Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **54**(3), 2012, 281–299 doi: 10.1111/j.1467-842X.2012.00680.x

CONFIDENCE INTERVALS FOR THE WEIGHTED SUM OF TWO INDEPENDENT BINOMIAL PROPORTIONS

GEOFFREY DECROUEZ[1,*] AND ANDREW P. ROBINSON[2]

University of Melbourne and ACERA

Summary

Confidence intervals for the difference of two binomial proportions are well known, however, confidence intervals for the weighted sum of two binomial proportions are less studied. We develop and compare seven methods for constructing confidence intervals for the weighted sum of two independent binomial proportions. The interval estimates are constructed by inverting the Wald test, the score test and the Likelihood ratio test. The weights can be negative, so our results generalize those for the difference between two independent proportions. We provide a numerical study that shows that these confidence intervals based on large-sample approximations perform very well, even when a relatively small amount of data is available. The intervals based on the inversion of the score test showed the best performance. Finally, we show that as for the difference of two binomial proportions, adding four pseudo-outcomes to the Wald interval for the weighted sum of two binomial proportions improves its coverage significantly, and we provide a justification for this correction.

Key words: border security; leakage survey; likelihood ratio test; quarantine inspection; score test; small sample; sum of proportions; Wald test.

1. Introduction

Construction of confidence intervals for the difference of proportions has been widely studied, due to its numerous applications in biostatistics and elsewhere, see e.g. Anbar (1983), Newcombe (1998a); Zhou, Tsao & Qin (2004). However, the construction of interval estimates for the sum of proportions, and more generally the weighted sum or proportions, has received much less attention.

Motivation for this study came from a recent undertaking to develop performance indicators in the operation of quarantine inspection, (Robinson; Decrouez & Cannon, ‘A regulator’s performance indicator’, pers. comm., 2012). Briefly, consider the following setup. A collection of N items is sequentially presented for inspection, where N is large. Then n_1 of these items are randomly selected with equal and known probability and inspected for contamination. Suppose that x_1 items are identified with quarantine contamination, and we assume that $x_1 \sim Bi(n_1, p_1)$. The inspections are known to be imperfect,

*Author to whom correspondence should be addressed.

[1] Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia.
e-mail: dgg@unimelb.edu.au

[2] ACERA and Department of Mathematics and Statistics, ACERA, Parkville, VIC 3010, Australia.
e-mail: andrewpr@ms.unimelb.edu.au

Acknowledgements. The authors are grateful to Alan Agresti and Rob Cannon for helpful advice and constuctive review comments, and to Kathleeen Quan of DAFF Biosecurity for provision of the inspection data. This research was partially supported by ACERA and ARC grant DP110103159.

but the probability of contamination being missed is unknown. Those x_1 items are cleaned; that is the contamination is removed.

Then, an independent, random, equal and known probability sub-sample of $n_2(n_2 \ll n_1)$ items is taken from the $n_1 - x_1$ items that were inspected and passed. These n_2 items are inspected thoroughly in a process that is assumed to detect all contamination. Suppose that x_2 items are identified as contaminated, and we assume that $x_2 \sim Bi(n_2, p_2)$. The re-sampling process is referred to as a leakage survey, but it is comparable to a gold-standard test procedure.

The x_1 out of n_1 items represent the rate at which contamination is approaching the border, whereas the x_2 out of n_2 items represent the rate at which inspections fail to capture contamination. In an operational setting these random variables can be assumed to be independent, and they may differ sharply.

An estimate of the proportion of items that has passed through the entire inspection process that are still contaminated is developed through the following process. There are now two streams of leaked items to consider: those among the n_1 inspected items that are not intercepted and those among the $N - n_1$ items that are not inspected. The count of contaminated units that leaks through the inspection is estimated by

$$\widehat{l} = \frac{n_1 x_2}{n_2},$$

so the estimated total number of contaminated units in the whole pathway is

$$\widehat{c} = \left(x_1 + \frac{n_1 x_2}{n_2}\right)\frac{N}{n_1}$$

and the estimated number that remains after the inspection is

$$\widehat{L} = \left(x_1 + \frac{n_1 x_2}{n_2}\right)\frac{N}{n_1} - x_1 - x_2$$

and, finally, the pathway-level leakage rate, expressed as the proportion of arriving items, is

$$\widehat{\theta} = \frac{x_1}{n_1}\frac{N - n_1}{N} + \frac{x_2}{n_2}\frac{N - n_2}{N}$$

which can be rewritten as $\widehat{\theta} = \alpha\widehat{p}_1 + \beta\widehat{p}_2$, where $\widehat{p}_1 = x_1/n_1$ is the estimate of the binomial parameter for the random variable x_1 defined as above, $\widehat{p}_2 = x_2/n_2$ is the estimate of the binomial parameter for the independent random variable x_2 defined as above, and the two constants are $\alpha = 1 - n_1/N$ and $\beta = 1 - n_2/N$ which depend only on known elements of the design.

It is of interest to provide an interval estimate for the leakage in order to provide the manager with information about the quality of the estimate.

In medical studies, the assessment of performance of a prediction model using a weighted average of sensitivity and specificity has received increased interest recently. The present study provides methodology for constructing confidence intervals for these novel measures. In Vach, Gerke & Høiland-Carlsen (2012), the success of a diagnosis study is defined using the so-called 'liberal criterion', expressed as the weighted average of sensitivity and specificity. Decision-analytic methods take into account the harm from unnecessary treatment or overtreatment using the net benefit, expressed as the weighted sum of true and false positive counts, see Vickers & Elkin (2006); Vickers *et al.* (2008). See also Newcombe (2001); Steyerberg *et al.* (2010) for assessment of these new measures of performance.

A general method for constructing confidence intervals is by the inversion of a test statistic. Suppose we wish to find a $100(1-\alpha)\%$ confidence interval for some parameter θ. Denote by $\widehat{\theta}$ the maximum likelihood estimate of θ, given independent and identically distributed observations $X_1,\ldots,X_n$. Then, under general assumptions, $n^{1/2}(\widehat{\theta}-\theta)$ is asymptotically normally distributed with zero mean and variance $I^{-1}(\theta)$, where $I(\theta)$ is the Fisher information for X_1. This setup suggests using the Wald test to test the null hypothesis $H_0:\theta=\theta_0$ against the two-sided alternative $\theta\neq\theta_0$. The statistic $W_n(\theta_0)=n(\widehat{\theta}-\theta_0)^2I(\theta_0)$ can then be used to construct a $100(1-\alpha)\%$ confidence interval, given by the set of θ_0 values such that $\mathbf{P}(W_n^2(\theta_0)\leq z_{\alpha/2}^2)=1-\alpha$, where z_α is the $(1-\alpha)$-th quantile of the normal distribution. Alternatively, if we denote by $l(\theta|x_1,\ldots,x_n)$ the log-likelihood of the data, and $S(\theta)=\partial l(\theta|x_1,\ldots,x_n)/\partial\theta$ the score function, the asymptotic chi-square distribution of the statistic $Y_n(\theta_0)=S^2(\theta_0)/I(\theta_0)$ can be used to provide a confidence interval for θ. Finally, one can invert a likelihood-ratio test, with statistic given by $Z_n(\theta)=-2(l(\theta_0|x_1,\ldots,x_n)-l(\widehat{\theta}|x_1,\ldots,x_n))$, which can be shown to have an asymptotic chi-square distribution.

These three methods, the Wald test, the score test and the likelihood-ratio test, can be used to provide confidence intervals for the weighted sum of two independent binomial proportions.

The rest of the paper is laid out as follows. In Section 2, we explain how to obtain such intervals. We outline and report a numerical study of the performance of these methods in Section 3, Section 4 provides a discussion, and Section 5 gives an application.

2. Interval estimators

Let X_1 and X_2 be two independent binomial random variables, with respective sample sizes n_1 and n_2, and probability of success p_1 and p_2. In this section, we construct confidence intervals for the weighted sum $\alpha p_1+\beta p_2$, where $\alpha\neq 0$, $\beta\neq 0$, by inverting three two-sided tests: the Wald test, the score test and the likelihood ratio test. To keep this study as general as possible, we do not assume that α and β are strictly positive. Therefore, confidence intervals constructed here match existing methods for the difference of two proportions in the particular case $\beta=1$ and $\alpha=-1$. Suppose without loss of generality that $|\alpha|\leq|\beta|$. Then $\alpha p_1+\beta p_2=\beta(\alpha/\beta p_1+p_2)$, where $|\alpha/\beta|\leq 1$. Therefore, we can restrict the construction of confidence intervals for the weighted sum of two binomial proportions $\alpha p_1+\beta p_2$ to the case where $-1\leq\alpha\leq 1$, $\alpha\neq 0$, and $\beta=1$. Let $\theta=\alpha p_1+p_2$ and $\psi=\alpha p_1-p_2$. The range of possible values of θ, the parameter of interest, is $[0, 1+\alpha]$ if $\alpha>0$ and $[\alpha,1]$ if $\alpha<0$.

2.1. Wald interval

The first interval is obtained from the Wald statistical test, which evaluates the standard error of the maximum likelihood estimate $\widehat{\theta}=\alpha\widehat{p}_1+\widehat{p}_2$, where $\widehat{p}_1=X_1/n_1$ and

$\widehat{p}_2 = X_2/n_2$. Denote by $z = z_{a/2}$ the (1 − *a*/2)-th quantile of the normal distribution. Based on a large-sample approximation, a confidence interval for $\widehat{\theta}$ is given by

$$\widehat{\theta} \pm z\left(\alpha^2\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}\right)^{1/2}. \tag{1}$$

We refer to this interval as the Wald interval. Wald intervals for a single proportion or for the difference of two proportions are known to perform poorly, and are usually well below the nominal coverage (Newcombe 1998a,b), although they are still widely used in practice because of their simplicity. Adding artificial outcomes is a simple and efficient way to get better coverage, as explained in Agresti & Coull (1998) and Agresti & Caffo (2000). In Section 2.3, we propose a justification for applying such a correction for sums and differences of proportions. The Wald interval also suffers from overshoot, that is, bounds of the calculated interval can be outside the range of possible values for θ.

2.2. Haldane and Jeffreys–Perks interval

The next two intervals are based on confidence intervals constructed for the difference of two binomial proportions that were introduced by Beal (1987). It is convenient to express the variance of $\widehat{\theta}$ in terms of $\theta = \alpha p_1 + p_2$ and $\psi = \alpha p_1 - p_2$,

$$\vartheta(\theta,\psi;u,v,\alpha) = u((\alpha+1-\theta)\theta + (\alpha-1-\psi)\psi) + v(\theta(\alpha-1) + \psi(\alpha+1) - 2\theta\psi),$$

where

$$u = \frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right), \quad v = \frac{1}{4}\left(\frac{1}{n_1} - \frac{1}{n_2}\right).$$

A confidence interval for θ comprises the set of values of θ such that

$$(\theta - \widehat{\theta})^2 \leq z^2\vartheta(\widetilde{\theta}, \widetilde{\psi}; u, v, \alpha), \tag{2}$$

where $\widetilde{\theta}$ and $\widetilde{\psi}$ are expressions for θ and ψ (see examples below), so that the bounds of the confidence interval correspond to equality in (2). For example, the Wald interval is obtained with $\widetilde{\theta} = \widehat{\theta}$ and $\widetilde{\psi} = \widehat{\psi}$, where $\widehat{\psi} = \alpha\widehat{p}_1 - \widehat{p}_2$ denotes the maximum likelihood estimate of ψ. Here we obtain another interval with $\widetilde{\theta} = \theta$ and $\widetilde{\psi} = \widehat{\psi}$. This new interval, given below, is the equivalent of the Wilson interval for the single binomial proportion, see Wilson (1927), which is known to be an improved confidence interval compared to the basic Wald interval. With $\widetilde{\theta} = \theta$ and $\widetilde{\psi} = \widehat{\psi}$, (2) reduces to a quadratic equation given by $a_2\theta^2 + a_1\theta + a_0 = 0$, with

$$\begin{aligned}
a_2 &= 1 + z^2u\\
a_1 &= -2\left(\widehat{\theta} + z^2(u(\alpha+1) + v(\alpha-1) - 2\widehat{\psi}v)/2\right)\\
a_0 &= \widehat{\theta}^2 - z^2\widehat{\psi}(u(\alpha-1-\widehat{\psi}) + v(\alpha+1)),
\end{aligned}$$

whose roots delimit a new confidence interval. The roots are

$$\frac{\widehat{\theta} + \frac{z^2}{2}(u(\alpha+1) + v(\alpha-1) - 2\widehat{\psi}v)}{1+z^2u} \pm z\frac{(\vartheta(\widehat{\theta}, \widehat{\psi}; u, v, \alpha) + z^2\Delta)^{1/2}}{1+z^2u}, \qquad (3)$$

where

$$\begin{aligned}\Delta = u^2\left(\frac{1}{4}(\alpha+1)^2 + \widehat{\psi}(\alpha-1-\widehat{\psi})\right) + v^2\left(\frac{1}{4}(\alpha-1)^2 - \widehat{\psi}(\alpha-1-\widehat{\psi})\right) \\ + uv\frac{(\alpha+1)(\alpha-1)}{2}\end{aligned}.$$

A closely related interval is obtained with $\widetilde{\theta} = \theta$ and

$$\widetilde{\psi} = \widehat{\psi}(\gamma) = \alpha\frac{n_1\widehat{p}_1 + \gamma + 1}{n_1 + 2\gamma + 2} - \frac{n_2\widehat{p}_2 + \gamma + 1}{n_2 + 2\gamma + 2},$$

where $\gamma \geq -1$, which is the posterior mean of ψ using a prior proportional to $(p_1(1-p_1)p_2(1-p_2))^\gamma$ on (p_1, p_2). Interval (3) corresponds to $\gamma = -1$ and is referred to as the Haldane interval, because the prior on (p_1, p_2) corresponds to the product of Haldane priors (Haldane 1945), which gives most weight to extreme values 0 and 1. Different values of γ lead to various confidence intervals. Following Beal (1987), we consider the interval obtained with $\widetilde{\theta} = \theta$ and $\widetilde{\psi} = \widehat{\psi}(-1/2)$, which is referred to as the Jeffreys–Perks interval. These two intervals can suffer from overshoot.

2.3. Modified Wald interval

It is well known that the Wald confidence interval for a single binomial proportion p with nominal coverage 0.95 can have a coverage closer to the nominal value after adding to the data four-pseudo observations, comprising two successes and two failures, see Agresti and Coull (1998). The justification provided by Agresti and Coull for adding four observations comes from the Wilson interval, which has a coverage close to the nominal level, and a midpoint that is not $\widehat{p}$, the maximum likelihood estimate of p, but instead $(X + z^2/2)/(n + z^2)$, where X is the binomial variate and n the sample size. For 95% confidence intervals, the midpoint is approximately $(X + 2)/(n + 4)$. Agresti & Coull proposed to adjust the Wald interval by replacing the number of observations n by $n + 4$, and the number of successes X by $X + 2$. Surprisingly, doing so improves the coverage dramatically. In Agresti & Caffo (2000), the authors investigated the improvement made by using a similar trick for the construction of confidence intervals for the differences of two proportions, but without giving a justification for it, other than from a Bayesian point of view. Here we provide a general development that justifies the strategy for both the sum and the difference of independent proportions.

The Haldane interval for the weighted sum is obtained from the solutions of a quadratic equation. Suppose for simplicity that $n_1 = n_2 = n$ and $\alpha = 1$. The midpoint of the Haldane interval

$$\widehat{\theta}\left(\frac{1}{1+z^2u}\right) + \frac{z^2u}{1+z^2u}$$

falls between $\widehat{\theta}$ and 1. The midpoint, re-expressed in terms of X_1 and X_2 becomes

$$\frac{X_1 + X_2 + z^2/2}{n + z^2/2} \approx \frac{X_1 + X_2 + 2}{n + 2} = \frac{X_1 + 1}{n + 2} + \frac{X_2 + 1}{n + 2},$$

The square of the standard deviation to be added/subtracted to the midpoint is equal to

$$\frac{1}{n + z^2/2}\left((\widehat{p}_1(1 - \widehat{p}_1) + \widehat{p}_2(1 - \widehat{p}_2)) \frac{n}{n + z^2/2} + \frac{1}{2}(1 - \widehat{\psi}^2)\frac{z^2}{z^2 + 2n}\right),$$

which is a weighted average of two terms, the first term being the variance of the sum of two proportions where the sample size n is replaced by $n + z^2/2$. This observation provides a motivation to adjust the Wald interval (1), with $\widehat{p}_1 = (X_1 + 1)/(n_1 + 2)$, $\widehat{p}_2 = (X_2 + 1)/(n_2 + 2)$ and n_1 and n_2 replaced by $n_1 + 2$ and $n_2 + 2$.

This calculation is similar if considering the Haldane interval for the difference of two proportions. It provides a nice justification for using such a correction, which is similar to the justification given by Agresti & Coull for the single proportion. Note that the adjusted Wald interval still suffers from possible overshoot.

2.4. Score interval

The log-likelihood function may be expressed in terms of θ and p_1,

$$l(\theta, p_1) = x_1 \ln p_1 + (n_1 - x_1)\ln(1 - p_1) + x_2 \ln(\theta - \alpha p_1) + (n_2 - x_2)\ln(1 - \theta + \alpha p_1).$$

The score functions for θ and p_1 are, respectively,

$$S_\theta(\theta, p_1) = \frac{\partial l(\theta, p_1)}{\partial \theta} = \frac{x_2 - n_2 p_2}{p_2 q_2},$$

$$S_p(\theta, p_1) = \frac{\partial l(\theta, p_1)}{\partial p_1} = \frac{x_1 - n_1 p_1}{p_1 q_1} - \alpha\frac{x_2 - n_2 p_2}{p_2 q_2},$$

where we denote for simplicity $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. The variance and covariance of the score functions are given by

$$J_{\theta,\theta}(\theta, p_1) = -\mathbf{E}\left(\frac{\partial^2 l(\theta, p_1)}{\partial \theta^2}\right) = \frac{n_2}{p_2 q_2},$$

$$J_{p,p}(\theta, p_1) = -\mathbf{E}\left(\frac{\partial^2 l(\theta, p_1)}{\partial p_1^2}\right) = \frac{n_1}{p_1 q_1} + \alpha^2\frac{n_2}{p_2 q_2},$$

$$J_{\theta,p}(\theta, p_1) = -\mathbf{E}\left(\frac{\partial^2 l(\theta, p_1)}{\partial \theta \partial p_1}\right) = -\alpha\frac{n_2}{p_2 q_2}.$$

We wish to construct a confidence interval for θ regardless of p_1. To do so, we treat p_1 as a nuisance parameter, and we denote by $\widetilde{p}_1(\theta) = \arg\max_{p_1} l(\theta, p_1)$ the maximum likelihood of p_1 given θ. Assuming θ is known, p_1 lies in the open interval

$$I^{+}_{\theta,\alpha} = \left(\max\left(0, \frac{\theta-1}{\alpha}\right), \min\left(1, \frac{\theta}{\alpha}\right)\right).$$

if $\alpha \in (0,1]$, and in

$$I^{-}_{\theta,\alpha} = \left(\max\left(0, \frac{\theta}{\alpha}\right), \min\left(1, \frac{\theta-1}{\alpha}\right)\right).$$

if $\alpha \in [-1,0)$. The estimate $\widetilde{p}_1(\theta)$ satisfies $S_p(\theta, p_1) = 0$, which is a cubic equation in p_1: $Q(p_1) = 0$, with

$$Q(x) = b_3x^3 + b_2x^2 + b_1x + b_0, \quad x \in I^{\pm}_{\theta,\alpha}, \tag{4}$$

and

$$b_3 = \alpha^2(n_1 + n_2),$$

$$b_2 = \alpha(n_1(1-2\theta) - \alpha(n_2 + x_1) + x_2 - \theta n_2),$$

$$b_1 = x_1\alpha(2\theta - 1) + n_1\theta(\theta - 1) + \alpha(n_2\theta - x_2),$$

$$b_0 = \theta x_1(1-\theta).$$

This equation is found to have three real roots. We take $\widetilde{p}_1(\theta) = y_1$ (given in Appendix A) and $\widetilde{p}_2(\theta) = \theta - \alpha\widetilde{p}_1(\theta)$.

The statistic $S_\theta(\theta, \widetilde{p}_1(\theta))$ has an asymptotic normal distribution with variance (see, e.g. section 4.5 in Davison 2003)

$$\begin{aligned}\mathrm{Var}(S_\theta(\theta, \widetilde{p}_1(\theta))) &= J_{\theta,\theta}(\theta, \widetilde{p}_1(\theta)) - \frac{J^2_{\theta,p}(\theta, \widetilde{p}_1(\theta))}{J_{p,p}(\theta, \widetilde{p}_1(\theta))} \\ &= \left(\alpha^2 \frac{\widetilde{p}_1(\theta)(1-\widetilde{p}_1(\theta))}{n1} + \frac{\widetilde{p}_2(\theta)(1-\widetilde{p}_2(\theta))}{n_2}\right)^{-1} \equiv v_\alpha(\theta, \widetilde{p}_1(\theta)).\end{aligned}$$

We can therefore use the profile score function to derive a confidence interval, whose bounds are the solutions to

$$\frac{S^2_\theta(\theta, \widetilde{p}_1(\theta))}{v_{\alpha,\beta}(\theta, \widetilde{p}_1(\theta))} = \frac{(x_2 - n_2\widetilde{p}_2(\theta))^2}{(\widetilde{p}_2(\theta)\widetilde{q}_2(\theta))^2 v_\alpha(\theta, \widetilde{p}_1(\theta))} = z^2, \tag{5}$$

which can be solved numerically. We refer this interval to as the score interval.

Note that, following the method given by Mee (1984) for the construction of a confidence interval for the difference of two proportions, we can construct an interval for the sum $\theta = \alpha p_1 + p_2$, whose bounds are solutions to

$$\frac{(\widehat{\theta}-\theta)^2}{\alpha^2\frac{\widetilde{p}_1(\theta)(1-\widetilde{p}_1(\theta))}{n1}+\frac{\widetilde{p}_2(\theta)(1-\widetilde{p}_2(\theta))}{n_2}}=\left(\alpha\frac{x_1}{n_1}+\frac{x_2}{n_2}-\theta\right)^2 v_\alpha(\theta,\widetilde{p}_1(\theta))=z^2, \qquad (6)$$

where v_α is defined above in (5) and $\widetilde{p}_1(\theta)$ is the profile estimate of p_1 given θ. We show that this interval is identical to the score interval in Appendix B.

Also, following the idea of Miettinen & Nurminen (1985), we can obtain another interval by replacing z^2 with $z^2(n_1+n_2)/(n_1+n_2-1)$. We refer to the latter interval as the score interval with adjusted variance.

2.5. Likelihood-ratio interval

Finally, we invert a likelihood ratio test. The statistic is given by

$$\lambda(\theta)=-2\log\left(\frac{\widetilde{p}_1(\theta)}{\widehat{p}_1}\right)^{x_1}\left(\frac{1-\widetilde{p}_1(\theta)}{1-\widehat{p}_1}\right)^{n_1-x_1}\left(\frac{\widetilde{p}_2(\theta)}{\widehat{p}_2}\right)^{x_2}\left(\frac{1-\widetilde{p}_2(\theta)}{1-\widehat{p}_2}\right)^{n_2-x_2},$$

where $\widetilde{p}_1(\theta)$ and $\widetilde{p}_2(\theta)$ are the maximum likelihood estimators of p_1 and p_2 under the constraint $\alpha p_1 + p_2 = \theta$ (that is, $\widetilde{p}_1(\theta)$ is the root of Q, defined in (4)). It can be shown that $\lambda(\theta)$ has an asymptotic chi-squared distribution (see, e.g. section 4.5 in Davison 2003). Since the likelihood ratio compares the likelihood estimated at the maximum-likelihood estimate with the likelihood under the null hypothesis, we reject the null hypothesis when this ratio is too large, which provides a confidence interval for θ that is given by $\{\theta|\lambda(\theta)\geq z^2\}$.

2.6. Other methods

There are other methods for the difference of two proportions that we could adapt to the present setting. Newcombe (1998a) proposed a simple method for the difference that can be easily implemented for the sum. The lower and upper bounds are, respectively,

$$L=\widehat{\theta}-z\left(\alpha^2\frac{l_1(1-l_1)}{n_1}+\frac{l_2(1-l_2)}{n_2}\right)^{1/2} \text{ and}$$
$$U=\widehat{\theta}+z\left(\alpha^2\frac{u_1(1-u_1)}{n_1}+\frac{u_2(1-u_2)}{n_2}\right)^{1/2},$$

where l_i and u_i are the roots of $(\widehat{p}_i-p_i)^2=z^2(p_i(1-p_i)/n_i)$, $i = 1, 2$. We found that the coverage is close to and above the nominal level, but that this interval tended to be too conservative for small and large values of p_1 and p_2. In order to keep this paper brief, the results are not reported here.

There also exist exact intervals for the difference of two binomial proportions in the sense that the tails of the joint binomial distribution (likelihood) are used to compute the bounds of the confidence interval, instead of using a normal approximation. However, these methods tend to be conservative. Existing methods have been proposed by Santner & Yamagami (1993), Chan & Zhang (1999), Agresti & Min (2001) and Coe & Tamhane (1993), to cite but a few.

Finally bootstrap confidence intervals can be considered. These include, but are not limited to, one-sided and two-sided percentile bootstrap confidence intervals, symmetric and short bootstrap confidence intervals, see, for example, Hall (1992). We do not report in the next section a detailed numerical study for these intervals. However, we performed simulations in the case of the two-sided equal-tailed bootstrap confidence interval see Hall (1992, p. 87) for a precise definition of this interval), whose performance is similar to the Wald interval. Note that the performance of the bootstrap can be explained by considering an Edgeworth expansion of the statistic of interest. Specifically, when the underlying distribution is smooth, the bootstrap can be seen as a device for skewness correction, by removing the first error term to the normal approximation present in the asymptotic expansion. However, for lattice distributions, for example in the case of a binomial or a Poisson distribution, an additional discontinuous term of the same order as the skewness term is added to the Edgeworth expansion, which takes into account the continuity correction needed when approximating a lattice distribution with a continuous one. Therefore, care is needed when constructing confidence intervals for the sum of two binomial proportions since bootstrap methods suffer from the presence of this additional term, see for example Hall (1992, p. 91). However, procedures have been derived to overcome these difficulties, such as smoothed bootstrap methods, see, for example Hall (1987), Hall & Zhou (2003) and Zheng & Loh (1995).

3. Evaluation of the methods

New measures of performance based on a weighted average of the sensitivity and specificity in diagnostic studies require the construction of confidence intervals for sums of binomial proportions. In medical studies the number of patients available can be small, see, for example, the diagnostic tests reported in Di Nisio *et al.* (2010). In view of this, we provide a numerical study of the performance of the different methods presented in the previous section for small values of the sample sizes n_1 and n_2. Specifically, we derive the exact coverage and expected length for all $(n_1 + 1)(n_2 + 1)$ combinations of the various pairs (n_1, n_2) considered, for $\alpha = 0.05$.

Exact coverage for a pair (p_1, p_2) is given by

$$\mathcal{C}(p_1,p_2) = \sum_{(x_1,x_2)} \prod_{j=1}^{2} \binom{n_j}{x_j} p_j^{x_j}(1-p_j)^{n_j-x_j} \mathbf{1}_{\{L(\mathbf{x}) \le \alpha p_1 + p_2 \le U(\mathbf{x})\}}(x_1,x_2),$$

where $\mathbf{x} = (x_1, x_2)$, $\mathbf{1}_A$ is the indicator function of A, and $L(\mathbf{x})$ and $U(\mathbf{x})$ denote, respectively, the lower and upper bound of the confidence interval calculated if we observe $\mathbf{x}$. We show in Appendix C. that for all the methods presented in Section 2, we have

$$\mathcal{C}(p_1,p_2) = \mathcal{C}(1-p_1, 1-p_2). \tag{7}$$

The exact average length is

$$\mathcal{L}(p_1,p_2) = \sum_{(x_1,x_2)} \prod_{j=1}^{2} \binom{n_j}{x_j} p_j^{x_j}(1-p_j)^{n_j-x_j}(U(\mathbf{x}) - L(\mathbf{x})).$$

We also compute the probability that the lower bound is below min(0, α),

$$\mathcal{U}(p_1,p_2) = \sum_{(x_1,x_2)} \prod_{j=1}^{2} \binom{n_j}{x_j} p_j^{x_j}(1-p_j)^{n_j-x_j} \mathbf{1}_{\{L(\mathbf{x})<\min(0,\alpha)\}}(x_1,x_2),$$

and that the upper bound is above max(1, 1 + α),

$$\mathcal{O}(p_1,p_2) = \sum_{(x_1,x_2)} \prod_{j=1}^{2} \binom{n_j}{x_j} p_j^{x_j}(1-p_j)^{n_j-x_j} \mathbf{1}_{\{U(\mathbf{x})>\max(1,1+\alpha)\}}(x_1,x_2) = \mathcal{U}(1-p_1,1-p_2),$$

which are identical by symmetry (see Appendix C). We also compute the mean distance of the coverage to the nominal value 0.95

$$\mathcal{D}(p_1,p_2) = |\mathcal{C}(p_1,p_2) - 0.95|.$$

and keep track of the number of times that the coverage is below 0.93

$$\mathcal{T}(p_1,p_2) = \mathbf{1}_{\{\mathcal{C}(p_1,p_2)<0.93\}}(p_1,p_2).$$

Tables 1 and 2 present the average of the quantities outlined in the previous section, denoted, respectively, by $\mathcal{C}$, $\mathcal{L}$, $\mathcal{D}$, $\mathcal{T}$ and $\mathcal{U}$ for values of (p_1,p_2) chosen uniformly from the unit square, $(p_1,p_2) \in \Lambda$, where

$$\Lambda = \{(p_1,p_2)|\, p_i = 0.01k, k = 1,\ldots,99, i = 1,2\},$$

for $(n_1,n_2) = (20,10)$, (20, 20), (30, 20), and (50,20), (α, β) = (1, 1) (Table 1) and (0.8, 0.6) (Table 2). We give also the smallest coverage $\mathcal{S}$ returned by a method over all $(p_1,p_2) \in \Lambda$,

$$\mathcal{S} = \inf_{(p_1,p_2)\in\Lambda} \mathcal{C}(p_1,p_2).$$

The value $\mathcal{O}$ is not presented since, by symmetry, $\mathcal{O} = \mathcal{U}$. In Figure 1, we present plots of the actual coverage for various values of p_1 and p_2. We repeated this numerical study for other values of α, and found that the seven methods perform similarly to the numerical values provided here.

4. Discussion

From Tables 1 and 2 it can be seen that the Wald interval is, on average, below the nominal level and performs poorly for small sample sizes. Moreover, Figure 1 shows that this coverage is very erratic, and that the poor performance of this interval applies not only to extreme values of p_1 and p_2, but for many of the possible combinations. Oscillations could be explained, for example, by considering an Edgeworth expansion of the coverage probability. Adding one success and one failure to each observation greatly improves the mean coverage, which stays above 0.95 for small values of p_1 and p_2, without the interval being too conservative, except when p_1 is very small (large) and p_2 very large (small). Also, the proportion of intervals with coverage below 0.93 is dramatically

TABLE 1

Coverage, length, distance and proportions of cases when the coverage is below 0.93, shortest coverage, and proportion of lower bound below 0, for (p_1, p_2) uniformly chosen in the unit square for a nominal coverage equal to 0.95, with $\alpha = 1$.

(n_1, n_2)	W	AW	H	JP	S	AS	L
				(20,10)			
$\mathcal{C}$	0.902	0.959	0.938	0.951	0.954	0.957	0.939
$\mathcal{L}$	0.577	0.590	0.557	0.573	0.583	0.593	0.575
$\mathcal{D}$	0.048	0.011	0.015	0.009	0.009	0.009	0.014
$\mathcal{T}$	0.925	0.003	0.154	0.036	0.003	0.001	0.183
$\mathcal{S}$	0.260	0.916	0.260	0.874	0.923	0.923	0.865
$\mathcal{U}$	0.034	0.012	0	0.004	0	0	0
				(20,20)			
$\mathcal{C}$	0.925	0.956	0.941	0.949	0.949	0.952	0.942
$\mathcal{L}$	0.485	0.490	0.471	0.478	0.485	0.491	0.483
$\mathcal{D}$	0.025	0.007	0.010	0.006	0.007	0.006	0.010
$\mathcal{T}$	0.414	$8{\cdot}10^{-4}$	0.091	0.023	0.013	$8{\cdot}10^{-4}$	0.076
$\mathcal{S}$	0.331	0.924	0.453	0.899	0.917	0.922	0.858
$\mathcal{U}$	0.019	0.006	0	0	0	0	0
				(30,20)			
$\mathcal{C}$	0.928	0.955	0.944	0.950	0.950	0.952	0.943
$\mathcal{L}$	0.444	0.449	0.434	0.440	0.445	0.450	0.443
$\mathcal{D}$	0.022	0.007	0.007	0.005	0.005	0.005	0.008
$\mathcal{T}$	0.331	$4{\cdot}10^{-4}$	0.060	0.011	0.001	$8{\cdot}10^{-4}$	0.049
$\mathcal{S}$	0.395	0.927	0.394	0.899	0.928	0.928	0.872
$\mathcal{U}$	0.013	0.004	0	0.001	0	0	0
				(50,20)			
$\mathcal{C}$	0.926	0.955	0.945	0.950	0.952	0.954	0.944
$\mathcal{L}$	0.406	0.411	0.399	0.404	0.407	0.410	0.404
$\mathcal{D}$	0.024	0.007	0.007	0.005	0.004	0.005	0.008
$\mathcal{T}$	0.431	$4{\cdot}10^{-4}$	0.051	0.007	$6{\cdot}10^{-4}$	$6{\cdot}10^{-4}$	0.043
$\mathcal{S}$	0.454	0.929	0.504	0.891	0.927	0.927	0.886
$\mathcal{U}$	0.009	0.004	0.001	0.002	0	0	0

Notes: Methods tested are: W, Wald interval; AW, Adjusted Wald; H, Haldane interval; JP, JP interval; S, Score interval; AS, Score with adjusted variance; L, Likelihood ratio.

TABLE 2

Coverage, length, distance and proportions of cases when the coverage is below 0.93, shortest coverage, and proportion of lower bound below 0 for (p_1, p_2) uniformly chosen in the unit square for a nominal coverage equal to 0.95, with $\alpha = 0.7$.

(n_1, n_2)	W	AW	H	JP	S	AS	L
(20,10)							
$\mathcal{C}$	0.886	0.960	0.938	0.952	0.955	0.958	0.941
$\mathcal{L}$	0.518	0.535	0.503	0.519	0.523	0.531	0.516
$\mathcal{D}$	0.064	0.013	0.018	0.012	0.008	0.009	0.016
$\mathcal{T}$	0.961	0.004	0.162	0.037	0.005	0.004	0.218
$\mathcal{S}$	0.260	0.906	0.396	0.874	0.897	0.903	0.855
$\mathcal{U}$	0.038	0.016	0.004	0.008	0	0	0
(20,20)							
$\mathcal{C}$	0.922	0.956	0.942	0.949	0.951	0.954	0.942
$\mathcal{L}$	0.417	0.422	0.405	0.412	0.418	0.424	0.415
$\mathcal{D}$	0.028	0.008	0.010	0.006	0.006	0.006	0.010
$\mathcal{T}$	0.491	0.002	0.086	0.016	0.001	$2{\cdot}10^{-4}$	0.094
$\mathcal{S}$	0.331	0.923	0.453	0.876	0.925	0.928	0.843
$\mathcal{U}$	0.019	0.006	0	0.002	0	0	0
(30,20)							
$\mathcal{C}$	0.922	0.956	0.944	0.950	0.952	0.955	0.943
$\mathcal{L}$	0.393	0.399	0.385	0.391	0.395	0.398	0.391
$\mathcal{D}$	0.028	0.008	0.008	0.006	0.005	0.006	0.009
$\mathcal{T}$	0.580	$2{\cdot}10^{-4}$	0.062	0.011	$2{\cdot}10^{-4}$	0	0.065
$\mathcal{S}$	0.395	0.927	0.553	0.892	0.929	0.930	0.844
$\mathcal{U}$	0.014	0.005	0	0.002	0	0	0
(50,20)							
$\mathcal{C}$	0.917	0.956	0.945	0.951	0.954	0.955	0.944
$\mathcal{L}$	0.372	0.378	0.366	0.372	0.373	0.375	0.369
$\mathcal{D}$	0.033	0.010	0.008	0.007	0.005	0.005	0.009
$\mathcal{T}$	0.732	0.001	0.052	0.007	0.001	0.001	0.080
$\mathcal{S}$	0.454	0.925	0.701	0.891	0.913	0.913	0.848
$\mathcal{U}$	0.010	0.005	0.001	0.002	0	0	0

Notes: Methods tested are: W, Wald interval; AW, Adjusted Wald; H, Haldane interval; JP, JP interval; S, Score interval; AS, Score with adjusted variance; L, Likelihood ratio.

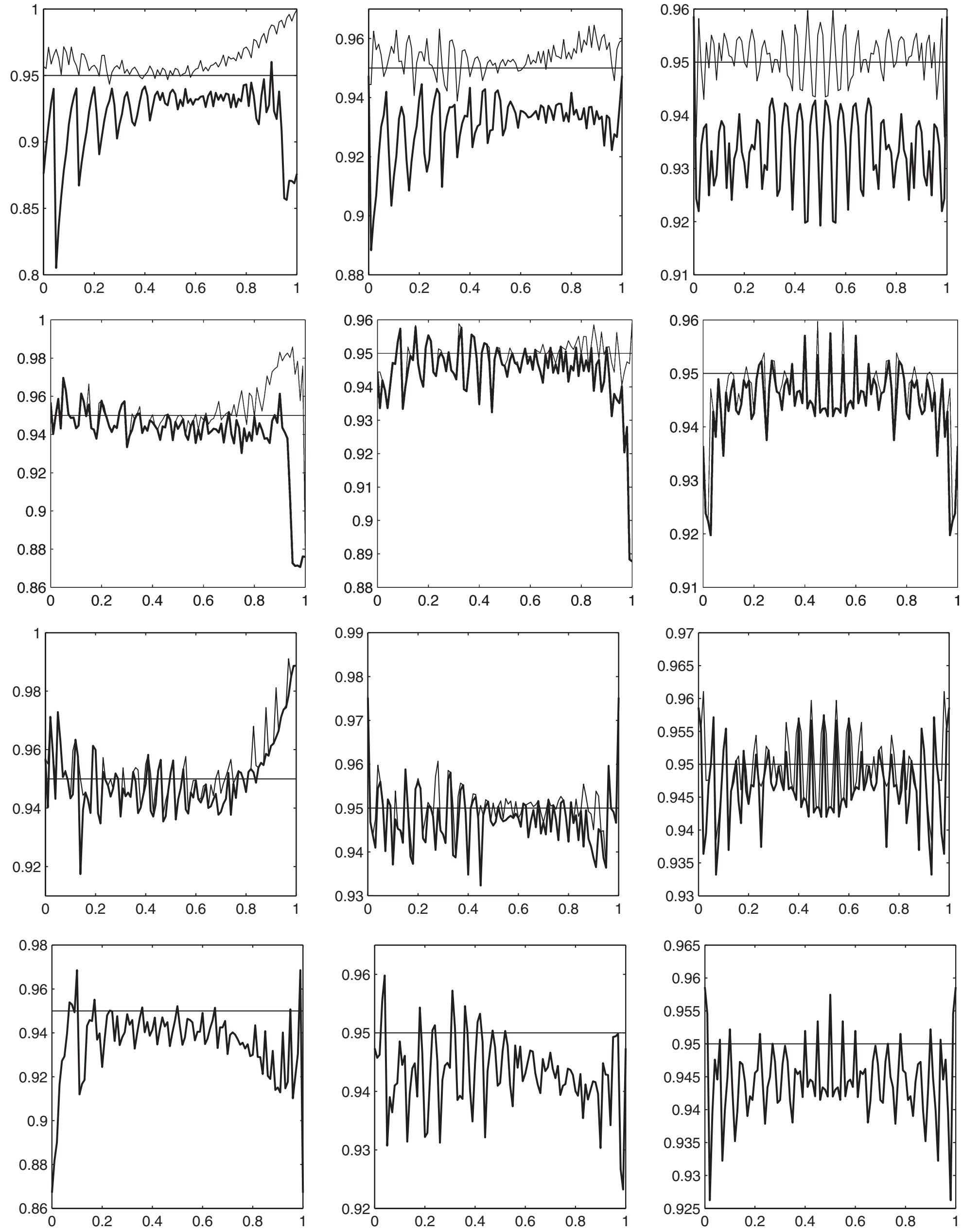

Figure 1. Coverage probability for the seven methods with $(n_1, n_2) = (20, 20)$ and $\alpha = 1$. The left, middle and right column correspond, respectively, to $p_1 = 0.1, 0.3$ and 0.5 and p_2 varying from 0 to 1. Top row are Wald interval (bold) and adjusted Wald (thin). Second row are Haldane (bold) and Jeffreys–Perks (thin) methods. Third row are score (bold) and score with adjusted variance method (thin). Bottom row corresponds to the likelihood ratio method.

reduced and close to zero even for very small sample sizes. However, the Wald intervals suffer from overshoot when x_i are both small or large and should not be used in these cases. The Wald interval performs similarly to the case of the single proportion, see Newcombe (1998a); Brown, Cai & DasGupta (2002), or the difference of two proportions, see Newcombe (1998b).

The Haldane and Jeffreys–Perks intervals also returned coverages close to the nominal value, the latter achieving better coverage in most cases and keeping $\mathcal{T}$ small. We can see in Figure 1 that the Jeffreys–Perks method provides better coverage than the Haldane interval when p_1 is very small (large) and p_2 very large (small). The overshoot is much reduced compared to the Wald intervals. The Jeffreys–Perks intervals are also, on average, shorter than the Wald intervals.

Intervals obtained with the score method are close to the nominal level, having the smallest $\mathcal{T}$ values among all six methods. It can be seen from Figure 1 that these two intervals tend to be conservative for large (small) values of p_1 and small (large) values of p_2. A nice feature of these two methods is that they do not return confidence intervals whose bounds are outside the possible range of values for θ, unlike the previous methods. Except for extreme cases where $x_1 = n_1$ and $x_2 = n_2$ or $x_1 = x_2 = 0$, the limits do not correspond to the point estimate $\widehat{\theta}$, which is a desirable property. Moreover, only a very small proportion of intervals have exact coverage below the value 0.93, a proportion which is further reduced when we adjust the variance.

We show in Appendix B that the bounds of the score interval satisfying (5) can be obtained by solving (6). During our calculations, it seems that (6) led to fewer numerical issues and we recommend the practitioner to implement the latter equation in order to get the bounds of the score interval.

Finally, the likelihood ratio-based interval does not produce bounds outside the possible values of θ, but suffers from poor coverage and a non-negligible proportion of intervals with exact coverage below the value 0.93. It is known that confidence intervals based on the likelihood ratio test for the difference of two proportions also have coverage below the nominal level, see Miettinen & Nurminen (1985); Newcombe (1998a). This method for the sum of two proportions should also be avoided.

In large sample situations, for example when $n_1 \geq 100{,}000$, and $n_2 \geq 1000$, such as in the application given in the next section, the normal approximation is very accurate and the seven methods perform well. Specifically, all intervals, including the Wald interval, have close to the nominal coverage probability, with a mean distance $\mathcal{D}$ of order no larger than 10^{-3}, and do not present any case where the coverage probability drops below 0.93. Moreover, there is no difference between the score and adjusted score methods as the factor in front of z is very close to 1.

As a conclusion, we recommend using the score interval with adjusted variance in small sample situations, unless simplicity of calculation is important, in which case we advocate the Jeffreys–Perks interval.

5. An application

We now use the score interval with adjusted variance to determine a confidence interval for the proportion of contaminated mail items passing inspection at the border. The data in Table 3 were kindly provided by DAFF Biosecurity. The data are the outcomes of

TABLE 3

Results of quarantine inspection of certain classes of mail items for 12 months at all mail facilities within Australia.

Pathway	N	n_1	x_1	n_2	x_2	$\widehat{\theta}(\%)$	$\widehat{\theta}_L(\%)$	$\widehat{\theta}_U(\%)$
EMS	3,628,993	3,059,169	5108	10,357	5	0.0744	0.0488	0.1374
Other articles	47,300,154	28,088,067	7071	31,537	9	0.0387	0.0274	0.0671
Parcels	3,196,962	2,862,399	7919	12,288	10	0.1100	0.0742	0.1810
Registered	845,007	748,559	139	4162	2	0.0499	0.0000	0.1749

Notes: N is the number of mail items in the pathway, and n_1 of them are inspected by x-ray or detector dogs, with x_1 items intercepted as having high biosecurity risk material. A manual leakage survey is performed in which n_2 items are inspected from all non-intercepted items, with x_2 items intercepted as having high biosecurity risk material. The estimate $\widehat{\theta}$ is the pathway-level leakage rate, defined in the introduction.

the quarantine inspection of certain types of mail articles for 12 months across all mail facilities in Australia. Inspection is performed by one of two instruments: x-ray and detector dogs. The leakage survey for these four pathways is performed by random selection followed by physically opening and checking the contents.

Our results provide useful insight into the relative biosecurity contamination rate of the four pathways. The approaching contamination rates are comparable for registered articles and other articles, slightly but not substantially higher for Express Mail Service (EMS), and slightly higher again for parcels. The interval estimates provide comfort that the low estimates of biosecurity contamination rate are statistically defensible.

Appendix A: Maximum likelihood estimate for the score method

Lemma 1. Let $-1 \leq \alpha \leq 1$, $\alpha \neq 0$, $n_i \geq 1$, and $x_i \in \{0,\ldots,n_i\}$, for $i = 1, 2$. Let $\theta \in (0,1+\alpha)$ if $\alpha > 0$ and $\theta \in (\alpha,1)$ if $\alpha < 0$. Then Q has three real roots, whose expressions are

$$y_i = 2p\cos(c_i) - b_2/(3b_3), \quad i = 1,2,3,$$

with

$$\begin{aligned} c_1 &= \left(\pi + \cos^{-1}(q/p^3)\right)/3, \\ c_2 &= \left(-\pi + \cos^{-1}(q/p^3)\right)/3, \\ c_3 &= \frac{1}{3}\cos^{-1}(q/p^3), \\ p &= \pm\left(b_2^2/(3b_3)^2 - b_1/(3b_3)\right)^{1/2}, \\ q &= b_2^3/(3b_3)^3 - b_1b_2/(6b_3^2) + b_0/(2b_3), \end{aligned}$$

where the sign of p is chosen so that p and q have the same sign. Moreover,

(i) If $x_1 \neq 0$, $x_1 \neq n_1$, $x_2 \neq 0$, and $x_2 \neq n_2$, there is a unique root in $I^{\pm}_{\theta,\alpha}$.

(ii) In any other case, there is at least one root in $\overline{I}^{\pm}_{\theta,\alpha}$, where $\overline{I}^{\pm}_{\theta,\alpha}$ denotes the closure of $I^{\pm}_{\theta,\alpha}$, where $I^{\pm}_{\theta,\alpha}*$ represents either $I^{+}_{\theta,\alpha}$ or $I^{-}_{\theta,\alpha}$, defined above formula (4). If there is more than one root in $\overline{I}^{\pm}_{\theta,\alpha}$, then at most one lies in $I^{\pm}_{\theta,\alpha}$, and the other one(s) are at (one of) the end points of $I^{\pm}_{\theta,\alpha}$.

Proof. Suppose $\alpha > 0$. Let $\theta \in (0,1 + \alpha)$. We have

$$\begin{aligned} Q(0) &= \theta x_1(1-\theta), \\ Q(1) &= (\alpha-\theta)(\alpha+1-\theta)(n_1-x_1), \\ Q((\theta-1)\alpha) &= \frac{1}{\alpha}(\theta-1)(\alpha+1-\theta)(n_2-x_2), \\ Q(\theta/\alpha) &= \theta(\theta-\alpha)x_2/\alpha, \end{aligned}$$

Suppose first that $x_1 \neq 0$, $x_1 \neq n_1$, $x_2 \neq 0$, $x_2 \neq n_2$.

(i) If $0 < \theta < \alpha$, then $I_{\theta,\alpha} = (0, \theta/\alpha)$, $Q(0) > 0$, $Q(1) > 0$, $Q(\theta/\alpha) < 0$ and $Q((\theta-1)\alpha) < 0$ so that one root lies in $(-\infty, 0)$, one in $I_{\theta,\alpha}$ and one in $(\theta/\alpha, 1)$.

(ii) If $\theta = \alpha$, then $I_{\theta,\alpha} = (0,1)$, and $Q(x)$ can be expressed as $Q(x) = \alpha(x-1)P(x)$, where

$$P(x) = \alpha(n_1+n_2)x^2 + ((n_1+x_2) - \alpha(x_1+n_1+n_2))x + (\alpha-1)x_1,$$

with $P(0) < 0$ and $P(1) = n_1 - x_1 + x_2 > 0$. Therefore P had one negative root and one root in (0,1). It follows that Q has one root in $(-\infty,0)$, one in $I_{\theta,\alpha}$ and its third root is 1.

(iii) If $\alpha < \theta < 1$, then $I_{\theta,\alpha} = (0,1)$, $Q(0) > 0$, $Q(1) < 0$, $Q(\theta/\alpha) > 0$ and $Q((\theta-1)\alpha) < 0$ so that one root lies in $(-\infty, 0)$, one in $I_{\theta,\alpha}$ and one in $(1, \theta/\alpha)$.

(iv) If $\theta = 1$, then $I_{\theta,\alpha} = (0,1)$ and $Q(x) = \alpha x R(x)$, with

$$R(x) = \alpha(n_1+n_2)x^2 + ((x_2-n_2) - n_1 - \alpha(n_2+x_1))x + n_2 - x_2 + x_1,$$

with $R(0) > 0$ and $R(1) = (\alpha-1)(n_1-x_1) < 0$. Thus R has one root in (0,1) and one root in $(1,\infty)$. Since $Q(\theta/\alpha) > 0$, it follows that Q has one root in $(1,\theta/\alpha)$, one in $I_{\theta,\alpha}$ and the third root is 0.

(v) If $1 < \theta < 1 + \alpha$, then $I_{\theta,\alpha} = ((\theta-1)/\alpha, 1)$, $Q(0) < 0$, $Q(1) < 0$, $Q(\theta/\alpha) > 0$ and $Q((\theta-1)\alpha) > 0$ so that one root lies in $(0,(\theta-1)/\alpha)$, one in $I_{\theta,\alpha}$ and one in $(1,\theta/\alpha)$.

In summary, Q has a unique root in $I_{\theta,\alpha}$ for all $\theta \in (0,1+\alpha)$ provided $x_1 \neq 0$, $x_1 \neq n_1$, $x_2 \neq 0$, and $x_2 \neq n_2$.

When x_1 and/or x_2 take extreme values, Q can be expressed as a product of a first order polynomial (since 0, 1, $(\theta-1)/\alpha$ and/or θ/α are obvious roots of Q in this case) with a quadratic polynomial whose sign at 1, $(\theta-1)/\alpha$ and θ/α permit us to locate the position of the remaining roots. The details are not presented here and are left to the reader. The cases $\alpha \in (-1, 0)$ and $\alpha = \pm 1$ can be treated similarly. Trigonometric expressions for the three real roots of Q can be found e.g. in Bronshtein *et al.* (2007). A numerical study shows that the value of $\widetilde{p}_1(\theta)$ corresponds to root y_1.

Appendix B: Comparison of two methods

Following the method given by Mee (1984) for the construction of a confidence interval for the difference of two proportions, we can construct an interval for the sum $\theta = \alpha p_1 + p_2$, whose bounds are solutions to

$$\frac{(\widehat{\theta} - \theta)^2}{\alpha^2 \frac{\widetilde{p}_1(\theta)(1-\widetilde{p}_1(\theta))}{n1} + \frac{\widetilde{p}_2(\theta)(1-\widetilde{p}_2(\theta))}{n_2}} = \left(\alpha\frac{x_1}{n_1} + \frac{x_2}{n_2} - \theta\right)^2 v_\alpha(\theta, \widetilde{p}_1(\theta)) = z^2,$$

where v_α was defined in Section 2.4 and $\widetilde{p}_1(\theta)$ is the profile estimate of p_1 given θ. We show that this interval is identical to the score interval. Since $\widetilde{p}_1(\theta)$ is a solution to $S_1(\theta, p_1) = 0$, we get

$$\alpha\frac{x_1}{n_1} - \theta = -\widetilde{p}_2(\theta) + \alpha^2 \frac{\widetilde{p}_1(\theta)\widetilde{q}_1(\theta)(x_2 - n_2\widetilde{p}_2(\theta))}{n_1\widetilde{p}_2(\theta)\widetilde{q}_2(\theta)},$$

which we substitute back in to (6) to get, after simple algebra,

$$\frac{(x_2 - n_2\widetilde{p}_2(\theta))^2}{(\widetilde{p}_2(\theta)\widetilde{q}_2(\theta))^2 v_\alpha(\theta, \widetilde{p}_1(\theta))},$$

which is exactly (5).

Appendix C: Proof of Equation (7)

We provide here a justification for why $\mathcal{C}(p_1, p_2) = \mathcal{C}(1 - p_1, 1 - p_2)$ and $\mathcal{O}(p_1, p_2) = \mathcal{U}(1 - p_1, 1 - p_2)$ hold. Let $\mathbf{x} = (x_1, x_2)$ and $\mathbf{n} - \mathbf{x} = (n_1 - x_1, n_2 - x_2)$. Clearly

$$\begin{aligned}
&\mathcal{C}(1 - p_1, 1 - p_2)\\
&\quad = \sum_{(x_1,x_2)} \prod_{j=1}^{2} \genfrac{}{}{0pt}{}{n_j}{x_j} (1 - p_j)^{x_j} p_j^{n_j - x_j} \mathbf{1}_{\{L(\mathbf{x}) \le \alpha(1-p_1)+1-p_2 \le U(\mathbf{x})\}}(\mathbf{x})\\
&\quad = \sum_{(x_1,x_2)} \prod_{j=1}^{2} \genfrac{}{}{0pt}{}{n_j}{x_j} p_j^{x_j} (1 - p_j)^{n_j - x_j} \mathbf{1}_{\{1+\alpha-U(\mathbf{n}-\mathbf{x}) \le \alpha p_1 + p_2 \le 1+\alpha-L(\mathbf{n}-\mathbf{x})\}}(\mathbf{n} - \mathbf{x}),
\end{aligned}$$

where we made the change of variable $x_j = n_j - x_j$, j = 1, 2 and used $\binom{n_j}{n_j - x_j} = \binom{n_j}{x_j}$. The equality $\mathcal{C}(p_1, p_2) = \mathcal{C}(1 - p_1, 1 - p_2)$ follows if we can show that

$$\mathbf{1}_{\{1+\alpha-U(\mathbf{n}-\mathbf{x}) \le \alpha p_1 + p_2 \le 1+\alpha-L(\mathbf{n}-\mathbf{x})\}}(\mathbf{n} - \mathbf{x}) = \mathbf{1}_{\{L(\mathbf{x}) \le \alpha p_1 + p_2 \le U(\mathbf{x})\}}(\mathbf{x}).$$

In other words, we want to show that $L(\mathbf{x}) = 1 + \alpha - U(\mathbf{n} - \mathbf{x})$ and $U(\mathbf{x}) = 1 + \alpha - L(\mathbf{n} - \mathbf{x})$, from which $\mathcal{O}(p_1, p_2) = \mathcal{U}(1 - p_1, 1 - p_2)$ would follow as well. The Wald/adjusted Wald/Haldane/Jeffreys–Perks intervals have an explicit expression, and it is just a straightforward calculation to check that the previous equalities hold. For intervals based on the score function, changing x_j to $n_j - x_j$ changes the score function for p_1 as follows

$$S_p(\theta, p_1) = -\left(\frac{x_1 - n_1 q_1}{p_1 q_1} - \alpha \frac{x_2 - n_2 q_2}{p_2 q_2}\right),$$

so that $\widetilde{p}_1(\theta)$ and $\widetilde{p}_2(\theta)$, the maximum likelihood estimates of p_1 and p_2 under the constraint $\alpha p_1 + p_2 = \theta$ when we observe x_1 and x_2, become $1 - \widetilde{p}_1(\theta)$ and $1 - \widetilde{p}_2(\theta)$ when observing $n_1 - x_1$ and $n_2 - x_2$.

Suppose we observe x_1 and x_2. The lower bound of the confidence interval $L(\mathbf{x})$ is such that $\alpha\widetilde{p}_1(L(\mathbf{x})) + \widetilde{p}_2(L(\mathbf{x})) = L(\mathbf{x})$. From the remark above, if we now observe $n_1 - x_1$ and $n_2 - x_2$, then one have $\alpha(1 - \widetilde{p}_1(L(\mathbf{x}))) + 1 - \widetilde{p}_2(L(\mathbf{x})) = 1 + \alpha - L(\mathbf{x})$, which corresponds to the upper bound of the confidence interval and $L(\mathbf{x}) = 1 + \alpha - U(\mathbf{n} - \mathbf{x})$ follows.

References

AGRESTI, A. & COULL, B.A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *Amer. Statist.* **52**, 119–126.

AGRESTI, A. & CAFFO, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *Amer. Statist.* **54**, 280–288.

AGRESTI, A. & MIN, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **57**, 963–971.

ANBAR, D. (1983). On estimating the difference between two probabilities, with special reference to clinical trials. *Biometrics* **39**, 257–262.

BEAL, S.L. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics* **43**, 941–950.

BRONSHTEIN, J.N., SEMENYAYEV, K.A., MUSIOL G. & MUEHL H. (2007). *Handbook of Mathematics.* 5th edn. Berlin: Springer.

BROWN, L.D., CAI, T. & DASGUPTA, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.* **30**, 160–201.

COE, P.R. & TAMHANE, A.C. (1993). Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities.. *Comm. Statist. Simulation Comput.* **22**, 925–938.

CHAN, I.S.F. & ZHANG, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* **55**, 1201–1209.

DAVISON, A.C. (2003). *Statistical Models.* Cambridge Cambridge University Press.

HALDANE, J.B.S. (1945). On a method of estimating frequencies. *Biometrika* **33**, 222–225.

HALL, P. (1992). The Bootstrap and Edgeworth Expansion. New York: Springer-Verlag.

HALL, P. (1987). On the bootstrap and continuity correction. *J. R. Statist. Soc. Ser. B Stat. Methodol.* **49**, 82–89.

HALL, P. & ZHOU, X-H. (2003). Effects of smoothing on distribution approximations. *Lect. Notes-Monogr. Ser. Probab., Stat. Appl: Papers in Honor of Rabi Bhattacharya* **41**, 169–186.

MEE, R. (1984). Confidence bounds for the difference between two probabilities. *Biometrics* **40**, 1175–1176.

MIETTINEN, O.S. & NURMINEN, M. (1985). Comparative analysis of two rates. *Statist. Med.* **4**, 213–226.

NEWCOMBE, R.G. (1998a). Interval estimation for the difference between independent proportions:comparison of eleven methods. *Statist. Med.* **17**, 873–890.

NEWCOMBE, R.G. (1998b). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statist. Med.* **17**, 857–872.

NEWCOMBE, R.G. (2001). Simultaneous comparison of sensitivity and specificity of two tests in the paired design: a straightforward graphical approach. *Statist. Med.* **20**, 907–915.

DI NISIO, M., VAN SLUISI, G.L., BOSSUYT, M.M., BÜLLER, H.R., PORRECA, E. & RUTJES, A.W.S. (2010). Accuracy of diagnostic tests for clinically suspected upper extremity deep vein thrombosis: a systematic review. *J. Thromb. Haemost.* **8**, 684–692.

SANTNER, T.J. & YAMAGAMI, S. (1993). Invariant small sample confidence intervals for the difference of two success probabilities.. *Comm. Statist. Simulation Comput* **22**, 33–59.

STEYERBERG, E.W., VICKERS, A.J., COOK, N.R., GERDS, T., GONEN, M., OBUCHOWSKI, N., PENCINA, M.J., KATTAN, M.W. (2010). Assessing the performance of prediction models. A framework for traditional and novel measures. *Epidemiology* **21**, 128–138.

VACH, W., GERKE, O. & HØILUND-CARLSEN P.F., (2012). Three principles to define the success of a diagnostic study could be identified. *J. Clinical Epidemiol.* **65**, 293–300.

VICKERS, A.J. & ELKIN, E.B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* **26**, 565–574.

VICKERS, A.J., CRONIN, A.M., ELKIN, E.B. & GONEN, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med. Inform. Decis. Making* **53**.

WILSON, E.B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *J. Amer. Statist. Assoc.* **22**, 209–212.

ZHOU, X-H., TSAO, M. & QIN, G. (2004). New intervals for the difference between two independent binomial proportions. *J. Statist. Plann. Inference* **123**, 97–115.

ZHENG, X. & LOH, W-Y. (1995). Bootstrapping binomial confidence intervals. *J. Statist. Plann. Inference* **43**, 355–380.