Article:

Laud, P.J. orcid.org/0000-0002-3766-7090 (2018) Comments on "One-tailed asymptotic inferences for the difference of proportions: analysis of 97 methods of inference" by Álvarez Hernández M, Martín Andrés A and Herranz Tejedor I. (2018). Journal of Biopharmaceutical Statistics, 28 (5). pp. 1018-1020. ISSN 1054-3406

https://doi.org/10.1080/10543406.2018.1489411

Comments on “One-tailed asymptotic inferences for the difference of proportions: analysis of 97 methods of inference” by Álvarez Hernández M, Martín Andrés A and Herranz Tejedor I. Journal of Biopharmaceutical Statistics (published online 02 Apr 2018)

Peter J. Laud

In their recent paper, Álvarez Hernández and colleagues present an evaluation of numerous methods for analysing the difference between two independent binomial proportions for one-tailed inference. I would like to offer a graphical presentation giving some insight into the criteria used for selecting the optimal method, and recommend a superior method that was omitted from the evaluation.

The parameter α^* (real error) is described as “traditional”. This parameter (or equivalently coverage probability for a confidence interval) is more often defined in the literature at each individual parameter space point (p_1, p_2), but the authors have defined it as a local maximum over the range of p_1 and p_2 for a fixed δ. Consequently the summary statistic $\overline{\Delta\alpha}$, which is based on the average of these local maxima over different values of δ, n_1 and n_2, is open to misinterpretation. This is further confused when evaluating the $\delta = 0$ case, when $\overline{\Delta\alpha}$ uses only a single value of α^* for each n_1, n_2 pair.

As described by Newcombe (1998), there is a fundamental choice to be made in selecting the best methods in the field of discrete data analysis, since the nominal significance level cannot be attained exactly. For a given nominal significance level (such as 5%), one may seek to obtain error rates that are either guaranteed to be *at most* 5% (the conservative criterion) or that are 5% *on average* (the proximate criterion). Naturally, there is a range of possibilities between these two extremes, and this is where the definition of "optimal" becomes problematic. For example, when applying a proximate criterion, it is sometimes preferred that average error rates are slightly below 5%, and this seems to be the position the authors are aiming for. However, in practice, the $\overline{\Delta\alpha}$ parameter results in a selected method that is further towards the conservative end of the scale than implied in the paper. When it comes to evaluating $\delta = 0$, the selection criterion almost becomes a strictly conservative one, resulting in the unexpected conclusion that the classic ZE0 method is "too liberal".

Furthermore, in the case of one-tailed errors in particular, it is not uncommon to observe a type of bias in error rates with systematic deviations from the nominal level that cancel each other out when averaged. Therefore $\overline{\Delta\alpha}$ is not a reliable indicator of optimal one-tailed coverage, regardless of whether the average is taken over local maxima or the whole parameter space. Cai (2005) discussed this bias for the one-sample case, without reference to any summary of mean coverage probabilities. The ideal for a one-tailed method is to have consistent error rates that are close to the target rate (whether in a proximate or conservative way) over as much of the parameter space as possible, not just on average. A confidence interval that achieves this is described as having central location (Newcombe, 2011), and arguably this property should form part of the selection criteria for both one- and two-tailed inference. For risk difference, this may be examined by separately averaging the error rates above and below the $p_1 = p_2$ diagonal, to assess equality of "mesial" and "distal" errors respectively. Alternatively, for methods that achieve proximate coverage, it is sufficient to summarise the percentage of all parameter space points that have an error rate within a set distance (e.g. ±0.1×) above or below the nominal error rate (% proximate).

Newcombe and Nurminen (2011) considered that occasional substantial dips in coverage (i.e. “failures”) need not be viewed with disquiet, and proposed evaluating performance on the basis of a moving average representation of actual error rates to smooth out the unavoidable spikes in coverage. This approach was extended to the two sample case in Laud and Dane (2014). The resulting smoothed probability surface plots provide a visual means of evaluating confidence interval performance, and they also help to illustrate the problems noted above regarding $\overline{\Delta\alpha}$.

The attached figure displays the unsmoothed and smoothed one-sided error rates (RNCP = right-sided non-coverage probability) for $n_1 = n_2 = 60$ with nominal one-sided $\alpha = 5\%$. Parameter space points with a “too conservative” real error below 4.5% are coloured in shades of yellow, fading to white at 0%, and “too liberal” error rates of over 5.5% appear as shades of red to black (note that this is much more strict than the 7% used in the definition of failures). Two shades of orange represent the desirable “proximate” range of error rates within ±0.5% of the target 5% error rate.

Blue circles are superimposed on the plots showing the local maxima at selected δs that form the basis of $\overline{\Delta\alpha}$. These reveal the clear pattern of change in error rates for AE5 across the range of δ (which can also be observed to some extent in Table 2 of the paper). It can also be seen that in places the performance of all three selected methods can be quite conservative, and alarmingly so for the ZP’0 method. It may also be noted that the continuity correction for AE5 has almost no effect, and does not achieve the conservative selection criterion, which is commonly the purpose of such corrections. This suggests that the authors' proposed continuity correction may have a different intended purpose.

The authors also analysed the hybrid "MNBL" method, proposed in Laud and Dane (2014). They have overlooked the superior method labelled in the same paper as "GNbc", which has subsequently been developed further as the skewness-corrected asymptotic score ("SCAS") method (Laud, 2017). Error rate plots for SCAS are shown in the figure alongside those for AE5 and ZP'0, clearly demonstrating the superior performance of this method. The incidence of failures (α^*>0.07) across the whole parameter space is comparable to the other methods at 0.2%. Similar patterns are observed in the surface plots for other combinations of n_1, n_2 and α. Further evaluations also confirm that error rates for SCAS are equally consistent for two-tailed inference.

Average power ($\bar{\theta}$) has not been extensively evaluated for SCAS, which is designed to prioritize the other aspects of performance in relation to error rates and interval location. However, for the example shown here, the calculated value of $\bar{\theta}$ for SCAS is slightly higher than the AE5 methods. According to my calculations, SCAS also has fewer failures than AE5 at the selected values of δ, so I believe it might have been selected as optimal if it had been included in the paper, despite the flaws in the selection criteria.

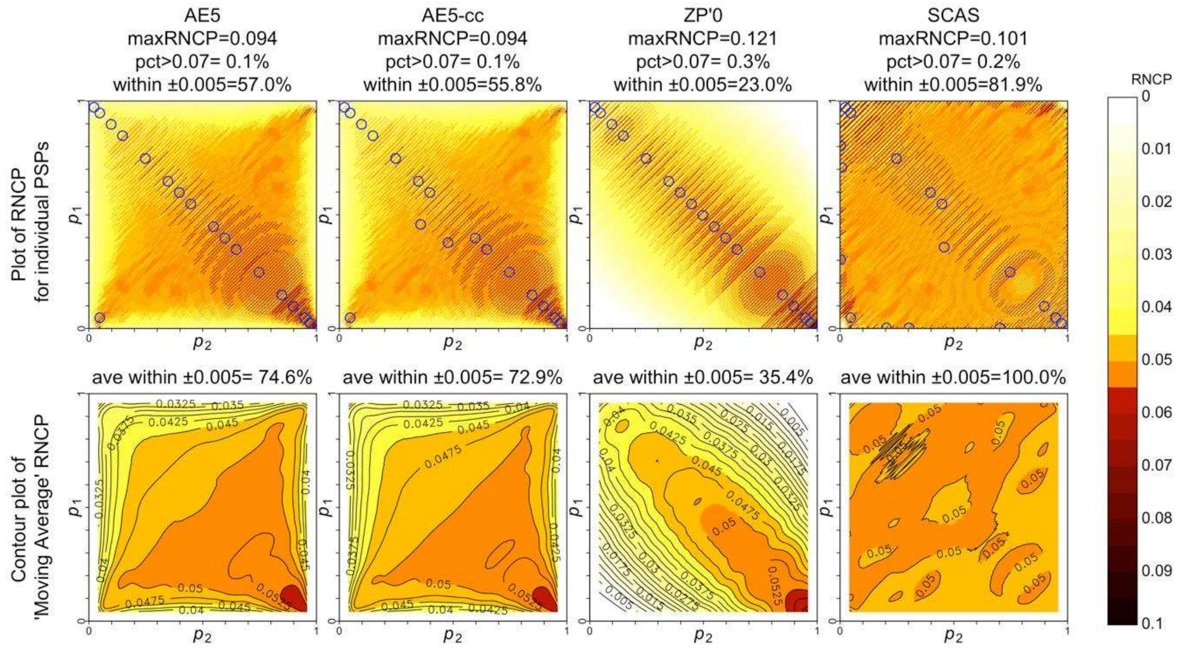
As well as superior error rates, the SCAS method has a number of other advantages:

1) No adjustments to the sample data are necessary;
2) The method is extended to stratified datasets;

3) The same underlying methodology is applicable to analysis of relative risk and odds ratio, and also to analysis of Poisson exposure-adjusted incidence rates, and even to the single sample case;

4) A versatile continuity correction may be used to obtain either strictly conservative, mostly conservative, or even conservatively proximate coverage, while maintaining central interval location (see Laud 2017 Supplementary Appendix).

5) Finally, SCAS is applicable to smaller sample sizes without too much reduction in performance. The authors of the above paper declare that “all methods work badly” with $n_1 < 60$. It is true that smaller sample sizes result in increased fluctuations in the probability surface leading to a slightly higher failure rate, but SCAS nevertheless achieves remarkably consistent proximate coverage at all sample sizes, when assessed on the basis of moving average error rates. This is helped by the n/(n-1) variance bias correction (Miettinen and Nurminen, 1985) that Álvarez Hernández *et al*. omit from the ZE0 method, presumably for the sake of consistency with the classic Pearson chi-squared test.



Surface plots of [top] right-sided non-coverage probability (RNCP) and [bottom] moving average RNCP for AE5, AE5-cc, ZP’0 and SCAS, with $n_1 = n_2 = 60$ and one-sided $\alpha = 5\%$.

REFERENCES

Álvarez Hernández M., Martín Andrés A., Herranz Tejedor I. (2018). One-tailed asymptotic inferences for the difference of proportions: analysis of 97 methods of inference. *Journal of Biopharmaceutical Statistics* (published online 02 Apr 2018).

Cai T. T. (2005). One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference* 131(1), 63-88.

Laud P. J., Dane A. (2014). Confidence intervals for the difference between independent binomial proportions: comparison using a graphical approach and moving averages. *Pharmaceutical Statistics* 13(5), 294-308.

Laud P. J. (2017). Equal-tailed confidence intervals for comparison of rates. *Pharmaceutical Statistics* 16(5), 334-348.

Miettinen O., Nurminen M. (1985). Comparative analysis of two rates. *Statistics in Medicine* 4(2), 213-226.

Newcombe R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 17(8), 857-872.

Newcombe R. G. (2011). Measures of location for confidence intervals for proportions. *Communications in Statistics – Theory and Methods* 40(10), 1743-1767.

Newcombe R. G., Nurminen M. M. (2011). In defence of score intervals for proportions and their differences. *Communications in Statistics – Theory and Methods* 40(7), 1271-1282.