Background

1. Confidence intervals were developed for NSS results in 2007. This paper provides further information on how the confidence intervals are calculated.

2. Professor Harvey Goldstein was asked to advise on how confidence intervals could be calculated which would enable them to be shown on the NSS results. His paper to HEFCE is given in Annex A.

3. For illustration of the size on these intervals on actual data, Annex B shows the Wilson confidence intervals for the institutional percentage who agree in NSS 2005 for English-based studies, computer science, and arts and design.

ANNEX A: Constructing confidence intervals for proportions

Single confidence intervals

1. For one proportion

The standard Normal confidence interval (CI) approximation is given by

$$\hat{p} \pm z_{\alpha}\sqrt{\hat{p}(1-\hat{p})/n}$$

where $\hat{p} = x/n$ is the observed proportion, *n* is the sample size and z_{α} is the standard Normal deviate corresponding to the two sided α percentage point.

Recent literature suggests, however, that this approximation, known as a Wald interval, can be very poor where the underlying proportion is close to 1 or 0 and/or when *n* is not large (see e.g. Brown et al., 2001). In particular the coverage probability (the proportion of times the interval actually includes the true value) can vary quite erratically as *n* changes, even for values of *n* as high as 100. Note that all intervals for a proportion, especially for small *n,* will be approximate due to the discrete nature of the binomial distribution.

An interval that has good properties for sample sizes down to about 40, and good but slightly conservative (i.e. a tendency to provide a longer interval) properties down to 30, is the so called Agresti-Coull interval (Brown et al., 2001). Its advantage is that it is very simple to compute and can be regarded as an adjusted Wald interval, and its conservative nature below 40 seems acceptable given the increasing likelihood of biases arising with smaller sample sizes. If a shorter interval with equivalent coverage probablilities is required then the 'Wilson interval' (Newcombe, 1998a) can be used and both formulae are given below.

Define the modified proportion and sample size

$$\tilde{p} = \frac{x + z_{\alpha}^{2}/2}{\tilde{n}}, \quad \tilde{n} = n + z_{\alpha}^{2}, \qquad (1)$$

The Agresti-Coull interval is given by

$$\tilde{p} \pm z_{\alpha}\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}} \qquad (2)$$

and the Wilson interval is given by

$$\tilde{p} \pm \frac{z_{\alpha}}{\tilde{n}}\sqrt{(n\hat{p}(1-\hat{p}) + z_{\alpha}^{2}/4)} \qquad (3)$$

2. For the difference between two proportions

As in the case of a single proportion, the standard 'Wald' interval given by

$$(\hat{p}_1 - \hat{p}_2) \pm z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

performs badly. Analogously to the single proportion we can construct a 'Wilson' CI that has good coverage down to a sample size of 30 (Newcombe, 1998b). We compute the quantities (l_1, u_1) which are respectively the smallest and largest values of

$$\frac{n_1(2x_1 + z_\alpha^2) \pm z_\alpha \sqrt{n_1^2 z_\alpha^2 + 4x_1 n_1^2 - 4x_1^2 n_1}}{2(n_1^2 + n_1 z_\alpha^2)}$$

and (l_2, u_2) which are respectively the smallest and largest values of

$$\frac{n_2(2x_2 + z_\alpha^2) \pm z_\alpha \sqrt{n_2^2 z_\alpha^2 + 4x_2 n_2^2 - 4x_2^2 n_2}}{2(n_2^2 + n_2 z_\alpha^2)}$$

And form the interval ($\hat{p}_1 \geq \hat{p}_2$)

$$\left\{(\hat{p}_1 - \hat{p}_2) - z_\alpha \sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}}, (\hat{p}_1 - \hat{p}_2) + z_\alpha \sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}}\right\} \quad (4)$$

Multiple comparisons

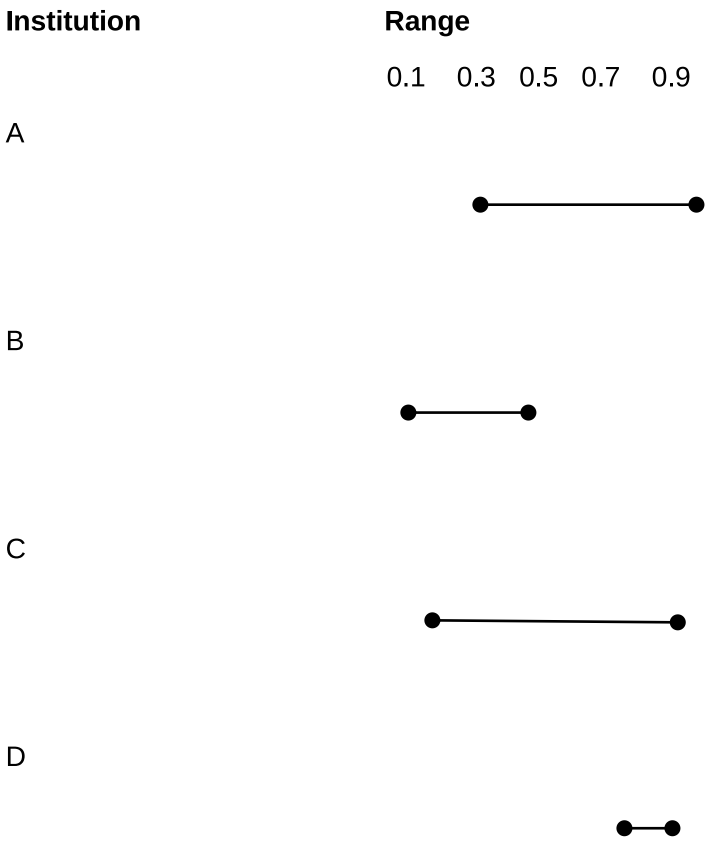
To avoid too much data snooping, and if we assume that someone will carry out *m* comparisons of a given type at any one time then, for a 95% joint interval, we should construct a 'Bonferroni' interval corresponding to the tail area α / m interval, where $\alpha = 0.05$ here. Thus, if *m*=5 we would use the equivalent of a 99% interval. The value of *m* could simply be the number of individual proportions requested by a user. This, however, deals only with the case of comparing each proportion with a single 'null' value – say the population mean proportion.

The Bonferroni interval can become very wide as *m* increases. For this reason many people prefer to use a criterion that controls the rate of *false* rejections (FDR) of a hypothesis out of all rejections when several tests are conducted in an experiment. This operates in terms of the type 1 error rate (α) and can be applied to a set of confidence intervals (Benjamini and Yekutieli, 2001). For a set of *m* independent intervals an approximate adjusted error rate for the separate intervals is given by $\alpha^* = \alpha(m+1)/2m$. A more conservative procedure that allows for the (expected) negative dependencies among the test statistics is given by $\alpha^* = \dfrac{\alpha(m+1)}{2m(\log_e(m) + 0.6)}$.

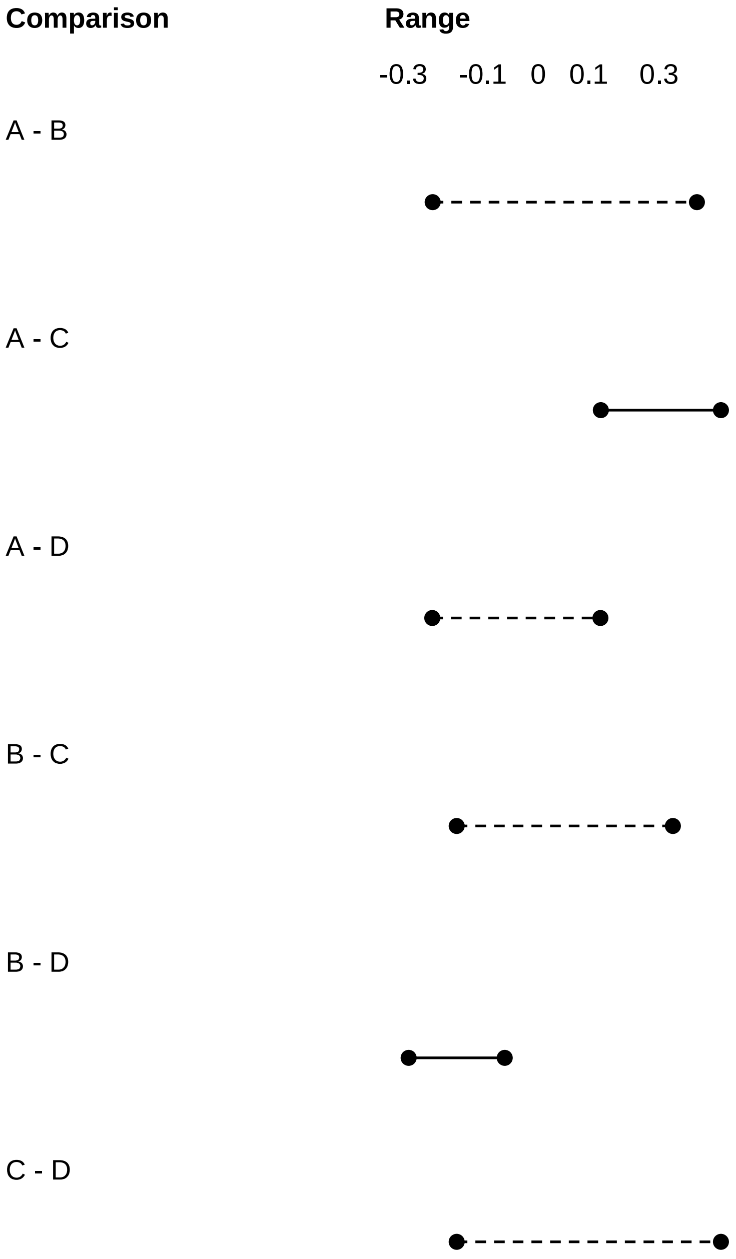
More precise adjustments are available but involve extra, on the fly, computation. Thus, for *m*=10 in the case of pairwise comparisons $\alpha^* = 0.028$ or 0.01 for the more conservative procedure. Since we will always have m<=5 it is proposed that we use the conservative FDR for the difference of proportions and the non-conservative FDR for the single proportions assuming *m*=5. (If it is simple to tailor to the number selected then we can use the actual value of *m* chosen.) That is for differences we construct our intervals In the case of scores that (possibly after transformation) have a Normal distribution we will apply a similar procedure, but using the standard formulae for a Normal confidence interval. For multiple comparisons we construct intervals using values of z_α derived as above, namely with a nominal value of 0.01 for a difference between mean scores and for a single score using $\alpha * = 0.03$. .

Presentation

One possibility for presentation is to divide the presentation into two parts – possibly side by side on the screen. For the single proportion we might have something like the following:



For comparing two proportions we might have something as follows:



Here we list all comparisons and give the intervals, with dashed lines where they overlap zero. The presentations would need to have associated textual explanations that need to be written.

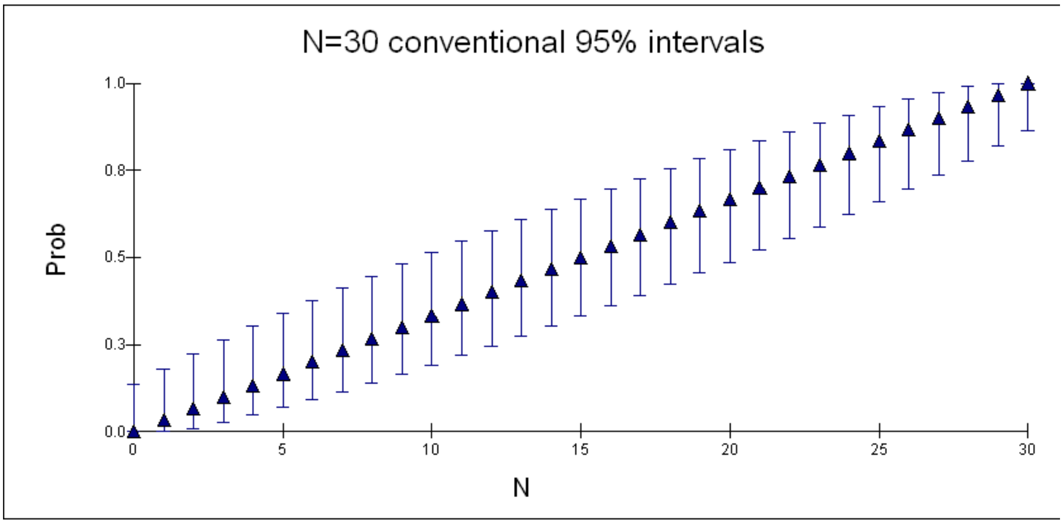
For comparing proportions we can construct a 4-way look-up table which gives the values of *n, p* for each pair to be compared with associated interval computed using the above formulae. Likewise we have a 2-way look up table for single proportions. We can use single value increments for *n* and for *p* increments of 0.02 should be adequate. It would be preferable to carry out the relatively straightforward calculations on the web site as a user requests a set of values.

Modelling

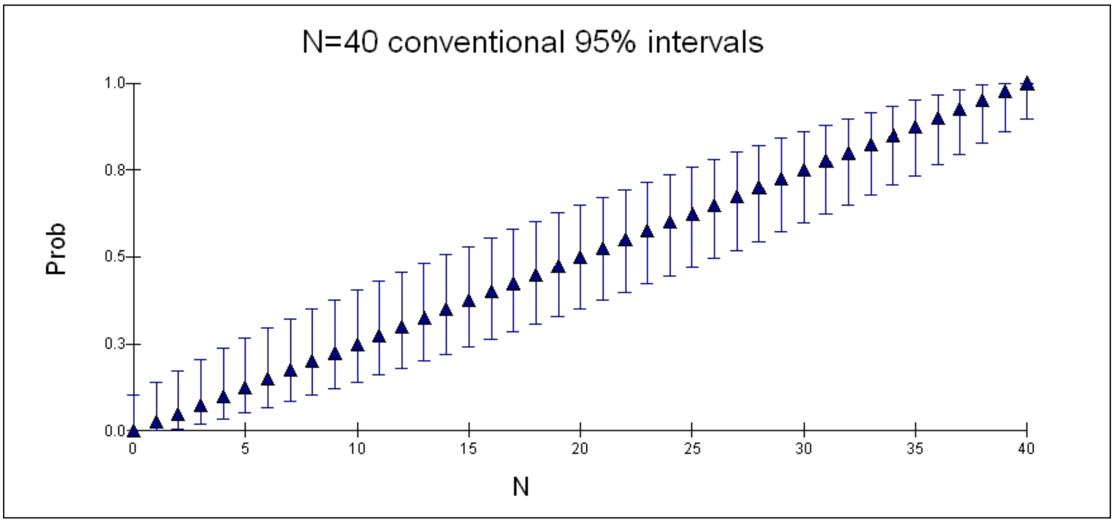
In all of the above it is assumed that we are using observed proportions or means – the latter possibly with some suitably pooled estimate of variance. At some stage, however, it may be worth considering a fully model based procedure whereby the proportions and means and associated intervals are model based estimates. Since the numbers are small, these should probably be based upon MCMC estimates rather than large sample ones. Each interval still needs to be adjusted for multiple comparisons. To carry out such a model based procedure some web based computing would be needed, not to fit the models but to estimate the residuals using model parameter estimates.

Examples

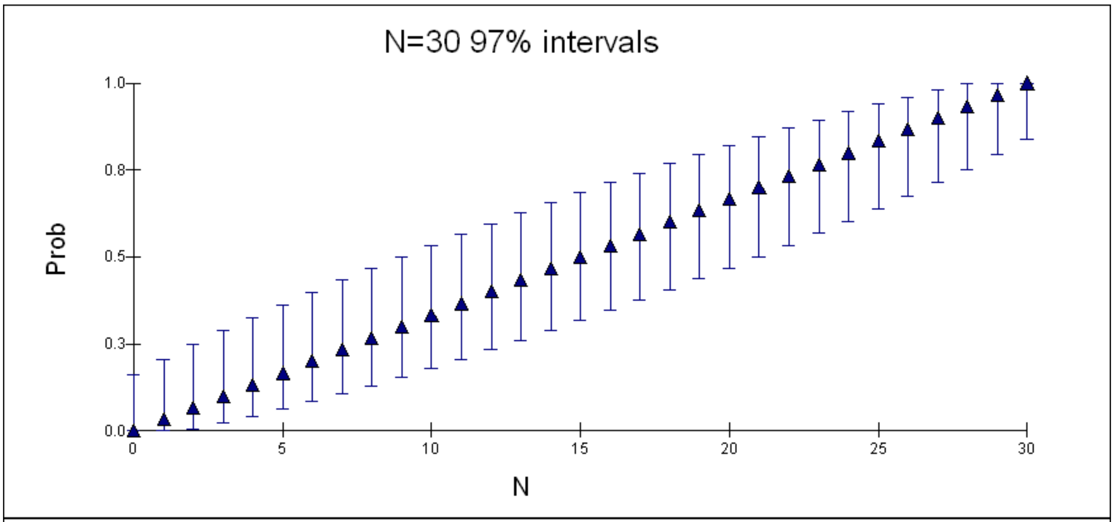
The first example is for conventional 95% intervals for N=30

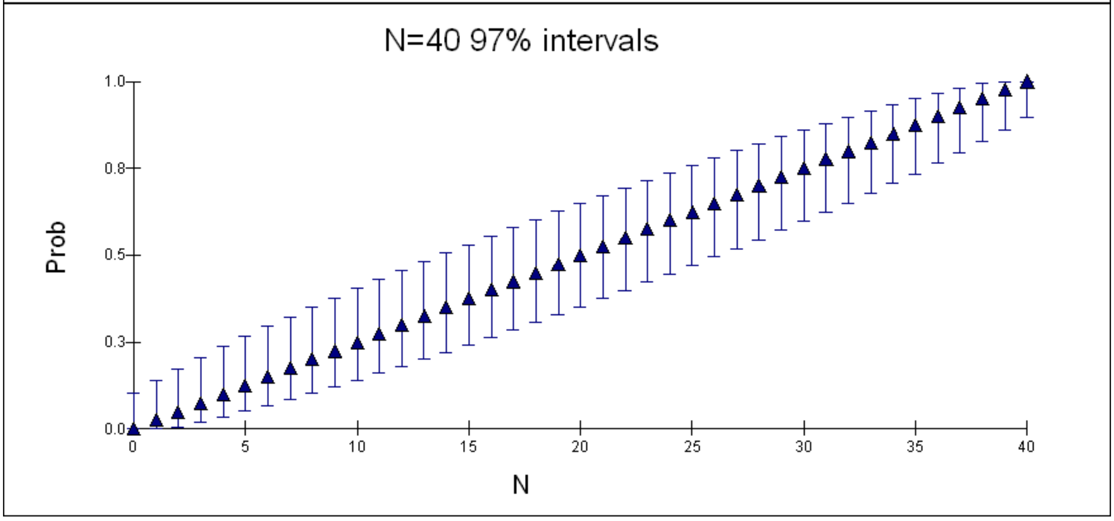


And for N=40



Now using Z=2.17 (0.03 two sided tail area) we obtain





For a pair of proportions we could have a large lookup table, unless it was feasible to carry out 'on the fly' computations which would be preferable. To illustrate, using a tail area of 0.01 we have:

1. Assuming both pairs have denominator N=30 and one numerator is 15 (p=0.5) then the range of values for the other such that the CI includes 0 is 6 – 24 (0.2 – 0.8). For both N=40 with numerator 20, it is 9 – 31 (0.23 – 0.78).
2. For both N=30 with one numerator 23 (p=0.77) then the CI range that includes zero is 14 – 29 (0.47 – 0.97)
   and for both N=40 and one numerator 31 (0.78) is 20 – 39 (0.50 – 0.98).

There is not a great deal of difference between N of 30 and 40 but in both cases we do have wide intervals.

Harvey Goldstein
29/03/07

References

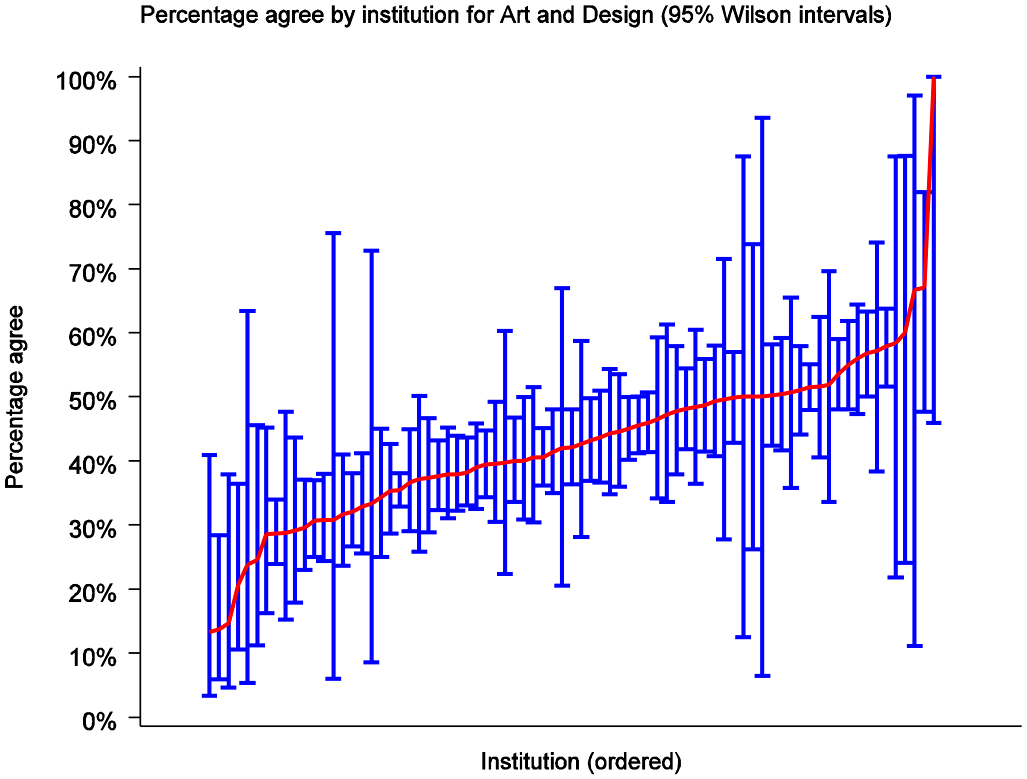
Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of Statistics, 29, 1165-1188

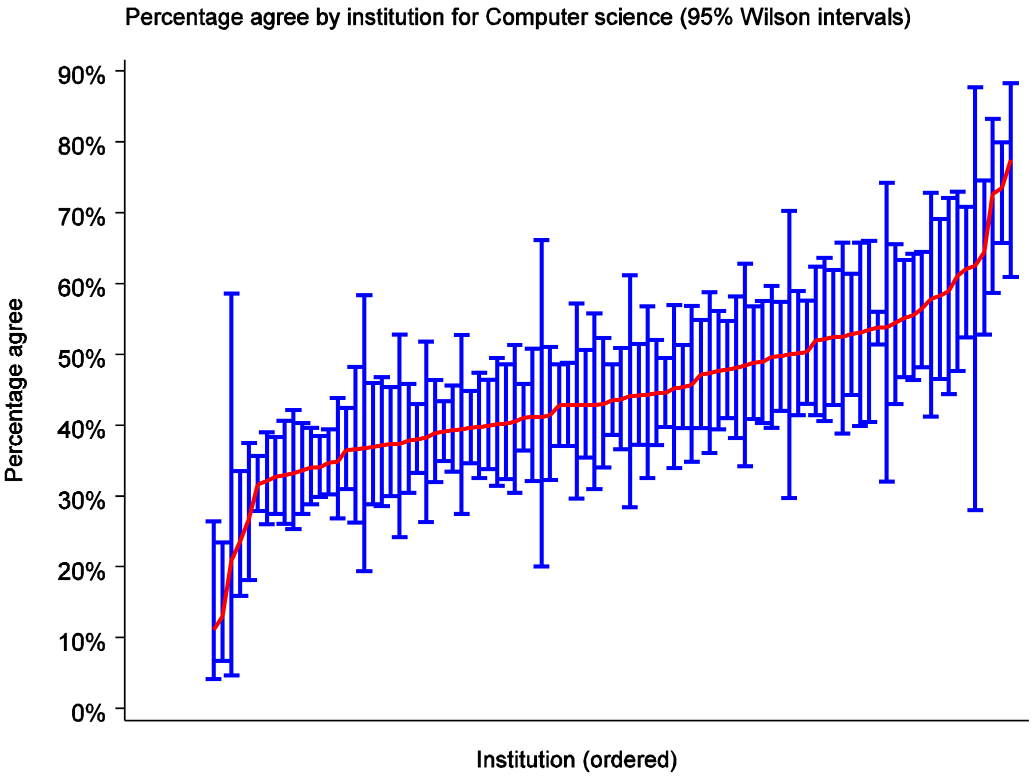
Brown, L. D., Cai, T., and DasGupta. A. (2001). Interval estimation for a binomial proportion. Statistical Science, 16, 101-117.

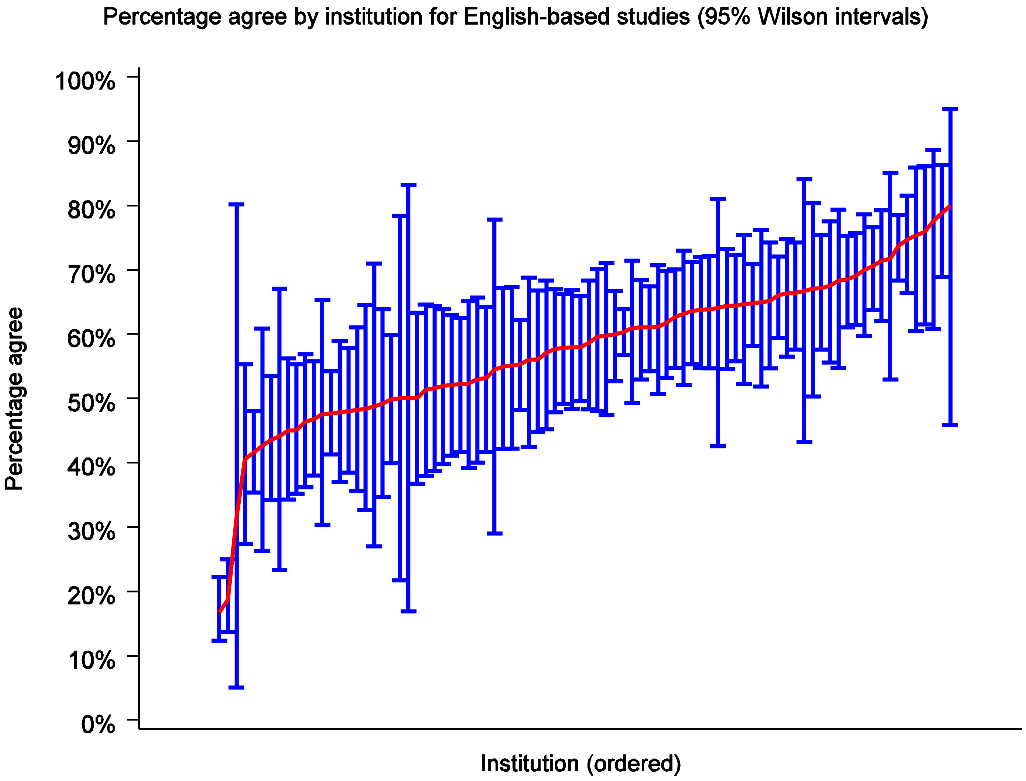
Newcombe, R. (1998a). Two sided confidence intervals for the single proportion: comparison of seven methods. Statistics in Medicine, 17, 857-872.

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

Annex B: Illustration of Wilson confidence intervals for percentage agree






Percentage agree by institution for English-based studies (95% Wilson intervals)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage agree
Institution (ordered)