'. If F_{ss} is less than 1, the weight of evidence (sample size) decreases, $n' < n$ and confidence intervals become broader (less certain).

If we decrease n in Equations (1) and (2), we obtain larger estimates of sample variance and wider confidence intervals. An adjusted n is easily generalised to contingency tests and other methods.[1]

In order to evaluate this method, we turn to some worked examples.

[1] Our approach is worth comparing to the *finite population correction* for random samples drawn from a small finite population. An observed value in a sample will tend to converge on the true population value as the sample becomes a greater proportion of the population. See Wallis (2012d). Singleton *et al.* (1988) proposed dividing sample size n by $v^2 = 1 - n/N$, where N is the finite population size. It is now standard to employ $v^2 = (N - n)/(N - 1)$. Either way, the effect is to increase n and thereby increase estimated confidence in results from the standard random-sample, infinite-population model.

	CL	CL(inter)	Words	*p*(inter)
ICE-GB	145,179	5,793	1,061,263	0.0399
spoken	90,422	5,050	637,682	0.0558
dialogue	57,161	4,686	376,689	0.0820
private	32,658	2,901	205,627	0.0888
direct conversations	29,503	2,617	185,208	0.0887
S1A-001	322	20	2,050	0.0621
S1A-002	328	19	2,055	0.0579
S1A-003	334	25	2,146	0.0749
S1A-090	326	44	1,968	0.1350

Table 1: Snippet of data table extracted from ICE-GB.

4. Example 1: interrogative clause probability, direct conversations

Drawing frequency statistics from ICE-GB for interrogative clauses (denoted by 'CL(inter)') and all clauses ('CL'), for each text in ICE-GB (approximately 2,000 words), we obtain a large hierarchical table, excerpted in Table 1. Here, "ICE-GB" represents the total across all of ICE-GB etc. There are 90 texts within the direct conversations category. This shows there are 29,503 clauses in 90 texts.

At this point we assume that each text is approximately the same size and we make no assumptions about their further subdivision (separately-sampled subtexts, speakers), etc. We return to this question in Section 6. We merely assume that texts are randomly drawn from a population of comparable texts. Our approach is therefore conservative.

Consider the probability that a clause is interrogative (p(CL(inter) | CL), or 'p(inter)' for short). For direct conversations,

$$\textit{observed probability } p = p(\text{inter}) = \frac{f(\text{CL(inter)})}{f(\text{CL})} = 0.0887.$$

If we make the standard assumption that the corpus is a random sample, we obtain the following standard deviation, variance and Wilson score intervals on this observation in the normal way (Equations (1) and (2), see also Wallis 2013).

$n = f(\text{CL}) = 29{,}503.$
standard deviation $s = 0.001655.$
Wilson interval $(w^-, w^+) = (0.0855, 0.0920)$ at a 95% error level.

However, this model assumes that the 29,503 clauses are drawn at random from the infinite population of English native-speaker direct conversations. The question is: *to what extent are these measures of uncertainty an underestimate?*

First, we consider each **text** as a sample drawn from the population of similar-size texts. To what extent are interrogative clauses more common in some texts than others, and is this variation merely due to chance? A visualisation may help.

A frequency distribution of p(inter) across 90 texts may be obtained by quantisation and summation, i.e. rounding each probability into a series of discrete intervals, and summing the number of texts falling into this interval. This "frequency distribution by text" will tend to be Binomial (see Figure 1).

Next, we compute a Normal approximation to this observed Binomial distribution. The mean of this distribution, $\bar{p}$, is simply the mean over the number of texts, so

$$\bar{p} = \frac{\sum p_i}{t'} = 0.0855.$$

Employing Equations (3) and (4) obtain

$$S_{ss} = \sqrt{\frac{\bar{p}(1-\bar{p})}{t'}} = 0.0299, \text{ and}$$
$$s_{ss} = \sqrt{\frac{\sum (p_i - \bar{p})^2}{t'-1}} = 0.0395.$$

The two Normal distributions for s_{ss}, both centred on $\bar{p}$, are shown in Figure 2, alongside the observed Binomial frequency distribution of texts.

The next step is to estimate the extent to which this distribution tends to 'cluster', i.e. the degree to which these interrogative clauses do not distribute randomly.

ratio $F_{ss} = S_{ss}^2/s_{ss}^2 = 0.5757$.
number of cases $n' = 17{,}045.28$.
standard deviation $s = 0.002178$,
95% Wilson interval $(w^-, w^+) = (0.0845, 0.0931)$.

The order of magnitude in this case is comparable to a suggestion we made on the **corp.ling.stats** blog in 2012: that dividing n by 2 would be a reasonable conservative estimate of the effect of compensating for within-text variance (Wallis 2012a). Naturally now we can be much more precise.

4.1 Alternative method: Fitting

One concern may be that Equation (4) does not appear to optimally match the observed discrete Binomial distribution (see Figure 2). An alternative method for estimating s_{ss} involves a computational search procedure.

By a process of minimising the sum of square differences between the observed distribution $f(p)$ and the desired Gaussian probability function, Z, we vary the standard deviation parameter s_{ss}.

$$e = \sum_{p=0\ldots1} (Z(\bar{p}, s_{ss}, p) - f(p))^2 \,. \quad (6)$$

This method uses the Gaussian probability function $Z(\bar{p}, s, p)$ with parameters mean $\bar{p}$, standard deviation s, and probability p. By minimising e, this obtains $s_{ss} \approx 0.0474$.

Note that, apart from relying on a computationally intensive search procedure, this method also depends on a complex Gaussian function and the quantised $f(p)$ distribution. In this case, the method of fitting obtains a greater estimate of spread than by direct computing using Equation (4).

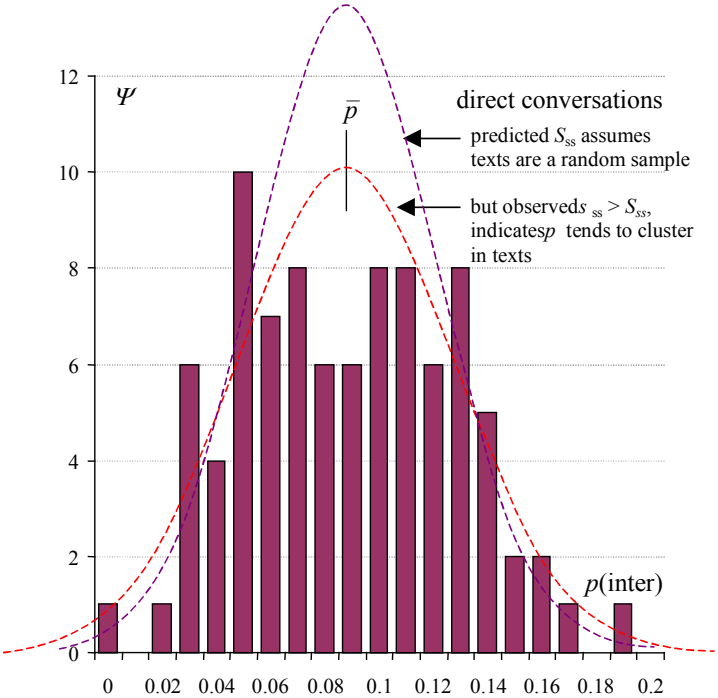

Figure 2: Estimating Normal approximations to the frequency distribution of subsamples.

Using the fitted estimate for s_{ss} we obtain

cluster-adjustment ratio F_{ss} = 0. 4008.
number of cases n' = 11,825.
standard deviation s = 0.002615.
95% Wilson interval (w^-, w^+) = (0.0837, 0.0940).

In this case, a scaling ratio of about 0.4 means that the variance for p(inter) will be approximately 2.5 times greater than would be expected if cases were drawn randomly from the population rather than whole texts, and the standard deviation and Wilson error intervals are about 1.6 times greater (the standard deviation being the square root of the variance).

5. Example 2: Clauses per word, direct conversations

The next example utilises a word-based baseline, a baseline that is common in corpus linguistics for many purposes. This baseline obtains *exposure* probabilities, i.e. the probability that a reader or hearer will be exposed to a particular word, sequence or construction. It does not give us *choice* (or utilisation) probabilities, unlike our first example. Word-based baselines also tend to be subject to between-genre sensitivity.

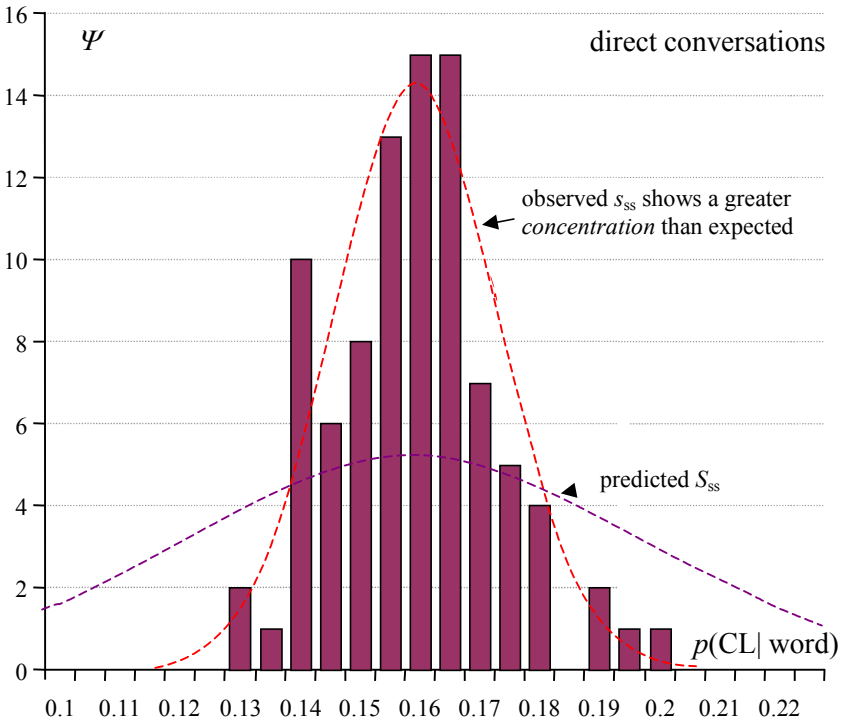

Figure 3: Frequency distribution per text, p(CL | word), ICE-GB direct conversations, with Gaussian curves based on predicted and observed measures of standard deviation.

Suppose we wish to study the mean length of clauses, i.e. the number of words per clause. This variable is not a probability, but we can perform a simple mathematical 'trick' to allow us to work with probabilities (Wallis 2012b). We take the reciprocal of the measure, i.e. *the number of clauses per word*, perform computations of variance, confidence intervals, statistical tests, etc., and finally invert this probability for citation and plotting purposes. Another way of expressing the number of clauses per word is that it is the probability that a random word is the first word in a clause.

Using the ICE-GB direct conversation data we obtain the following

observed probability p = p(CL) = f(CL)/f(word) = 0.159297.
number of cases n = f(word) = 185,208.
standard deviation s = 0.00850.
95% Wilson interval (w^-, w^+) = (0.15921, 0.15940).

This is a small interval due to the large number of words, n, obtained. Next, we examine the distribution of texts. Does the rate of clauses per text vary more than would be expected by chance?

distribution mean $\bar{p}$ = 0.159277.

predicted standard deviation $S_{ss} = 0.038573$.
observed standard deviation $s_{ss} = 0.013781$ (Equation (4)).
cluster-adjustment ratio $F_{ss} = S_{ss}^2/s_{ss}^2 = 7.8341$.

This result is, initially at least, rather surprising. So far we have assumed that the predicted standard deviation (calculated by assuming subsamples are proper random samples), would be **smaller** than the observed standard deviation, based on the distribution of the subsamples themselves. But in this case the predicted standard deviation is nearly three times **greater** than the observed! A slight increase might be explained by rounding errors. But this is a substantial increase. What is going on?

The fitting method for s_{ss} obtains almost exactly the same result in this case. This is not an artefact of the method of calculation of standard deviation. The explanation must lie elsewhere.

The Binomial model for a variable p assumes that p is free to vary from 0 to 1. Logically, in our example, this would mean that *every word in the corpus could be the first word in a clause*. But this is impossible in practice. However many times we sample clauses, max(p) will be less than 1.

If the range of p is compressed, then the Binomial model, and approximations to it (1), necessarily overestimate the variance of p, and the degree of overestimation can be substantial (Wallis 2012c). Many linguistic variables, including word-based probabilities, do not maximise at 1. It is unrealistic to assume that an entire corpus would consist of a given word, structure, etc., yet many methods make precisely this freedom-to-vary assumption. A free choice between lexical or grammatical alternates may range from 0 to 1, but the number of clauses per word is not such a variable.

By examining the frequency distribution of subsample mean probabilities, we observe the actual behaviour of p and how it distributes, and obtain s_{ss}. As a side-effect of our method, therefore, we are able to compensate for this kind of overestimation of intervals. We are entitled to increase n by F_{ss} and reduce confidence intervals still further.

number of cases $n' = 1{,}450{,}933$.
standard deviation $s = 0.000304$.
95% Wilson interval $(w^-, w^+) = (0.159277, 0.159318)$.

The argument is that we are drawing more information from the sample, and thus we are entitled to improve the precision of our tests, confidence intervals, etc. This decreases the absolute variance.

6. Uneven-size subsamples

Note that the subsample distribution mean, $\bar{p}$ (averaged over subsamples) and the sample-wide mean, p (averaged over instances) are not necessarily the same. The reason they may differ is that subsamples can vary in size. However, Equation (4) assumes that all subsamples are the same size.

In the previous examples, subsample sizes were fairly consistent, so this is not really an issue. However, for robustness, we should allow for cases where source texts may vary in size (i.e. texts are permitted to vary in word length). In ICE-GB, for example, some texts are composed of shorter independent subtexts. Moreover, even if the number of words per text was exactly constant, for any given research question the number of instances abstracted from the text, n_i, may vary. In examining the tendency for a clause to be interrogative, we should take heed of the fact that the number of clauses per word also varies between texts (as we saw in Figure 3).

A better estimate of between-subsample variance may be obtained by generalising Equation (4) to uneven-size subsamples. As a result, $\bar{p} = p$, which is an immediate simplification.

Equation (4'), which obtains the internal variance of the set, can be generalised to the following probabilistically-weighted formula.

$$s_{ss}^2 = \sum p(x_i)(p_i - p)^2 \,, \qquad (7')$$

where prior probability $p(x_i) = n_i / n$. The prior probability is the chance that a random instance would be found in subsample i. For mathematical rigour, we can show that this 'ANOVA-type' approach is applicable to Binomial probabilities. Where the prior is an ideal Binomial distribution for P (Wallis 2013), Equation (7') simplifies to Equation (3)[2].

$$s_{ss} = \sqrt{\sum nCr.P^r(1-P)^{(n-r)}(r/n-P)^2} \equiv \sqrt{P(1-P)/n} \,.$$

However, as we have already noted, to take into account the fact that the subsamples are observed rather than ideal, the number of degrees of freedom is not t', but $t' - 1$, and Equation (7') must be adjusted accordingly:

$$s_{ss}^2 = \frac{t'}{t'-1}\sum p(x_i)(p_i - p)^2 \,, \qquad (7)$$

where t' is the number of non-empty samples (i.e. where $n_i > 0$). Using Equation (7) in place of (4), we obtain s_{ss} = 0.0397 (Example 1) and 0.013856 (Example 2). In the examples we have seen, standard deviations increase by ~0.5%, which is negligible.

7. Example 3: Interrogative clause probability, all ICE-GB data

So far we have considered the effect of measuring the between-subsample variation for a variable within a single ICE-GB text category.

Next, we consider a situation of sampling across many text genres for a variable known to vary substantially across them.

ICE-GB contains both spoken and written English. The probability of selecting an interrogative clause varies widely between different text categories.

We can expect that question-clauses in written texts and monologues are likely to be much less frequent than in direct conversations. But if our method is to be robust and scalable, we need to consider research questions like these.

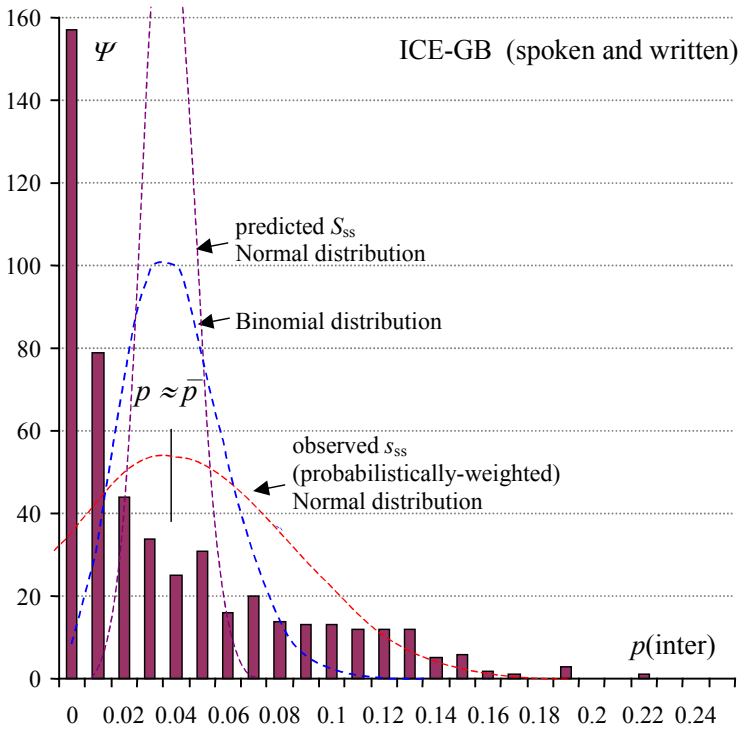

Figure 4: Frequency distribution for p(inter), all ICE-GB texts, with Gaussian curves for predicted and observed measures of standard deviation.

Using all 500 ICE-GB text categories we obtain the following:

observed probability $p = p(\text{CL}) = f(\text{CL})/f(\text{word}) = 0.0399$.
number of cases $n = f(\text{CL}) = 145{,}179$.
standard deviation $s = 0.000514$.
95% Wilson interval $(w^-, w^+) = (0.038908, 0.040922)$.

[2] A proof on an integer scale is found in Wackerly *et al* (2008:107). Equation (3) is the variance of the Binomial distribution on a probabilistic scale obtained by division by n, hence the difference $(r/n - P)$ rather than $(r - nP)$.

Employing the probabilistically-weighted computation using p obtains:

$S_{ss} = 0.008753,$
$s_{ss} = 0.044462,$
ratio $F_{ss} = S_{ss}^2/s_{ss}^2 = 0.0388.$

The variation of p between different texts is much larger than the variation expected by the random sampling assumption. See Figure 4. This result entails a reduction of n by approximately $1/26^{th}$. Following rescaling, we obtain

$p = p(\text{CL}) = 0.0399.$
number of cases $n' = 6{,}107,$
standard deviation $s = 0.002505,$
95% Wilson interval $(w^-, w^+) = (0.035276, 0.045107).$

This reduction in n amounts to multiplying the interval width by a factor of five. However, the initial interval was tiny, so the new interval is still small. If you compare the resulting sample size n with Example 1 (direct conversation data only), despite the much more diverse sample, the standard deviation of p has increased by approximately 50%.

Whereas a method based on the ratio of variances should be sound and robust, there remains one potential concern. Where p approaches 0 or 1, the Normal curve deviates from the Binomial (see upper curves in Figure 4). Similarly the lower dotted line in Figure 4 does not follow the distribution. Nonetheless, provided the distribution is approximately Binomial, Equation (7) should still obtain a meaningful estimate of variance.

This method of estimating variance merges each set of subsamples into a single set and then estimates the variance of this larger set. This works on the assumption that the average of multiple Binomial distributions tends to a Binomial distribution. Wallis (forthcoming 2020) has an extended discussion of what this means for corpus sampling as well as discussing solutions to this problem for small samples and free choices.

8. Example 4: Rate of transitive complement addition

As a final example, we will take a grammatical variable and apply it to the text categories in ICE-GB. A typical variable is illustrated by the two *Fuzzy Tree Fragments* (FTFs, Nelson *et al*. 2002) in Figure 5. Figure 5a retrieves the baseline set (cases of non-passive clauses), whereas Figure 5b retrieves the subset where the verb phrase is followed by a non-passive, infinitive transitive complement clause containing a subject. The variable in this case, p, may be expressed as the tendency to add such transitive complements to the upper, 'host', clause.

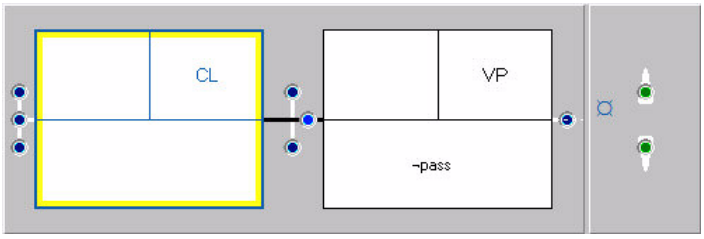

Figure 5a. Baseline FTF: a clause containing a verb phrase (VP) that is not passive.

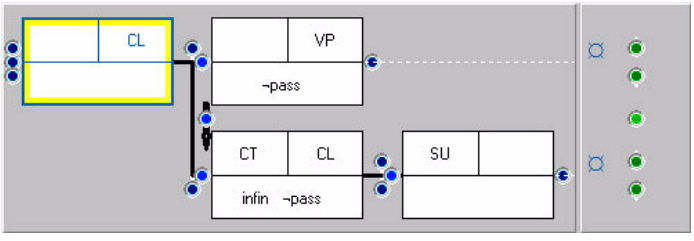

Figure 5b. Subset FTF: VP followed by an infinitive transitive complement clause containing a subject.

Figure 6 plots the distribution of p with Wilson intervals across ICE-GB genre categories. The thin 'I'-shaped error bars represent the conventional Wilson score interval for p, assuming random sampling. The thicker error bars represent the adjusted Wilson interval obtained by the

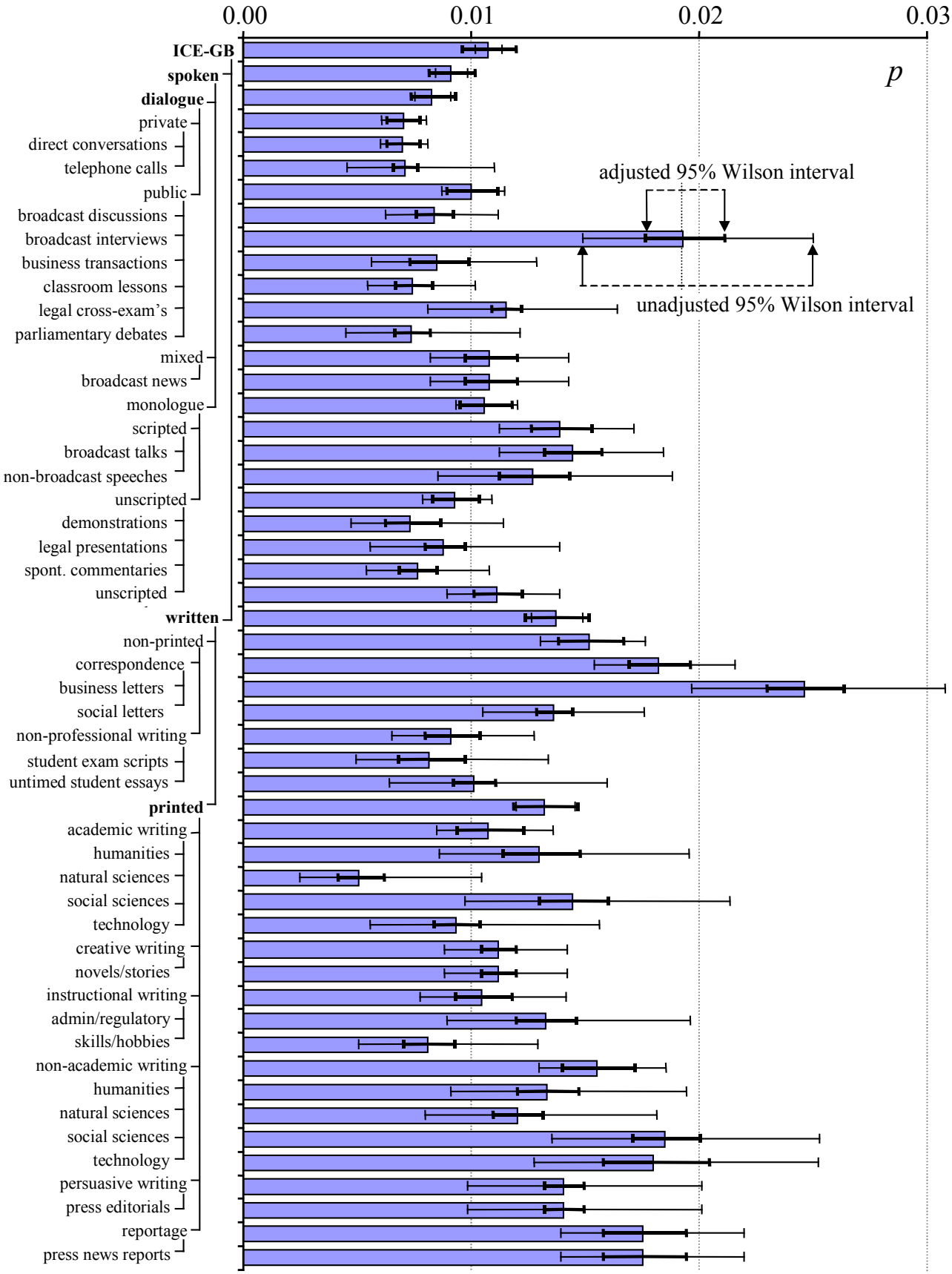

Figure 6: Probability of following a VP with a transitive complement clause, with Wilson score confidence intervals, before and after adjustment, over text categories of ICE-GB.

probabilistically-weighted method of Equation (7). These results are tabulated in Table 2.

The figure reinforces observations we made earlier. Within a single text type, such as *broadcast interviews*, p has a compressed range and cannot plausibly approach 1. (Note that mean p does not

	Texts	Rate	Size	Wilson		Ratio	Adjusted Wilson	
	t'	p	n	w^-	w^+	F_{ss}	w^-	w^+
ICE-GB	500	0.0107	115,846	0.0101	0.0113	0.2504	0.0096	0.0119
spoken	300	0.0091	74,973	0.0084	0.0098	0.4660	0.0081	0.0101
dialogue	180	0.0082	48,399	0.0075	0.0091	0.6732	0.0073	0.0093
private	100	0.0070	27,817	0.0061	0.0080	1.8354	0.0063	0.0077
direct conversations	90	0.0070	25,124	0.0060	0.0081	1.9261	0.0063	0.0077
telephone calls	10	0.0071	2,693	0.0045	0.0110	35.5575	0.0065	0.0076
public	80	0.0100	20,582	0.0087	0.0114	1.4248	0.0089	0.0112
broadcast discussions	20	0.0083	5,290	0.0062	0.0111	9.1723	0.0075	0.0092
broadcast interviews	10	0.0192	2,913	0.0148	0.0249	8.2502	0.0176	0.0210
business transactions	10	0.0085	2,595	0.0056	0.0128	7.4040	0.0073	0.0099
classroom lessons	20	0.0074	5,127	0.0054	0.0102	8.3633	0.0066	0.0083
legal cross-examinations	10	0.0115	2,613	0.0081	0.0163	37.6379	0.0108	0.0122
parliamentary debates	10	0.0073	2,044	0.0045	0.0121	22.5232	0.0066	0.0082
mixed – broadcast news	20	0.0108	4,554	0.0081	0.0142	6.6857	0.0097	0.0120
monologue	100	0.0105	22,020	0.0093	0.0120	1.4040	0.0095	0.0117
scripted	30	0.0138	6,216	0.0112	0.0171	4.8230	0.0126	0.0152
broadcast talks	20	0.0144	4,318	0.0112	0.0184	7.9494	0.0132	0.0157
non-broadcast speeches	10	0.0126	1,898	0.0085	0.0187	10.7933	0.0112	0.0143
unscripted	70	0.0092	15,804	0.0079	0.0109	2.1267	0.0083	0.0103
demonstrations	10	0.0073	2,604	0.0047	0.0114	7.4505	0.0062	0.0086
legal presentations	10	0.0088	2,055	0.0055	0.0138	21.8150	0.0079	0.0097
spontaneous commentaries	20	0.0076	4,205	0.0054	0.0107	10.2608	0.0068	0.0085
unscripted speeches	30	0.0111	6,940	0.0089	0.0138	5.2884	0.0101	0.0122
written	200	0.0137	40,873	0.0126	0.0148	0.6445	0.0123	0.0151
non-printed	50	0.0151	10,857	0.0130	0.0176	2.5855	0.0137	0.0166
correspondence	30	0.0181	7,220	0.0153	0.0215	5.2043	0.0168	0.0195
business letters	15	0.0245	3,018	0.0196	0.0307	11.0307	0.0229	0.0262
social letters	15	0.0136	4,202	0.0105	0.0175	19.6104	0.0128	0.0144
non-professional writing	20	0.0091	3,637	0.0065	0.0127	6.7799	0.0080	0.0103
student examination scripts	10	0.0081	1,851	0.0049	0.0133	8.1914	0.0068	0.0097
untimed student essays	10	0.0101	1,786	0.0064	0.0159	25.8299	0.0092	0.0110
printed	150	0.0131	30,016	0.0119	0.0145	0.8603	0.0118	0.0146
academic writing	40	0.0107	6,259	0.0084	0.0136	3.1691	0.0094	0.0122
humanities	10	0.0129	1,702	0.0086	0.0195	10.1561	0.0113	0.0147
natural sciences	10	0.0051	1,382	0.0025	0.0104	14.1730	0.0042	0.0062
social sciences	10	0.0144	1,670	0.0097	0.0213	14.7718	0.0130	0.0159
technology	10	0.0093	1,505	0.0055	0.0156	24.3490	0.0084	0.0103
creative writing – novels/stories	20	0.0111	5,926	0.0088	0.0141	11.8186	0.0104	0.0119
instructional writing	20	0.0104	3,929	0.0077	0.0141	6.7585	0.0093	0.0117
administrative/regulatory	10	0.0132	1,820	0.0089	0.0195	16.2057	0.0119	0.0146
skills/hobbies	10	0.0081	2,109	0.0050	0.0129	11.8617	0.0070	0.0092
non-academic writing	40	0.0154	7,646	0.0129	0.0184	3.0017	0.0139	0.0171
humanities	10	0.0133	1,960	0.0091	0.0194	14.0089	0.0120	0.0147
natural sciences	10	0.0120	1,837	0.0079	0.0181	21.2347	0.0109	0.0131
social sciences	10	0.0184	2,061	0.0135	0.0252	15.1523	0.0170	0.0200
technology	10	0.0179	1,788	0.0127	0.0252	6.9958	0.0157	0.0204
persuasive – press editorials	10	0.0140	2,070	0.0098	0.0200	33.7119	0.0132	0.0149
reportage – press news reports	20	0.0174	4,186	0.0139	0.0219	4.8168	0.0157	0.0193

Table 2: Hierarchical table of results for following a verb phrase with a transitive complement, with Wilson intervals before and after adjustment, adjustment ratio F_{ss} and other variables. Shaded cells indicate contraction of intervals.

exceed 0.03 in any genre.) The observed between-text distribution is smaller than that predicted by Equation (3), and, armed with this information, we are able to reduce the 95% Wilson score interval for p. This degree of compression (i.e., the plausible value of max(p)) may also differ by text genre.

However, the reduction due to range-compression is offset by a countervailing tendency: pooling genres increases the variance of p. The distribution of texts across the entire corpus consists of the sum of the spoken and written distributions (p = 0.0091, 0.0137 respectively), and so on.

The Wilson interval for the mean p averaged over all of ICE-GB approximately doubles in width (F_{ss} = 0.2509), and the intervals for *spoken*, *dialogue*, *written* and *printed* (marked in bold in Figure 6) also expand, albeit to lesser extents. Other intervals contract (F_{ss} > 1, shaded), tending to generate a more consistent set of intervals over all text categories.

9. Conclusions

In this paper we derived and demonstrated a mathematically defensible method for adjusting the estimated variance of a sample composed of contiguous subsamples which is highly generalisable to a wide range of approaches. Where we find that subsamples are more widely spread than would be expected by chance random sampling, our approach simply reduces the weight of evidence, represented by sample size n, supporting an observation p.

The power of this approach lies in its integration with other robust methods. It can be combined with continuity corrections, and employed in logistic regression, model-fitting and meta-testing. One might even conceivably apply it in conjunction with a finite population correction.

As a by-product of the procedure of examining actual distributions of p, we also compensate for another issue, namely that many 'probabilistic' variables conventionally measured by linguists are not plausibly able to reach 100% saturation (Wallis 2012c). Without this compensation, standard methods can vastly overestimate the range of values for p: in our final example we found confidence intervals that were more than 5 times oversized. Correcting for this problem, which is rarely discussed in the literature, increases the statistical sensitivity of tests substantially.

This paper is based on sample-wide estimation. Unlike the case interaction approach, it is not built up from individual instances. We rely on less information than may be obtained by such models. On the other hand, we rely on more information than a Mann-Whitney approach. Instead of rank positions, we use the actual scores per text, and we can make claims about our observation p. We extended our method to subsamples of uneven sizes, which is not trivial to address with a rank order statistic. Our approach does not dispense with the Binomial model, but adjusts it, so it suffers no loss of generality. Thus, for example, it is a simple matter to compare any pair of adjusted Wilson intervals in Figure 6 with Newcombe's method (Wallis 2013).

Our approach is likely to be more conservative ("weaker", or a less powerful test) than an optimal case interaction model, and is also less conservative ("stronger") than the Mann-Whitney U test.

One direction of future and ongoing work remains the development of case interaction models. As we noted earlier, these models have a number of potential desirable characteristics for linguistic study in their own right. Our method is premised on observing the distributional consequences of many interactions, rather than examining each interaction in turn. To draw a medical analogy, it is the difference between plotting the epidemiological consequences of contagion rather than modelling the process of contagion itself. The method described in the present paper can clearly be employed to parameterise such case interaction models. It gives us an order of magnitude estimate of the maximum extent of the sum of between-case interdependence measures.

This paper is an initial sketch of the approach. The work in this paper is developed considerably further in Wallis (forthcoming 2020), and it is to that book that the reader is directed.

References

Brezina, V. and M. Meyerhoff, 2014. Significant or random?: A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19:1, 1-28.

Sheskin, D.J. 1997. *Handbook of Parametric and Nonparametric Statistical Procedures*. 1st Edition. Boca Raton, Fl: CRC Press.

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

Wackerly, D.D., Mendenhall, W. and Scheaffer, R.L. 2008. *Mathematical Statistics with Applications*. 7th Ed. Belmont, USA: Brooks/Cole.
Wallis, S.A. and B. Aarts, 2007. *Final Report to EPSRC: Next Generation Tools for Linguistic Research in Grammatical Treebanks*. London: Survey of English Usage. Published online at http://www.ucl.ac.uk/english-usage/projects/next-gen/report.htm
Wallis, S.A. 2012a. Random sampling, corpora and case interaction, *corp.ling.stats*. London: Survey of English Usage. http://corplingstats.wordpress.com/2012/04/15/case.
Wallis, S.A. 2012b. Reciprocating the Wilson interval, *corp.ling.stats*. London: Survey of English Usage. http://corplingstats.wordpress.com/2012/11/25/recip.
Wallis, S.A. 2012c. Freedom to vary and significance tests, *corp.ling.stats*. London: Survey of English Usage. http://corplingstats.wordpress.com/2012/09/30/free.
Wallis, S.A. 2012d. Inferential statistics – and other animals, *corp.ling.stats*. London: Survey of English Usage. http://corplingstats.wordpress.com/2012/04/30/inferential.
Wallis, S.A. 2013. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics* 20:3, 178-208.
Wallis, S.A. forthcoming 2020. *Statistics in Corpus Linguistics Research*. New York and London: Routledge.