Score intervals for the difference of two binomial proportions

Markku M. Nurminen, Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland

Robert G. Newcombe, Department of Primary Care & Public Health, Cardiff University, Cardiff, United Kingdom

Abstract

In this communication, we reconsider the approach to the constrained maximum likelihood estimation of asymptotic score intervals for the difference of two independent binomial proportions proposed by Miettinen and Nurminen (MN) in 1985. Then we review the subsequent biometric literature regarding the accuracy of the coverage probabilities of the different methods and respond to the unfavourable criticism that was presented recently against the MN method versus competing methods. We recall criteria appropriate to the evaluation of the compared methods and provide new Monte Carlo simulation results to supplement those of Newcombe's 1998 evaluation. Our conclusion is that the MN likelihood score interval estimates are highly recommended on account of both theoretical arguments and empirical evidence for use in epidemiology and medical research practice.

A general procedure for asymptotic interval estimation described by Cox and Hinkley\(^1\) may be based on a test statistic which contrasts the parameter of interest \(P\) to its empirical value \(p\), usually taken as a maximum likelihood estimate (MLE), and which has the variance of \(p\) as its denominator. These interval estimates are obtained by inverting a chi-squared (1 d.f.) test corresponding to a large-sample Gaussian distribution rendered free of any nuisance parameters. This principle of setting approximate confidence intervals entails that the limits will depend on the actual (unknown) value of \(P\) and, therefore, the estimation procedure should acknowledge this functional relation. Following the theory in the context of independent (unpaired) proportions (risks) or incidence densities (rates) under a binomial or Poisson distribution model, Miettinen and Nurminen\(^2\) developed “score” intervals for a comparative analysis of two rates (without or with stratification) for application in epidemiology and medical research. The Miettinen-Nurminen (MN) method, as this has become known in the biometric literature, has been theoretically demonstrated to possess desirable statistical properties.\(^3\)-\(^7\) The accuracy of the MN “score” intervals was empirically evinced for the three comparative parameters, viz. rate difference \((RD)\), rate ratio \((RR)\), and odds ratio \((OR)\), in extensive simulation studies by the original collaborators.\(^8\) Many applied statisticians,\(^9\)-\(^20\) have since corroborated these findings by quantifying the empirical coverage probabilities. Furthermore, the score interval for a difference of proportions based on individually matched (paired) data\(^21\) also has excellent coverage properties.\(^22\)
In the specific case of the \( RD \) parameter, a fairly recent Monte Carlo simulation study\(^{18} \) convincingly verified the good coverage performance of the MN method:

"Then we review several methods of calculating confidence intervals for the difference of two incidence rates assuming that the number of events come from a Poisson distribution. The methods considered include Wald’s method, the two-by-two table method, the Miettinen and Nurminen (MN) method, and the conditional MN method. For all but the MN method, explicit confidence intervals can be obtained.

For the MN method, some numerical iterations are required. The properties of these methods are evaluated by simulations. The results show that the MN method outperforms the other three in terms of the coverage of the confidence interval, especially when the rates and therefore the number of events are small."

Judged against this assessment, it strikes us as odd that Santner \textit{et al.}\(^{19} \) in a small-sample study for \( RD \) came to the opposite conclusion:

"The first method we compare is an asymptotic method based on the score statistic (AS) as proposed by Miettinen and Nurminen [1985]. Newcombe [1998] has shown that under certain asymptotic set-up, confidence intervals formed from the score statistic perform better than those formed from the Wald statistic. ... Although AS is claimed to perform reasonably well, it performs the worst in this study; about 50\% of the time it fails to achieve nominal coverage even with moderately large sample sizes from each binomial treatment. ... We also recommend that neither the AS nor the SY [Santner and Yamagami\(^{12} \)] methods be used in practice."

This is a perplexing situation for three reasons. First of all: are the original authors\(^2 \) incorrect when they claim that their method is an appropriate statistical procedure? Second, which one of the above-mentioned conflicting assessments should the users of the MN method — for instance, by means of the StatXact program\(^{23} \) — trust? Thirdly, what is the explanation of the apparent discrepancy between the conclusions and recommendation of Santner \textit{et al.}\(^{19} \) and the other researchers who have examined this issue? We submit a likely explanation. All the computational exercises in the Santner \textit{et al.} study\(^{19} \) were, no doubt, performed technically impeccably, given that they were represented by professional statisticians and computing scientists. But, the numerical results illustrated by the one-dimensional distribution plots were conceptually misinterpreted.

Recalling ´first principles’, there are different approaches to the construction of "likelihood-based" intervals by inverting a chi-square test. Miettinen and Nurminen\(^2 \) considered three statistics that centered on: 1) a simple difference or contrast; 2) the efficient score; and, 3) the profile log likelihood (which occurs in the likelihood ratio statistic). The chi-square functions proposed in their paper for the difference of two binomial proportions \( RD = P_1 - P_2 \) focused on the contrast \( P_1 \) versus \( P_2 + RD \). The corresponding empirical contrast or ´difference’ score is \( p_1 - (p_2 + RD) = p_1 - p_2 - RD \), with its expectations equal to zero and variance estimate involving the parameter-constrained MLEs \( \hat{P}_1 \) and \( \hat{P}_2 \) that depend on \( RD = \hat{P}_1 - \hat{P}_2 \). This ´difference score’ has the virtue of simplicity, but it is not based on any principle of suggesting optimality in terms of total capture of the comparative information in the data. Theory, as outlined by Cox and Hinkley,\(^2 \) suggests as the ´efficient score’ the derivative of the log-likelihood with respect to the parameter at issue (here \( RD \)). However, the ´efficient score’ approach was found to be questionable for stratified series in general (Eq. 32 in Ref. 2) due to aberrant results for strata with overall (marginal) rates at the boundary values of 0 or 1, at least without resorting to Bayesian techniques. The performance of the
likelihood-ratio approach was also somewhat inferior to the proposed one. For these theoretical and empirical considerations the proposed MN method was based on the ‘difference score’ (Eq. 8-9, in Ref. 2). This, in fact, was the ‘asymptotic score’ statistic evaluated by Santner et al., not the ‘efficient score’ statistic, with the important omission of the ostensibly trivial variance bias correction (which we discuss below). With incidence density data, however, the MN interval that they derived under a Poisson sampling model is in fact consistent with the latter score statistic.

Next, one may think of several ways to gauge the accuracy of constructed confidence intervals, but the main function of these stochastic bounds is that they should include the particular value of the parameter \( P \) to be estimated. The coverage probability (CP) is defined as \( \Pr[L \leq P \leq U] \), where \( L \) and \( U \) are the calculated lower and upper limits that, in combination, form the interval \([L, U]\). The aim is that \( L \leq P \leq U \) should occur with probability \( 1 - \alpha \), and \( L < P \) and \( U < P \) each with probability \( \alpha/2 \). It is not possible to attain this property for all \( P \), because of the discontinuous behaviour of the assumed binomial or Poisson sampling distribution. The crucial issue then is whether the nominal CP, \( 1 - \alpha \), is meant to be a minimum or an average. Some confidence interval methods seek to align the minimum CP with \( 1 - \alpha \), others the mean CP. Modern statistical practice has moved very much in the direction of regarding \( 1 - \alpha \) as representing average coverage. When complex Bayesian models are fitted using MCMC (Markov Chain Monte Carlo) software such as WinBugs (Bayesian inference Using Gibbs Sampling), the coverage achieved is implicitly an average.

Appropriate coverage is the fundamental requirement. Among interval methods that produce acceptable coverage, another important criterion is that we seek to achieve this with minimum interval width. Having compared eleven alternative methods using several criteria including CP and average interval width, Newcombe remarked that,

"A likelihood-based approach [Miettinen-Nurminen, 1985] has been suggested as theoretically most appealing [Professor G.A. Barnard, personal communication, 1992], by definition it already incorporates an important aspect of symmetry about the maximum likelihood estimate (MLE) \( p \). There is no question of either continuity 'correction' or mid-p modification to adjust this method's coverage properties systematically."

Yet another criterion is the equal-tailed probabilities for a two-sided confidence interval. As remarked by Miettinen, "Such a two-sided interval, with the 'nonconfidence' (\( \alpha \)) equally split, is designed not only to cover the actual parameter value (unknown) in 100(1 - \( \alpha \))% of applications, but also to have the two types of noncoverage occur with equal frequencies. In 100(\( \alpha/2 \))% of applications the lower bound, and with this same frequency the upper bound falls below the actual value of the parameter."

This frequency behavior of a 100(1-\( \alpha \))% confidence interval does not mean that one may be 100(1 - \( \alpha \))% confident that the actual value lies within a particular interval already known, or that the probability (credibility) of its falling below (above) the interval is 100(\( \alpha/2 \))%.

Gart and Nam concluded that,

"It is found that the method based on likelihood scores (Miettinen and Nurminen, 1985) performs best in achieving the nominal confidence coefficient."

A meaningful interpretation of interval accuracy (that is, magnitude as measured by the closeness of the average achieved coverage probability to the nominal value) can be more important than interval location (that is, equal allocation of the tail probabilities) or minimum width (that is, the precision of the average
estimate). Results obtained by both Miettinen & Nurminen\textsuperscript{2} and Newcombe\textsuperscript{13} show that the MN method has favorable location properties.

We recognize that the MN interval is not set up in terms of tail areas, but, instead, it is based on the likelihood function. Our interpretation is that that they should be rather be regarded as leading to a different kind of interval estimate, which should be be called ‘likelihood interval’ to distinguish it from a ‘confidence interval’ or indeed from a ‘Bayes (credible) interval’. As already pointed out, there are different ways of extracting the information from the likelihood function. Hence the terminology varies regarding what is specifically meant by the general notion of ‘likelihood-based’ interval estimation.

The detailed results of the original Monte Carlo simulation study\textsuperscript{8} demonstrated that the confidence coefficient of the proposed (MN) 95 percent interval for $RD$ was 94 to 96 percent, whereas that of the standard interval (with $RD$ estimated by $p_1 - p_2$\textsuperscript{2}, Eq. 7 was 65 to 95 percent. The most inaccurate usual (Wald-type) intervals were obtained when the sample sizes differed a great deal from each other. For example, in the typical epidemiologic situation when the number of exposed subjects $n_1 \ll n_2$ the number on non-exposed subjects, the confidence coefficient associated with the usual limit was around 90 percent rather than the nominal 95%. For two-sided intervals, which are the regularly used ones in epidemiology, a conventional optimality requirement in constructing confidence intervals in symmetrical situations is to allot equal probability at either tail area. On account of the asymmetry of the underlying sampling distribution the error rates of the symmetric usual interval either fell short of the nominal level (2.5%) or exceeded it far more, both at the lower and upper end, as compared to the corresponding rates associated with the proposed interval. Pointed discrepancies were especially evident in lopsided samples with extreme parameter values.

In the assessment of the conservatism (Cox and Hinkley\textsuperscript{1}, Section 7.2 Expression 6, p. 210) of the compared methods, the mean widths of all the simulated intervals matched one another quite well. But, in instances where the empirical intervals did not cover the true $RD$, the usual large-sample method tended to be more biased (wider) than the proposed MN method.

Recently, Pradhan and Banerjee\textsuperscript{24} proposed a method for estimating a confidence interval for two independent binomial proportions based on profile likelihood, where the likelihood is weighted by noninformative Jeffreys prior. Simulations showed that this profile method performs well compared to the simpler Wilson-type score method proposed by Newcombe.\textsuperscript{14}

Incidentally, while sampling theory just does not provide for exact intervals for $RD$ and $RR$, a complex solution may be found by the use of Bayesian formulations: Nurminen and Mutanen\textsuperscript{25} derived exact Bayesian posterior probability intervals corresponding to ignorance priors as analogues of “frequentist” confidence intervals. For the three parameters ($RD$, $RR$, $OR$), the priors are reasonable in that they apportioned an equal amount of density on either side of the null-hypothesis value. In the particular case of $RD$, the prior density was assumed to be a symmetric triangular distribution in the range $-1 \leq RD \leq +1$ with its tine density (Schmidt\textsuperscript{26}) of 0.5 located at the null value $RD = 0$.

Santner at al.\textsuperscript{19} took the proximity the actual achieved coverage as the primary criterion for the performance of the interval systems compared. In their computations, they used the “exact” achieved coverage, this being based on precise enumeration of the assumed underlying distribution of a product of two binomials. Given the “long-run frequency” interpretation of probability, the suggested parameter-conditional interval estimates obtained by using the asymptotic MN method for $RD$, incontrovertibly showed, in the Santner at al.\textsuperscript{19}, Fig.\textsuperscript{1} study, coverage probabilities much nearer to the nominal 90 percent for all the specifications of the two proportions and sample sizes compared to the other
methods. The four ‘exact’ intervals had average CPs ranging from 91 to over 95 percent.

However, “exact” does not mean attaining a CP equal to the nominal confidence level for all RD and sample size n constituting the parameter space. 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 CPs exceeding the nominal level as little as possible. This stance favors methods that are strictly conservative and lead the assessors to rank the performance of the asymptotic MN method in a pejorative manner. Hence it is hardly surprising that Santner et al.’s comments come across as unfavourable, and that they found the MN method performing 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 assessed. We would not like to make the concession to their standpoint that is implied by saying that the MN interval is of a different kind to the conventional confidence interval, meaning that conventionally confidence intervals are necessarily designed to align minimum CP with 1 - α.

We recall the verdict that stated (notation modified), "While it is of some interest to examine minₐ₀<ᵦ<₁ CP, either for chosen n or for all n, in that a value « 1 - α should be regarded as contraindicating the method, nevertheless to average a CP further, over a pseudo-prior [density function] f(P) for P, or a series of representative points in the (n, RD) space sampled accordingly, does no harm, and is arguably more appropriate."

The MN method of interval estimation for RD is based on a test statistic that is approximately Gaussian distributed in a large sample (of finite size n >1). To make the Bernoulli variance of the empirical rate difference contrast asymptotically unbiased, Miettinen and Nurminen applied a correction factor, 2 ≥ n/(n-1) > 1, as a multiplier in the denominator of the chi-square test. Unfortunately, this factor was omitted both from the score statistic of the AS method in Santner et al. and from the same equation in StatXact. The formula they present for the AS method is, in fact, the one described by Mee, not the MN method. These two “score-based methods” are identical apart from use of a factor n/(n-1). It follows that the MN interval always includes the Mee interval. The difference in width is, of course, greater for smaller sample sizes than for larger ones. We believe that the main explanation for Santner et al.’s findings is that they have confused these two look-alike methods. The numerical evaluations in the Appendix indicate that there are some (n₁, n₂) pairs for which the Mee interval gives undercoverage for over 50% of the PSPs but the MN interval does not.

This omission rendered the Mee intervals to become narrower compared to the MN intervals computed using the multiplier. Therefore, the numerical evaluation of the AS method carried out by Santner et al. is severely biased. This correction factor is of importance, especially when evaluating performance accuracy in small samples. A single example illustrates this point: for the proportion-type rates p₁ = 4/5 and p₂ = 2/5, the MN interval for the RD computed with the factor is (-0.228, 0.794) and without it (-0.199, 0.783), respectively. Even in large samples, in the general case of stratified series when the stratum-specific marginal numbers nᵢ are small, the factor nᵢ/(nᵢ-1) becomes quite relevant. In the special case of matched pairs, it takes on the value 2.

The most recent simulation study by Lang highlights the advantage of using the score and profile likelihood intervals, including the Miettinen-Nurminen intervals, rather than the Wald intervals.

"The consensus in the literature is that the score and profile likelihood intervals have better coverage properties than Wald intervals. ... Aligning with findings in
the existing literature, the score and profile likelihood intervals have coverage probabilities reasonably close to the nominal confidence level, even when the estimand [risk difference parameter] is closer to the upper boundary of the parameter space and the sample size is small. ... This supports the recommendation that rather than Wald, score or profile likelihood intervals should be used."

Final methodological points concern the computation of the likelihood-based intervals using the MN methodology. Stamey et al.\textsuperscript{29} determined that for all parameter configurations considered in a single two-by-two table, the profile likelihood interval performs adequately in terms of keeping best the coverage probability close to nominal with minimum expected width. But, to derive the restricted likelihood estimators, they augmented the data with the unobservable (latent) number of cases and non-cases and the used the EM algorithm to determine profile MLEs form the likelihood equation. They resorted to this procedure because, "No apparent closed form profile-likelihood estimators are available from [the likelihood for the product of two binomials] when any parameter is held fixed." Similarly, Mee\textsuperscript{27} noted, "The [Mee] interval (1) has the disadvantage of requiring iterative computation for the endpoints and for the evaluation of $\tilde{P}_1$ and $\tilde{P}_2$ at each iterate of the endpoints." As already pointed out, closed-form solution to the likelihood for the difference parameter is known. Thus, for any specified value of $RD$ the restricted MLEs of the two proportions $\tilde{P}_1$ and $\tilde{P}_2$ can be obtained directly, not needing a simultaneous iterative solution to pairwise equations; only the computations of the confidence intervals themselves require a trial-and-error solution.\textsuperscript{30}

When extending the statistical inference to stratified two-by-two samples, the expanded Cochran-Mantel-Haenszel (CMH) –type weighting strategy proposed by Miettinen and Nurminen requires a circular procedure to implement. This is nested within another iterative procedure for finding confidence limits. Lu\textsuperscript{30} improved the CMH weighting system so that it does not require such a complicated iterative scheme. Compared to the weighting strategy in the original paper, the new strategy proposed by Lu does not affect the performance of the resulting confidence intervals while significantly reducing the computational time. Thus, given the modern ease of computing, there is no longer a defensible reason in favor of a distinctly inferior approach to interval estimation of $RD$, neither in this context, nor in the interval estimation of risks differences in case-base studies.\textsuperscript{31}

Closing remarks. Considering the widely diverging stances on the issue of coverage probability (see Brown et al.,\textsuperscript{32} with discussion), and being cognizant that the fundamental purpose of applied statistics is to communicate a comprehensible quantitative appraisal of the precision of empirical findings, we stress the desirability of conceptual simplicity, methodological transparency, and consistency with principles of the general theory. The better the users of statistical methods see how they function to achieve the desired objectives, the more likely the practitioners will be satisfied in applying these methods anew.

In review of the above-cited biometric literature, we regard the arguments that have been presented to justify aligning typical rather than minimum coverage with the nominal level are very cogent ones — along with our own argument that what really matters is not the theoretical coverage probability at a point value for the parameter, rather than an average over an interval of parameter values which represents a reasonable expression of prior uncertainty.

In conclusion, it appears that the score-based intervals, including the MN method, have gained their place as appropriate tools for useful application in the epidemiologic and medical research practice. A wider availability of these
References


Appendix

Coverage of interval estimates for a difference of proportions

By R.G. Newcombe

The results presented here were produced in response to issues raised by Santner et al. Their article considered several interval estimation methods for a difference of two unpaired proportion, including the asymptotic score statistic (AS) method which they attributed to Miettinen and Nurminen. Their evaluation involved generating 10,000 parameter space points (PSPs) for each of seven \((n_1, n_2)\) pairs, namely (5, 5), (15, 15), (30, 30), (5, 15), (15, 25), (25, 35) and (20, 50). The primary presentation of results was a graphical display of the frequency distributions of coverage probabilities and expected lengths for intervals designed to have 90% two-sided coverage. For some methods, Santner et al. found that over 50% of PSPs resulted in an achieved coverage probability lower than the nominal 1-\(\alpha\). Consequently, in these cases, they also reported the percentage of PSPs with undercoverage. They reported that the AS method had undercoverage for over 65% of PSPs with \(n_1 = n_2 = 15\), and for over 58% of PSPs with \(n_1 = n_2 = 30\).

Newcombe’s original evaluation concentrated mainly on intervals designed to have 95% two-sided coverage, but it also obtained some results for intervals with 90% nominal coverage. The results presented were based on 9,200 PSPs, 40 for each of 230 \((n_1, n_2)\) pairs. These included all pairs with \(n_1 = n_2 = n\) for 5 \(\leq n \leq 50\), plus a deliberate representative selection of pairs with unequal values of \(n_1\) and \(n_2\) in this range — as it happens, these did not include any of Santner’s four unequal pairs. This degree of replication is clearly too small to determine whether Santner et al.’s claims are justified, so the computations are repeated here using their seven \((n_1, n_2)\) pairs, each with 10,000 PSPs.

One way in which our evaluation differs from theirs is the algorithm to generate the two proportions \((p_1, p_2)\). Santner et al. generated these from two independent uniform distributions. While this is obviously the simplest choice, we believe that it has limited relevance to the differences encountered in practice, for instance in clinical trials. In our trial of sentinel node biopsy (SNB) for staging of clinically early breast cancer, arm morbidity occurred in approximately 20 percent of patients on the SNB regime, compared to approximately 70 percent of the patients allocated to the control regime of axillary dissection. But this very large difference, \(p_1 - p_2 = -0.50\), corresponding to a number needed to treat (NNT) of 2, is exceptional in size. It would be very unreasonable to plan the size of a clinical trial on the assumption that such a large difference would occur. In general research practice, when two proportions get compared, they are broadly similar in magnitude. With this viewpoint in mind, we chose to generate random PSPs in a more appropriate way. Defining the parameter of interest as \(RD = P_1 - P_2\) and the nuisance parameter as \(P = (P_1 + P_2)/2\), we chose \(P\) from Uniform \((0,1)\) and then \(RD > 0\) from U(0,1-\(|2P-1|\)).

When \(P_1\) and \(P_2\) are generated from independent uniform distributions, \(RD\) has a triangular distribution, so the median value of \(|RD|\) is simply 1/3. With the chosen sampling scheme, the median \(RD\) was 0.193 in Newcombe’s original evaluation (0.186 in the rerun here), which seems much more realistic

Table A gives the achieved coverage of 90% intervals, taken from Newcombe’s original evaluation. These computations relate to a total of 9,200 PSPs, 40 for each \((n_1, n_2)\) pair. Figures are shown for 6 of the 11 methods evaluated. The first block gives the mean and minimum coverage for all 9,200 PSPs, exactly as in Table III of the reference, and also the proportion of PSPs with achieved coverage above the nominal 1-\(\alpha\) = 0.9. The
subsequent blocks give the results for the subsets of PSPs with \( n_1 = n_2 = 5, 15 \) and 30, and are consequently each based on only 40 PSPs.

In Table A, the shaded cells are of particular interest. Taking all 9,200 PSPs together, adequate coverage occurred for 62% of PSPs using the MN interval, but only for 48% of PSPs when the Mee\(^{27}\) interval is used. When \( n_1 = n_2 = 15 \) or 30, we get a similar effect — of course, the degree of replication here is insufficient to say whether this proportion really is greater or less than 50% for the MN interval.

### Table A. Coverage of Intervals with Confidence Coefficient \( \alpha = 90\% \).*

<table>
<thead>
<tr>
<th>Criterion \ Method</th>
<th>Wald</th>
<th>Mee</th>
<th>M &amp; N</th>
<th>Exact</th>
<th>Mid-P</th>
<th>Sq &amp; Add</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All (9,200 PSPs)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean CP</td>
<td>0.833</td>
<td>0.908</td>
<td>0.911</td>
<td>0.931</td>
<td>0.912</td>
<td>0.916</td>
</tr>
<tr>
<td>Min CP</td>
<td>0</td>
<td>0.806</td>
<td>0.806</td>
<td>0.886</td>
<td>0.837</td>
<td>0.823</td>
</tr>
<tr>
<td>Proportion of PSPs with CP &gt; 1-( \alpha )</td>
<td>0.027</td>
<td>0.479</td>
<td>0.620</td>
<td>0.960</td>
<td>0.54</td>
<td>0.674</td>
</tr>
</tbody>
</table>

| 5 & 5 (40 PSPs) |      |     |       |       |       |          |
| Mean CP         | 0.695| 0.938| 0.949 | 0.974 | 0.948 | 0.929    |
| Min CP          | 0.311| 0.851| 0.875 | 0.934 | 0.865 | 0.847    |
| Proportion of PSPs with CP > 1-\( \alpha \) | 0  | 0.8 | 0.9 | 1 | 0.9 | 0.65 |

| 15 & 15 (40 PSPs) |      |     |       |       |       |          |
| Mean CP         | 0.825| 0.906| 0.913 | 0.949 | 0.915 | 0.921    |
| Min CP          | 0.233| 0.870| 0.872 | 0.915 | 0.876 | 0.863    |
| Proportion of PSPs with CP > 1-\( \alpha \) | 0.075| 0.375| 0.475 | 1 | 0.55 | 0.8  |

| 30 & 30 (40 PSPs) |      |     |       |       |       |          |
| Mean CP         | 0.851| 0.903| 0.905 | 0.936 | 0.906 | 0.914    |
| Min CP          | 0.366| 0.876| 0.876 | 0.908 | 0.878 | 0.883    |
| Proportion of PSPs with CP > 1-\( \alpha \) | 0.025| 0.4 | 0.475 | 1 | 0.475 | 0.65 |

*Newcombe's\(^{13}\) 1998 evaluation. Based on 9,200 PSPs, 40 for each \((n_1, n_2)\) pair.
In recognition of the severe limitation of the above figures, Table B shows similar results for Santner's three equal sample size pairs, and also the four with unequal sample sizes, using identical methodology but with 10,000 PSPs in each case. Results are shown for 95% as well as 90% intervals. Santner's % 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. The percentage of PSPs conservative 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 of course simply too wide, whereas even the otherwise excellent though computationally laborious corresponding mid-$P$ interval doesn't fare quite so well by this criterion.

**Table B.** Proportion of Parameter Space Points with Coverage Probability $>1-\alpha$.*

<table>
<thead>
<tr>
<th>$(n_1,n_2) \backslash$ Method</th>
<th>Wald</th>
<th>Mee</th>
<th>M &amp; N</th>
<th>Exact</th>
<th>Mid-P</th>
<th>Sq &amp; Add</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>90% intervals</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 &amp; 5</td>
<td>0.001</td>
<td>0.714</td>
<td>0.856</td>
<td>0.998</td>
<td>0.820</td>
<td>0.710</td>
</tr>
<tr>
<td>15 &amp; 15</td>
<td>0.024</td>
<td>0.432</td>
<td>0.566</td>
<td>1</td>
<td>0.607</td>
<td>0.710</td>
</tr>
<tr>
<td>30 &amp; 30</td>
<td>0.035</td>
<td>0.375</td>
<td>0.526</td>
<td>1</td>
<td>0.549</td>
<td>0.688</td>
</tr>
<tr>
<td>5 &amp; 15</td>
<td>0.018</td>
<td>0.687</td>
<td>0.808</td>
<td>1</td>
<td>0.773</td>
<td>0.749</td>
</tr>
<tr>
<td>15 &amp; 25</td>
<td>0.014</td>
<td>0.451</td>
<td>0.603</td>
<td>0.946</td>
<td>0.500</td>
<td>0.661</td>
</tr>
<tr>
<td>25 &amp; 35</td>
<td>0.030</td>
<td>0.389</td>
<td>0.544</td>
<td>0.928</td>
<td>0.430</td>
<td>0.621</td>
</tr>
<tr>
<td>20 &amp; 50</td>
<td>0.024</td>
<td>0.492</td>
<td>0.647</td>
<td>0.970</td>
<td>0.476</td>
<td>0.660</td>
</tr>
<tr>
<td><strong>95% intervals</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 &amp; 5</td>
<td>0</td>
<td>0.814</td>
<td>0.949</td>
<td>0.999</td>
<td>0.890</td>
<td>0.735</td>
</tr>
<tr>
<td>15 &amp; 15</td>
<td>0.001</td>
<td>0.549</td>
<td>0.722</td>
<td>1</td>
<td>0.705</td>
<td>0.720</td>
</tr>
<tr>
<td>30 &amp; 30</td>
<td>0.003</td>
<td>0.431</td>
<td>0.593</td>
<td>1</td>
<td>0.623</td>
<td>0.723</td>
</tr>
<tr>
<td>5 &amp; 15</td>
<td>0.003</td>
<td>0.768</td>
<td>0.878</td>
<td>1</td>
<td>0.854</td>
<td>0.721</td>
</tr>
<tr>
<td>15 &amp; 25</td>
<td>0.008</td>
<td>0.590</td>
<td>0.753</td>
<td>0.956</td>
<td>0.602</td>
<td>0.687</td>
</tr>
<tr>
<td>25 &amp; 35</td>
<td>0.005</td>
<td>0.476</td>
<td>0.645</td>
<td>0.936</td>
<td>0.506</td>
<td>0.645</td>
</tr>
<tr>
<td>20 &amp; 50</td>
<td>0.020</td>
<td>0.634</td>
<td>0.794</td>
<td>0.979</td>
<td>0.584</td>
<td>0.704</td>
</tr>
</tbody>
</table>

* Based on 10,000 PSPs for each $(n_1, n_2)$ pair.