In Defence of Score Intervals for Proportions and their Differences

Robert G. Newcombe\textsuperscript{a}; Markku M. Nurminen\textsuperscript{b}

\textsuperscript{a} Department of Primary Care & Public Health, Cardiff University, Cardiff, United Kingdom

\textsuperscript{b} Department of Public Health, Hjelt Institute, University of Helsinki, Helsinki, Finland

Online publication date: 08 February 2011
IN DEFENCE OF LIKELIHOOD-BASED INTERVALS FOR PROPORTIONS AND THEIR DIFFERENCES*

ROBERT G. NEWCOMBE¹, MARKKU M. NURMINEN²

¹ Cardiff University, United Kingdom
² University of Helsinki, Finland

*Abstract for the Nordic-Baltic Biometric Conference, Turku, Finland, 5-6 June 2011.

Miettinen and Nurminen (1985) developed interval estimation methods for a comparative analysis of two binomial proportions for application in epidemiology and medical research. The MN method, as it has become known in the biometric literature, has been theoretically and empirically demonstrated to possess desirable statistical properties. A recent article by Santner et al (2007) asserted that the MN likelihood-based interval for a difference of two unpaired proportions may have inadequate coverage.

We re-visit the properties of likelihood-based intervals for proportions and their differences. Published results indicate that these methods produce a mean coverage probability slightly above the nominal confidence level. We argue (Newcombe and Nurminen, 2011) that it is appropriate to align the mean rather than the minimum coverage with the nominal one, based on a moving average over an interval of parameter values which represents a reasonable expression of prior uncertainty. We will show under various simulation scenarios that the poor coverage properties claimed by Santner et al. actually relate to an ostensibly similar but in fact inferior version of the method (Mee, 1984).

In conclusion, likelihood-based intervals, including the MN method, have gained a well-deserved place as appropriate tools for useful application in epidemiologic and medical research practice. We call for a wider availability of these procedures in statistical software, including StatXact (2007), based on a correct test statistic that employs the unbiased estimator of variance in the MN interval method.

References
Introduction

This communication is a response to an article authored by Santner et al. (2007) who criticized the score method for calculating confidence intervals for differences between two independent binomial proportions, originally developed by Miettinen and Nurminen (MN) (1985).

Objectives

(1) It is necessary to distinguish between different versions of score intervals when discussing their coverage probability, with special respect to Santner et al.’s assertion that the MN intervals may have inadequate coverage.

(2) In the evaluation of the performance of the various methods it is more appropriate to consider the moving average of coverage probabilities instead of the minimum of coverage probabilities in view of the oscillating pattern typically occurring for binomial confidence intervals.
Asymptotic and Exact Approaches to the Construction of Interval Estimates for Binomial Proportions

A general procedure for asymptotic interval estimation described by Cox and Hinkley (1974), the score method, is based on a test statistic which contrasts the parameter of interest $\theta$ — to its empirical value $\hat{\theta}$, usually taken as a maximum likelihood estimate, and which has the variance of $\hat{\theta}$ as its denominator. These interval estimates are obtained by inverting a chi-squared test with 1 degree of freedom corresponding to a large-sample Gaussian distribution rendered free of any nuisance parameters.

Evaluative studies commonly report both mean and minimum coverage probability. We argue in this article that it is appropriate to align the mean coverage with the nominal level, $1-\alpha$. Confidence interval methods such as the exact method (Clopper & Pearson, 1934) that align the minimum coverage with $1-\alpha$ may be regarded as leading to unnecessarily wide intervals, or conversely, as requiring unnecessarily large sample sizes in order to attain a desired degree of precision, as shown below.
Score Intervals for the Difference between Independent Proportions

Miettinen and Nurminen (1985) derived score intervals for three effect size measures commonly used to compare two independent binomial proportions, their difference, ratio, and the odds ratio. This study (for elaboration, see Nurminen, 1984; Miettinen, 1985), and several subsequent ones (Nurminen, 1986, 1988, 1990; Nurminen and Miettinen, 1990; Nurminen and Nurminen, 1990; Newcombe, 1998b; Liu et al, 2006) confirmed that this approach produces average coverage probabilities close to 1-\(\alpha\).

We show that the poor coverage properties claimed by Santner et al actually relate to an ostensibly similar but in fact inferior version of the score interval proposed by Mee (1984). The crucial difference between the Mee and Miettinen-Nurminen (MN) formulations is that in the latter the variance incorporates a correction factor \((n_1 + n_2) / (n_1 + n_2 - 1)\), where \(n_1\) and \(n_2\) denote the two sample sizes. This seemingly small difference has a substantial impact on coverage. In a matched-pairs design, with \(n_1 = n_2 = 1\), the factor equals 2.
Santner et al.’s comparison

Exact methods:
SY: Santner and Yamagami (1993)
CT: Coe and Tamhane (1993)
AM: Agresti and Min (2001)
CZ: Chan and Zhang (1999)

Asymptotic score methods, AS:
Mee (1984)
Miettinen and Nurminen (1985)

Coverages of 90% Nominal Intervals for $n_1=30$, $n_2=30$
Generation of Parameter Space Points

One essential way in which our evaluation differs from that of Santner et al. is the algorithm used to generate the proportions \((p_1, p_2)\). They generated these from two independent uniform distributions. While this is obviously the simplest choice, it has limited relevance to the differences encountered in practice, e.g. in clinical trials. It would be very unreasonable to plan the size of a clinical trial on the assumption that such a large difference would occur. In practice, when two proportions get compared, they are generally broadly similar in magnitude.

We generated random parameter space points (PSPs) in a more appropriate way. Defining the risk difference parameter as \(RD = p_1 - p_2\) and the nuisance parameter as \(P_a = (p_1 + p_2)/2\), we first chose \(P_a\) from \(U(0, 1)\) and then \(RD > 0\) from \(U(0, 1 - |2P_a - 1|)\).
Proportion of parameter space points with coverage probability greater than $1-\alpha$ based on 10,000 parameter space points for each sample size $(n_1, n_2)$ pair.

<table>
<thead>
<tr>
<th>90% intervals</th>
<th>Wald</th>
<th>Mee</th>
<th>MN</th>
<th>Exact</th>
<th>Mid-P</th>
<th>Square &amp; add</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1 = n_2 = 5$</td>
<td>0.0005</td>
<td>0.7135</td>
<td>0.8564</td>
<td>0.9983</td>
<td>0.8203</td>
<td>0.7096</td>
</tr>
<tr>
<td>$n_1 = n_2 = 15$</td>
<td>0.0242</td>
<td>0.4321</td>
<td>0.5658</td>
<td>1</td>
<td>0.6070</td>
<td>0.7099</td>
</tr>
<tr>
<td>$n_1 = n_2 = 30$</td>
<td>0.0354</td>
<td>0.3746</td>
<td>0.5264</td>
<td>1</td>
<td>0.5492</td>
<td>0.6880</td>
</tr>
<tr>
<td>$n_1 = 5, n_2 = 15$</td>
<td>0.0178</td>
<td>0.6867</td>
<td>0.8082</td>
<td>0.9997</td>
<td>0.7729</td>
<td>0.7488</td>
</tr>
<tr>
<td>$n_1 = 15, n_2 = 25$</td>
<td>0.0144</td>
<td>0.4514</td>
<td>0.6027</td>
<td>0.9456</td>
<td>0.4999</td>
<td>0.6606</td>
</tr>
<tr>
<td>$n_1 = 25, n_2 = 35$</td>
<td>0.0299</td>
<td>0.3888</td>
<td>0.5437</td>
<td>0.9275</td>
<td>0.4297</td>
<td>0.6214</td>
</tr>
<tr>
<td>$n_1 = 20, n_2 = 50$</td>
<td>0.0242</td>
<td>0.4916</td>
<td>0.6466</td>
<td>0.9700</td>
<td>0.4764</td>
<td>0.6595</td>
</tr>
</tbody>
</table>

| 95% intervals |
|---------------|-------|--------|--------|-------|-------|--------------|
| $n_1 = n_2 = 5$ | 0      | 0.8140 | 0.9494 | 0.9991 | 0.8899 | 0.7351       |
| $n_1 = n_2 = 15$ | 0.0010 | 0.5485 | 0.7222 | 1     | 0.7047 | 0.7200       |
| $n_1 = n_2 = 30$ | 0.0026 | 0.4306 | 0.5926 | 1     | 0.6232 | 0.7231       |
| $n_1 = 5, n_2 = 15$ | 0.0032 | 0.7676 | 0.8777 | 0.9999 | 0.8539 | 0.7205       |
| $n_1 = 15, n_2 = 25$ | 0.0078 | 0.5897 | 0.7525 | 0.9561 | 0.6018 | 0.6873       |
| $n_1 = 25, n_2 = 35$ | 0.0052 | 0.4759 | 0.6454 | 0.9358 | 0.5056 | 0.6447       |
| $n_1 = 20, n_2 = 50$ | 0.0197 | 0.6335 | 0.7938 | 0.9787 | 0.5837 | 0.7040       |
Results on Coverage Probabilities

Santner’s percentage of PSPs conservative criterion turns out to be particularly sensitive to detect differences in coverage properties between different methods. Clearly, for several combinations of $n_1$, $n_2$, and $1-\alpha$, the MN interval has coverage over $1-\alpha$ for more than 50% of PSPs whereas the Mee interval has coverage over $1-\alpha$ for less than 50% of the same PSPs.

This criterion shows two methods in a favorable light here, the MN interval and, arguably even more so, the square-and-add interval. The “exact” interval (based on profiling out the nuisance parameter) is simply too wide, whereas even the otherwise excellent though computationally laborious mid-P interval does not fare quite so well by this criterion.

In fact, Santner et al. used the minimum proximity criterion, according to which the nominal level is taken to represent an absolute minimum, but with coverage probabilities (CPs) exceeding the nominal level as little as possible. This stance favors methods that are strictly conservative, and led the assessors to rank the performance of the asymptotic MN method in a pejorative manner.

Thus it is hardly surprising that Santner et al.’s comments come across as unfavorable, and that they found the MN method performed poorly according to their implied criterion of strict conservatism. However, their stated primary criterion of nearness of achieved coverage, interpreted literally would mean that the MN method was the best of the five they assessed.
Moving Average Coverage Probabilities for the Miettinen-Nurminen Confidence Interval

Moving average Coverage probabilities for Miettinen-Nurminen 95% CI with m=15, n=15
- Top slice (CP>nominal conf level)
  - meanCP=0.9542, minCP=0.8601, pctCons=63.1%

Moving average Coverage probabilities for Miettinen-Nurminen 95% CI with m=15, n=15
- Bottom slice (CP<=nominal conf level)
  - meanCP=0.9542, minCP=0.8601, pctCons=63.1%

Courtesy of Pete Laud, AstraZeneca UK Limited
Conclusions

We emphasize the desirability of a comprehensible quantitative appraisal of the precision of empirical findings in applied research. The better the users of statistical methods see how these function to achieve the desired objectives, the more likely the practitioners will be satisfied in applying them again.

We regard the arguments in favor of aligning typical rather than minimum coverage with the nominal level as very cogent ones. What really matters is not the theoretical coverage probability at a point value for the parameter, but rather an average over an interval of parameter values which represents a reasonable expression of prior uncertainty.

We call for a wider availability of score-based intervals in statistical software. These have gained a well-deserved place as appropriate tools for useful application in epidemiological and medical research practice. A correct form of the MN statistic, one that employs the unbiased estimator of variance will be incorporated into the next release of StatXact (Cyrus Mehta, personal communication, 2011).