Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Inference for a binomial proportion in the presence of ties

Paul H. Garthwaite [a] & John R. Crawford [b]

[a] Department of Mathematics and Statistics, The Open University, Milton Keynes, MK7 6AA, UK

[b] School of Psychology, University of Aberdeen, Aberdeen, AB24 2UB, UK

Available online: 02 Aug 2011

To cite this article: Paul H. Garthwaite & John R. Crawford (2011): Inference for a binomial proportion in the presence of ties, Journal of Applied Statistics, 38:9, 1915-1934

To link to this article: http://dx.doi.org/10.1080/02664763.2010.537649

Journal of Applied Statistics
Vol. 38, No. 9, September 2011, 1915–1934

Taylor & Francis
Taylor & Francis Group

Inference for a binomial proportion in the presence of ties

Paul H. Garthwaite[a]* and John R. Crawford[b]

[a]*Department of Mathematics and Statistics, The Open University, Milton Keynes MK7 6AA, UK;* [b]*School of Psychology, University of Aberdeen, Aberdeen AB24 2UB, UK*

(*Received 03 November 2009; final version received 15 October 2010*)

We suppose a case is to be compared with controls on the basis of a test that gives a single discrete score. The score of the case may tie with the scores of one or more controls. However, scores relate to an underlying quantity of interest that is continuous and so an observed score can be treated as the rounded value of an underlying continuous score. This makes it reasonable to break ties. This paper addresses the problem of forming a confidence interval for the proportion of controls that have a lower underlying score than the case. In the absence of ties, this is the standard task of making inferences about a binomial proportion and many methods for forming confidence intervals have been proposed. We give a general procedure to extend these methods to handle ties, under the assumption that ties may be broken at random. Properties of the procedure are given and an example examines its performance when it is used to extend several methods. A real example shows that an estimated confidence interval can be much too small if the uncertainty associated with ties is not taken into account. Software implementing the procedure is freely available.

Keywords: coverage; Clopper–Pearson interval; credible interval; discrete distribution; multiple ties; Wald interval

1. Introduction

The problem that motivated the present research arises commonly in the practice of medicine, psychology and education. Scores $y_1, \ldots, y_n$ are obtained for a random sample of n people from a population of controls. A case has a score $y^{\#}$ and the case is to be compared with the controls. Scores are discrete so $y^{\#}$ may equal one or more of the y_i. However, the scores relate to an underlying quantity-of-interest that is continuous so an observed score can be treated as the rounded value of some underlying continuous score. This means that it is reasonable and desirable to break ties. For example, on a memory test or on an anxiety scale a case's score may equal the scores of some controls, while the case's underlying memory ability or anxiety level does not exactly equal the memory ability or anxiety level of any control, as both these constructs are assumed to be continuous variables. The task we address is to make inferences about the

*Corresponding author. Email: p.h.garthwaite@open.ac.uk

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664763.2010.537649
http://www.informaworld.com

proportion of controls, p say, who have a lower underlying score than the case. In the absence of ties, this task is essentially the standard problem of making inferences for a binomial proportion: inferences are to be made about the probability of success (p) on the basis of n trials, where the ith trial is a success if y_i is less than $y^{\#}$ and it is a failure if y_i is greater than $y^{\#}$.

To make inferences about p when ties are present, some assumption (or assumptions) about ties is necessary. We make the assumption that, in breaking ties, all possible results from tie-breaking are equally likely. Specifically, suppose the score of the case equals that of t controls. If the case and these controls were given further tests so as to break all ties, then the case and controls could be ranked from 1 to $t+1$. Our assumption is that the case is equally likely to have any one of these ranks. When this is true, we say that the ties satisfy the *random-ranking assumption*. More precisely, let S denote the number of controls in the sample whose scores are less than $y^{\#}$ and let T denote the number whose scores tie with $y^{\#}$. Also, let X denote the (unknown) number whose *underlying* score is less than y. Then ties satisfy the random-ranking assumption if

$$\Pr(X = x | S = s, T = t) = \begin{cases} \dfrac{1}{t+1} & \text{for } x = s, s+1, \ldots, s+t. \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Typically, this assumption is reasonable unless the case has the joint-lowest (or joint-highest) score, when it may be clear from circumstances that the case is quite likely to have a more extreme underlying score than any of the controls. To illustrate the latter, suppose psychological tests are given to a case because the case has suffered a head injury. If the case gets the minimum possible score on a test, there may be a distinct possibility that brain injury is affecting the case's performance and that the case's underlying score is well outside the normative range. Even if a few controls also obtain the minimum possible score, this may still seem likely. In such situations, the methods proposed here should not be used. Alternative strategies include using Bayesian methods with informative prior distributions. This approach has been used to handle zero binomial outcomes [9,16] and might be extended to handle ties.

Difficulties caused by ties have been considered in the context of hypothesis testing, but have not been addressed in the context of confidence interval estimates. In the hypothesis testing, common ways of handling ties are either (a) to ignore them and reduce the sample size, or (b) divide the ties in two, treating half of them as successes and the other half as failures [5,8,11]. Approach (a) can be sensible, as, for example, in problems where McNemar's test should be used. However, it is inappropriate when forming a confidence interval for the proportion of controls who have a smaller underlying score than the case, or other questions addressed here. Approach (b) is an over-simplification. It ignores the uncertainty that arises from ties and can lead to confidence intervals that are too narrow. Our assumption leads to methods that are clearly better than (a) or (b) for the contexts we consider.

A substantial body of work concerns the sign test and its treatment of ties [4,12,14]. Rather than p, this work focuses on slightly different parameters: p_+, p_0 and p_-, where these denote the probability of a success, the probability of a tie, and the probability of a failure, respectively. The aim has been to test whether p_+ equals p_-, rather than to make inferences about $p = p_+ + (1/2)p_0$, focusing on the asymptotic efficiencies of different test procedures. The work uses approaches (a) and (b) for handling ties and also various other approaches. For example, Woodbury *et al.* [17] suppose that a sequence of paired trials are conducted to compare two treatments. The result of each trial may favour one of the treatments or it may be a tie, and each tie is broken at random. If there are t ties, then this situation has t separate ties where each tie involves two items, rather than the situation we consider, where there is a multiple tie between t controls and a case. This latter situation does not seem to have been considered before.

We address the task of forming confidence interval for a binomial proportion when there are ties and these ties satisfy the random-ranking assumption. The simpler case in which there are no ties has attracted much attention. This is largely because problems of discreteness result in confidence intervals that seldom have the nominal coverage probability. That is, confidence intervals for a binomial proportion whose *nominal* coverage probability is, say, 0.95 (i.e. 95% confidence intervals) seldom have an *actual* coverage probability of 0.95. Some of the methods that have been proposed for forming confidence intervals can guarantee having a coverage that is no smaller than the nominal probability. However, these methods have been criticized as being too conservative, erring too heavily on the side of caution. Other methods may not guarantee being conservative, but may be conservative for most values of p and may typically give coverage probabilities that are close to the nominal value. Good reviews of methods are given by Agresti and Coull [1], Newcombe [10] and Brown *et al.* [2].

To cover a number of methods of forming confidence intervals relatively succinctly, we group them as Wald-form methods, exact-form methods, score-form methods and Bayesian methods. In the absence of ties, Wald-form methods include the Wald method (with and without a continuity correction) and Agresti–Coull intervals. Exact-form methods include the Clopper–Pearson intervals and mid-p intervals, while score-form methods include the Wilson score method (with and without a continuity correction). Bayes methods require a prior distribution that is usually assumed to be a beta distribution, which is the conjugate distribution. We describe these methods, extend them to handle ties, and examine their coverage.

In Section 2, we give a general procedure for extending a method of forming confidence intervals so that it can handle ties that satisfy the random-ranking assumption. We also give conditions under which the procedure is no more liberal than the method it extends. In Section 3, we briefly review methods of forming interval estimates for a binomial proportion and extend them to handle ties by applying our procedures. In Section 4, we examine the coverage probabilities of intervals and compare our procedure when ties are present with the equivalent method when ties are absent. We find that intervals are typically more conservative when ties are present. In Section 5, we give a practical example in which we compare our procedure for handling ties with naive alternative methods. The proposed procedure has been implemented in freely available software and details are provided in Section 6. Some concluding comments are given in Section 7. The work reported here underpins methods of forming confidence intervals that are proposed in Crawford *et al.* [7] and Crawford *et al.* [6].

2. General procedure

Throughout we assume that an equal-tailed $100(1-2\alpha)\%$ confidence interval is required. Thus if l and u are the lower and upper confidence limits, then $(0, u)$ and $(l, 1)$ are both $100(1-\alpha)\%$ confidence intervals for p. We refer to these as the upper-tailed and lower-tailed $100(1-\alpha)\%$ confidence intervals, respectively. The reason for introducing these one-tailed intervals is that they are defined for $0 \le \alpha \le 1$, while equal-tailed intervals are only defined for $0 \le \alpha \le 0.5$. To handle ties, we will consider confidence intervals that have many different levels of confidence, even though we only construct $100(1-2\alpha)\%$ confidence intervals.

In most statistical applications where confidence intervals are used, first a nominal confidence level is specified (say 95%) and then the confidence limits are determined. Here, in contrast, for the upper limit we wish to specify a value u and ask:

> What confidence level should be associated with the interval $(0, u)$? That is, for what value of γ is $(0, u)$ an upper-tailed $100(1-\gamma)\%$ confidence interval?

Let $\gamma(x, u)$ denote the answer to this question. For the lower confidence limit, l, we equivalently define $\eta(x, l)$ to be the value of η for which $(l, 1)$ is a $100(1-\eta)\%$ lower-tailed confidence interval. The quantities $\gamma(x, u)$ and $\eta(x, l)$ form the building blocks of our procedure for handling ties.

The procedure requires ties to satisfy the random-ranking assumption. Also, the methods of forming confidence intervals that we extend to handle ties must satisfy the following regularity conditions.

(a) Suppose there are no ties and (l_1, u_1) and (l_2, u_2) are equal-tailed confidence intervals with confidence levels of $100(1 - 2\alpha_1)$ and $100(1 - 2\alpha_2)$, respectively. We require $l_1 \leq l_2$ and $u_2 \leq u_1$ if $\alpha_1 < \alpha_2$.
(b) Some standard methods of forming confidence intervals can give upper limits that are above 1 if they are not reduced to equal 1, and lower limits that are below 0 if they are not increased to 0. For these methods, $\gamma(x, u)$ and $\eta(x, l)$ are not uniquely defined when $u = 1$ and $l = 0$, respectively. However, we assume that $\gamma(x, u)$ is uniquely defined when $u \neq 1$ and that $\eta(x, l)$ is uniquely defined when $l \neq 0$. For definiteness, we put $\gamma(x, u) = 0$ when $u = 1$ and $\eta(x, l) = 0$ when $l = 0$.
(c) We further assume that, for any x, $\gamma(x, u)$ is a continuous function of u for $u \in [0, 1)$ and $\eta(x, l)$ is a continuous function of l for $l \in (0, 1]$.

The procedure is as follows. When s successes and t ties have been observed, but x is unknown, we set the upper limit of the confidence interval equal to u^*, where u^* is the smallest value that satisfies $\sum_{x=s}^{s+t} \gamma(x, u^*)/(t + 1) \leq \alpha$. As $\gamma(x, u)$ is continuous for $u \in [0, 1)$, this uniquely defines $\gamma(x, u^*)$ and

$$\sum_{x=s}^{s+t} \frac{\gamma(x, u^*)}{t + 1} = \alpha, \tag{2}$$

if $u^* \neq 1$. The intuitive idea is that x is equally likely to take any of the values $s, \ldots, s + t$ and α is the average of $\gamma(x, u^*)$ as x ranges over these values. For the lower limit, we set l^* equal to the largest value that satisfies $\sum_{x=s}^{s+t} \eta(x, l^*)/(t + 1) \leq \alpha$. For $l^* \neq 0$,

$$\sum_{x=s}^{s+t} \frac{\eta(x, l^*)}{t + 1} = \alpha. \tag{3}$$

Then (l^*, u^*) is the $100(1 - 2\alpha)\%$ confidence interval. A separate search is conducted for each limit. We adopt a simple search strategy in which the range that contains the limit is repeatedly halved (twenty times), so the limit is estimated to an accuracy of 0.5^{20}. Each search is very fast on a computer. To tailor our procedure to a given method of forming confidence intervals only requires $\gamma(x, u)$ and $\eta(x, l)$ to be specified.

In the remainder of this section, we examine properties of the coverage of our procedure. In the absence of ties, the definition of a confidence interval means that, in principle,

$$\gamma(x, u) = P(X \leq x | p = u). \tag{4}$$

Several methods of forming confidence intervals will only satisfy Equation (4) approximately, either because they use asymptotic approximations or because they aim for better coverage overall by being a little liberal for some values of p. (The discreteness of X does not prevent Equation (4) from holding as, given x, $\gamma(x, u)$ is a continuous function of u.) We define $\psi(x, u)$ by

$$\psi(x, u) = \gamma(x, u) - P(X \leq x | p = u). \tag{5}$$

If we were seeking a $100(1 - \alpha)\%$ upper-tailed confidence interval and $\alpha = \gamma(x, u)$, then $1 - P(X \leq x | p = u)$ would be the actual coverage of the method of forming confidence intervals when $p = u$. Also, $\psi(x, p)$ would be the difference between the nominal and actual coverages.

When ties are present, let U^* be the random variable that denotes the upper limit of the $100(1-\alpha)\%$ upper-tailed confidence interval given by our procedure. U^* depends on S and T. For any given p, by definition the coverage is $P(U^* \geq p)$. The following proposition relates this coverage to the $\psi(x, u)$. A proof of the proposition is given in the appendix.

Proposition 1 *Given p and $T = t$, let $s^*(t)$ denote the largest value of S such that*

$$\sum_{x=s^*(t)}^{s^*(t)+t} \frac{\gamma(x, p)}{t+1} \leq \alpha. \tag{6}$$

Also, let m denote the largest value taken by $s^*(t) + t$ as t varies and suppose $P(S = s|t) \geq P(S = s-1|t)$ for all $s \leq m$ and any t. If ties satisfy the random-ranking assumption, then

$$P(U^* \geq p) \geq (1-\alpha) + \sum_{t=0}^{n} \sum_{x=s^*(t)}^{s^*(t)+t} P(T=t) \cdot \frac{\psi(x, p) + h(x, t)}{t+1}, \tag{7}$$

where Equation (2) *gives the value of u^* that is taken by U^* when $S = s$ and $T = t$, and*

$$h(x, t) = P(X \leq x) - P(X \leq x|t). \tag{8}$$

It is difficult to construct a situation where $P(S = x|T) \geq P(S = x-1|T)$ for some $x \leq m$, as m is almost inevitably in the lower tail of the distribution of S. (Exceptions arise when n is very large and either p or $1-p$ is very small.) Hence the conditions of the proposition generally hold if ties satisfy the random-ranking assumption. An equivalent result to Proposition 1 holds for the lower limit.

It is helpful to explore further the relationship given by Equation (7). For $i \leq t$, let $A_i = \sum_{j=1}^{i} P[S = s^*(t) - i + j|t]$ and $B_i = \sum_{j=1}^{t+1} P[S = s^*(t) - i + j|t]\{i/(t+1)\}$. If $s^*(t) + t$ is in the lower tail of the distribution of S, then B_i will usually be greater than A_i unless $t = 0$. Typically, the difference is noticeable for non-zero t and increases with t. The relevance of A_i and B_i is that they determine the tightness of the inequality in Equation (7): by modifying the proof of Proposition 1 it can be shown that $P(U^* \geq p)$ exceeds the right-hand side of Equation (7) by at least

$$\sum_{t=0}^{n} \left\{ P(T=t) \sum_{i=1}^{t} (B_i - A_i) \right\}. \tag{9}$$

In contrast, Equation (8) suggests that $h(x, t)$ will generally be small. Coupled with the fact that $\sum_{t=0}^{n} \sum_{x=s^*(t)}^{s^*(t)+t} P(T=t)h(x, t) = 0$, it follows that

$$\sum_{t=0}^{n} \sum_{x=s^*(t)}^{s^*(t)+t} \frac{P(T=t)h(x, t)}{t+1} \tag{10}$$

is likely to be very small. Usually then, the quantity in Equation (9) will be positive and non-trivial in size while the quantity in Equation (10) is likely to be small (and it may well be negative). Whenever the quantity in Equation (9) is greater than the quantity in Equation (10),

$$P(U^* \geq p) \geq (1-\alpha) + \sum_{t=0}^{n} \sum_{x=s^*(t)}^{s^*(t)+t} P(T=t) \cdot \frac{\psi(x, p)}{t+1}. \tag{11}$$

We believe that Equation (11) will typically hold, although it does depend on the distribution of the number of ties. To illustrate that it need not hold, suppose that the number of ties is almost certain to exceed $1/\alpha$ and that p is very close to 1. Then the ties give a probability of more than $1-\alpha$ that X is less than n. Hence $P(U^* \geq p)$ is less than $1-\alpha$, since $p \approx 1$ and the upper confidence interval is less than 1 whenever X is less than n. If the method chosen to form confidence intervals in the absence of ties is never liberal, then $P(U^* \geq p) < 1-\alpha$ implies that Equation (11) does not hold. The key feature of this situation is that there are a large number of ties and the value of p is extreme. In such situations, the assumption that ties should be broken at random (the random-ranking assumption) will usually be suspect.

When Equation (11) does hold, it clarifies the relationship between the coverage of a method of forming confidence intervals (in the absence of ties) and the coverage of the procedure constructed from that method in order to handle ties. It gives the following results, which are proved in the appendix.

Proposition 2 *Suppose Equation* (11) *holds and that a given method for forming confidence intervals* (*when there are no ties*) *is conservative for all confidence levels. Then the proposed procedure for forming confidence intervals from that method when ties are present is also conservative.*

Proposition 3 *Suppose Equation* (11) *holds and we have constants* λ_1 *and* λ_2, *with* $\lambda_2 \geq 0$, *such that* $\psi(x,p) \geq \lambda_1 - \lambda_2\gamma(x,p)$ *for all* x *and* p. *Then the proposed procedure for forming a* $100(1-\alpha)\%$ *upper-tailed confidence interval has coverage that is no less than* $(1-\alpha)+\lambda_1-\lambda_2\alpha$.

Propositions 2 and 3 suggest that the procedure for forming confidence intervals in the presence of ties will tend to have a greater average coverage than the method from which it is derived. The coverage will also tend to increase as the number of ties increases (from Equation (9)). The examples in Section 4 will illustrate this. Proposition 3 does not hold for negative values of λ_2 because of the discreteness of a binomial distribution. Given p, if for every value of t there existed an integer $s^*(t)$ for which $\sum_{x=s^*(t)}^{s^*(t)+t}\gamma(x,p)/(t+1)=\alpha$ (rather than the inequality in Equation (6)), then the proposition would hold for all λ_2. Discreteness should only have a marginal effect, so Proposition 3 will not be far from holding for all λ_2.

The procedure described here is useful, not only when sampling is from a binomial distribution, but in other situations where there are multiple ties that can be broken at random. Moreover, the proofs of Propositions 1–3 make little use of the precise form of the sampling model, so that they hold more generally (e.g. in forming confidence intervals for the parameter of a Poisson distribution).

3. Specific methods

In this section, we review specific methods for forming confidence intervals in the absence of ties and extend the methods to handle ties by giving explicit formulae for $\gamma(x,u)$ and $\eta(x,l)$.

3.1 *Wald-form methods*

Let z_α denote the $1-\alpha$ quantile of a standard normal distribution. Suppose there are no ties and x successes in n trials. Then the $100(1-2\alpha)\%$ confidence interval given by Wald-form methods is

$$\tilde{p} \pm \left[z_\alpha \left\{ \frac{\tilde{p}(1-\tilde{p})}{\tilde{n}} \right\}^{1/2} + c_1 \right], \tag{12}$$

where $\tilde{p}$ is a point estimate of p and $0 \leq c_1 < 1$. In the simple Wald method (the standard method taught in most elementary statistics textbooks), $\tilde{n}=n$, $\tilde{p}=x/n$ and $c_1=0$. A continuity correction is added by setting $c_1 = 1/(2\tilde{n})$. For 95% confidence intervals, Agresti and Coull [1] suggest adding two to both the number of successes and the number of failures (i.e. put $\tilde{n} = n + 4$ and $\tilde{p} = (x+2)/\tilde{n}$), as this improves the coverage probability of intervals. This could also prove useful for other levels of coverage, as the pathological cases where $x = 0$ or $x = n$ may have a smaller adverse impact on coverage if two is added to the number of successes and the number of failures. For other levels of confidence, Brown *et al.* [2] generalize the Agresti–Coull interval by setting $\tilde{n} = n + z_\alpha^2$, $\tilde{p} = (x + z_\alpha^2/2)/\tilde{n}$ and $c_1 = 0$. The formula in Equation (12) can yield confidence limits that are outside the feasible range of [0, 1], so they may require truncation. However, the Agresti–Coull method in particular gives 95% confidence intervals that have reasonably accurate coverage for moderately large n; Brown *et al.* [2] recommend its use for $n \geq 40$. At the same time though, results in Tuyl *et al.* [13] suggest that adding two successes and two failures is not helpful when $x = 0$.

To extend the Wald-form methods to handle ties, we define $\gamma(x, u^*)$ by

$$u^* = \tilde{p} + \left[z_{\gamma(x,u^*)} \left\{ \frac{\tilde{p}(1-\tilde{p})}{\tilde{n}} \right\}^{1/2} + c_1 \right] \tag{13}$$

when $\tilde{p}(1-\tilde{p}) \neq 0$. To handle pathologies, when $\tilde{p}(1-\tilde{p}) = 0$ put

$$\gamma(x, u^*) = \begin{cases} 0 & \text{if } u^* < \tilde{p} + c_1 \text{ or if } u^* = \tilde{p} + c_1 \text{ and } \tilde{p} = 1 \\ 1 & \text{if } u^* > \tilde{p} + c_1 \text{ or if } u^* = \tilde{p} + c_1 \text{ and } \tilde{p} = 0. \end{cases}$$

With these choices, if $u^* \neq \tilde{p} + c_1$, then $\gamma(x, u^*)$ is continuous as $\tilde{p}(1-\tilde{p}) \to 0$ while, if $u^* = \tilde{p} + c_1$, then $\gamma(x, u^*)$ takes its minimal value when u^* takes its maximal value, and vice versa. For the lower limit, define $\eta(x, l^*)$ by

$$l^* = \tilde{p} - \left[z_{\eta(x,l^*)} \left\{ \frac{\tilde{p}(1-\tilde{p})}{\tilde{n}} \right\}^{1/2} + c_1 \right] \tag{14}$$

when $\tilde{p}(1-\tilde{p}) \neq 0$ and, when $\tilde{p}(1-\tilde{p}) = 0$, put

$$\eta(x, l^*) = \begin{cases} 1 & \text{if } l^* < \tilde{p} - c_1 \text{ or if } l^* = \tilde{p} - c_1 \text{ and } \tilde{p} = 1 \\ 0 & \text{if } l^* > \tilde{p} - c_1 \text{ or if } l^* = \tilde{p} - c_1 \text{ and } \tilde{p} = 0. \end{cases}$$

Then, when there are s successes and t ties, the confidence limits are obtained via our general procedure, using Equations (2) and (3).

An advantage of Wald-form methods is their simplicity, though this benefit is largely lost with our procedure when ties are present, as searches are needed to find the confidence limits. Wald-form methods can have poor coverage, so the main reason for using them when there are ties, rather than using alternative methods, would be their greater familiarity to most people.

Extending the method of Brown *et al.* [2] to handle ties is messy. With their method, the adjustments that give $\tilde{n}$ and $\tilde{p}$ depend upon the confidence level, while the confidence level depends upon $\tilde{n}$ and $\tilde{p}$. Hence an additional layer of iteration is required. We will not consider Brown *et al.*'s method further here, as it is not widely used when there are no ties, and ties stop it from being a comparatively simple method to use.

3.2 *Score-form methods*

The $100(1-2\alpha)\%$ confidence interval for p given by score-form methods consists of all θ that satisfy

$$\frac{|\theta-(x/n)|-c_2}{\{\theta(1-\theta)/n\}^{1/2}} \leq z_\alpha, \tag{15}$$

where $0 \leq c_2 < 1$. Setting $c_2 = 0$ gives the Wilson [15] score-form method and putting $c_2 = 1/(2n)$ gives a continuity correction to the score-form method. Confidence limits can be found by inverting Equation (15) to give bounds on the value of θ [9]. The theoretical appeal of the Wilson interval is that it is equivalent to an equal tail score test of the hypothesis $H_0: p = \theta$. Assuming the normal approximation to the binomial is used, H_0 is rejected at the 2α significance level if and only if θ is within the $100(1-2\alpha)\%$ confidence interval. Equivalence is retained if both the hypothesis test and confidence interval method use a continuity correction. (Score tests are based on the log likelihood at the value of the parameter under the null hypothesis, whereas Wald tests are based on the log likelihood at the parameter's maximum-likelihood estimate.) Although score form methods of forming confidence intervals are not strictly conservative, their coverage probability tends to be close to the nominal value even for very small sample sizes [8]. Their use is widely recommended because of their good coverage [9,10].

To extend the methods to handle ties requires $\gamma(x, u^*)$ and $\eta(x, l^*)$. They are defined by

$$z_{\gamma(x,u^*)} = \frac{u^*-(x/n)-c_2}{\{u^*(1-u^*)/n\}^{1/2}} \quad \text{if } u^*(1-u^*) \neq 0.$$

and

$$z_{\eta(x,l^*)} = \frac{(x/n)-l^*-c_2}{\{l^*(1-l^*)/n\}^{1/2}} \quad \text{if } l^*(1-l^*) \neq 0.$$

If $u^*(1-u^*) = 0$, then

$$\gamma(x,u^*) = \begin{cases} 0 & \text{if } u^* < \dfrac{x}{n}+c_2 \text{ or if } u^* = \dfrac{x}{n}+c_2 \text{ and } \dfrac{x}{n}+c_2 = 1 \\ 1 & \text{if } u^* > \dfrac{x}{n}+c_2 \text{ or if } u^* = \dfrac{x}{n}+c_2 \text{ and } \dfrac{x}{n}+c_2 = 0. \end{cases}$$

If $l^*(1-l^*) = 0$, then

$$\eta(x,l^*) = \begin{cases} 1 & \text{if } l^* < \dfrac{x}{n}-c_2 \text{ or if } l^* = \dfrac{x}{n}-c_2 \text{ and } \dfrac{x}{n}+c_2 = 1 \\ 0 & \text{if } l^* > \dfrac{x}{n}-c_2 \text{ or if } l^* = \dfrac{x}{n}-c_2 \text{ and } \dfrac{x}{n}+c_2 = 0. \end{cases}$$

3.3 *Exact-form methods*

When there are x successes and no ties, the equal-tailed $100(1-2\alpha)\%$ confidence interval given by exact-form methods is (l, u) where u is the largest value in [0, 1] for which

$$c_3 P(X \leq x|p=u) + (1-c_3)P(X \leq x-1|p=u) \geq \alpha \tag{16}$$

and l is the smallest value in [0, 1] for which

$$c_3 P(X \geq x|p=l) + (1-c_3)P(X \geq x+1|p=l) \geq \alpha, \tag{17}$$

where $1 \geq c_3 > \alpha$. Setting $c_3 = 1$ gives the Clopper–Pearson method and putting $c_3 = 1/2$ gives the mid-p method. The Clopper–Peason method is guaranteed to be conservative and is commonly treated as the gold-standard method [1]. It inverts the hypothesis test $H_0: p = \theta$ where the test uses exact binomial probabilities; θ is within the Clopper–Pearson interval if and only if H_0 is not rejected at the 2α level of significance. Since the Clopper–Pearson interval inverts a test of the same hypothesis as the Wilson interval but uses exact probabilities, rather than a normal approximation, the Clopper–Pearson interval might be expected to be preferable to the Wilson interval, other than for reasons of computational convenience. However, unlike the Wilson method, the Clopper–Pearson method is widely regarded as being too conservative. For example, Agresti and Coull [1, p. 119], comment that its '…actual coverage probability can be much larger than the nominal confidence level unless n is quite large, and we believe it is inappropriate to treat this approach as optimal for statistical practice'. In comparison, mid-p intervals are better regarded as they give coverage probabilities that are generally close to the nominal value; Brown *et al.* [2, p. 115], comment that 'they are known to have good coverage and length performance'.

Tail-areas of a binomial distribution and a beta distribution are related. If $f_B(\xi; a, b)$ is the probability density function of a beta distribution with parameters a and b, then Equations (16) and (17) can be expressed as

$$c_3 \int_u^1 f_B(\xi; x+1, n-x)\,\mathrm{d}\xi + (1-c_3)\int_u^1 f_B(\xi; x, n-x+1)\,\mathrm{d}\xi \geq \alpha \tag{18}$$

and

$$c_3 \int_0^l f_B(\xi; x, n-x+1)\,\mathrm{d}\xi + (1-c_3)\int_0^l f_B(\xi; x+1, n-x)\,\mathrm{d}\xi \geq \alpha, \tag{19}$$

respectively. These expressions are useful for computation and also help compare exact-form confidence intervals with Bayesian credible intervals.

To extend exact-form methods to handle ties, we simply put

$$\gamma(x, u^*) = c_3 \int_{u^*}^1 f_B(\xi; x+1, n-x)\,\mathrm{d}\xi + (1-c_3)\int_{u^*}^1 f_B(\xi; x, n-x+1)\,\mathrm{d}\xi$$

and

$$\eta(x, l^*) = c_3 \int_0^{l^*} f_B(\xi; x, n-x+1)\,\mathrm{d}\xi + (1-c_3)\int_0^{l^*} f_B(\xi; x+1, n-x)\,\mathrm{d}\xi.$$

3.4 *Bayesian credible intervals*

Bayesian credible intervals may be regarded as confidence intervals when estimating a binomial proportion. They probably will not meet the strict definition of a confidence interval, in that their actual coverage for a given p may sometimes be less than their nominal coverage, but this is true of many well-recommended methods of forming confidence intervals for a binomial proportion.

For a Bayesian analysis, a prior distribution for p must be specified. Most commonly, it is assumed that the prior distribution is a beta distribution as this is the conjugate distribution for binomial sampling: if the prior distribution is beta(a, b) and x successes and no ties are observed in n trials, then the posterior distribution is another beta distribution, beta($x + a$, $n - x + b$). Then (l, u) is the equal-tailed $100(1 - 2\alpha)\%$ credible interval for p if u is the largest value in

[0, 1] for which

$$\int_{u}^{1} f_B(p; x + a, n - x + b)\,\mathrm{d}p \geq \alpha.$$

and l is the smallest value in [0, 1] for which

$$\int_{0}^{l} f_B(p; x + a, n - x + b)\,\mathrm{d}p \geq \alpha.$$

Values of a and b are generally chosen to be non-informative, with either $a = b = 1/2$, or $a = b = 0$, or $a = b = 1$. Setting a and b equal to $1/2$ gives Jeffreys' prior and with this choice the endpoints of the credible interval are usually very close to the endpoints of the mid-p confidence interval (obtained when $c_3 = 1/2$ in Equations (18) and (19)). The mean of a beta($x + a$, $n - x + b$) distribution is $(x + a)/(n + a + b)$, so setting $a = b = 0$ gives a posterior mean that equals the sample mean, conducive with a non-informative prior distribution. At the same time, putting $a = b = 1$ gives the uniform distribution whose flat shape might also be considered non-informative. In the literature on forming interval estimates for binomial proportions though, Jeffreys' prior is generally used.

Coverage is only a frequentist property and not a Bayesian concept. However, if a method is to be judged by the simple average of its coverage as p ranges over its possible values, then the Bayesian method that uses a *uniform* prior will seem perfect, with an average coverage exactly equal to the nominal confidence level. This is because a uniform prior is equivalent to letting p range over its possible values, with each value equally likely, precisely as in the method of determining average coverage. Credible intervals derived from Jeffreys' prior typically also have good coverage. This is perhaps unsurprising as a Jeffreys' prior and a uniform prior will usually lead to similar posterior distributions and, in addition, credible intervals for Jeffreys' prior are similar to mid-p intervals. Using Bayesian methods with Jeffrey's prior is widely recommended as a means of forming interval estimates [2,3].

When there is a multiple tie, our general method for forming confidence intervals can be applied. If the prior distribution is beta(a, b), we put

$$\gamma(x, u^*) = \int_{u^*}^{1} f_B(\xi; x + a, n - x + b)\,\mathrm{d}\xi$$

and

$$\eta(x, l^*) = \int_{0}^{l^*} f_B(\xi; x + a, n - x + b)\,\mathrm{d}\xi.$$

4. Coverage probabilities

We wish to examine the performance of our procedures for handling ties and to compare different methods of forming confidence intervals in the presence of ties. The primary measure of performance is the coverage of confidence intervals. Ideally, the coverage should exactly equal the nominal confidence level for every value of p, with equal coverage in each tail, as the intention is to form equal-tailed confidence intervals. When the coverage does not equal its nominal level then:

(a) It is better for the coverage to be too large rather than too small, as that is consistent with the definition of a confidence interval.
(b) The length of the confidence interval is of interest, as much narrower confidence intervals might be considered reasonable compensation for intervals that are too liberal for some values of p.

Given p and n, let L and U be random variables that denote the lower and upper confidence limits. Their values depend on X (if there are no ties) or S and T (if there are ties) and the coverage probability, $C_n(p)$, is given by

$$C_n(p) = \Pr(L \leq p \leq U).$$

Ideally, $C_n(p)$ should equal $1 - 2\alpha$ for $100(1 - 2\alpha)\%$ confidence intervals. In the literature, interest has focused on $C_n(p)$ but here we also consider $\Pr(p < L)$ and $\Pr(p > U)$, referring to these probabilities as the lower-tail coverage and the upper-tail coverage. These tail coverages should each equal α. We consider tail coverage because the effect of ties on the coverage of one tail is more transparent than their effect on the coverage of an interval.

As is well known, for almost any fixed value of p, coverage probabilities will not equal their nominal value for any of the methods of forming confidence intervals considered in Section 3, even when there are no ties. To illustrate, we consider Wilson's method (without a continuity correction). From Equation (15), the upper-tail coverage for a given n and p, $\mathrm{UC}_n(p)$ say, is given by

$$\mathrm{UC}_n(p) = \int_p^1 f_B(\xi; x+1, n-x)\,\mathrm{d}\xi,$$

where x is the largest integer for which $(x/n - p)\{p(1-p)/n\}^{-1/2} \leq z_\alpha$. The upper graph in Figure 1 plots $\mathrm{UC}_n(p)$ against p for the case where there are twenty trials (i.e. $n = 20$), no ties are possible and $\alpha = 0.025$. The jagged appearance of the plot arises because the number of successes has only 21 possible values (there are 20 spikes in the plot), so there are only 21 possible confidence intervals, each of which has a positive probability of occurring. Coverage changes as p moves from just inside a confidence limit to just outside it. For example, when there are six successes in 20 trials, the Wilson confidence interval for p is (0.1455, 0.5190). When $p = 0.5190$ (the upper limit), the probability of six successes is 0.0269. Hence the upper tail coverage is 0.0269 greater when $p = 0.5189$ than when $p = 0.5191$, as the former point is within the interval (0.1455, 0.5190) while the latter point is outside it.

The top graph also shows that the upper-tail coverage is zero for small values of p. This is because the upper confidence limit is 0.1611 when there are no successes so the upper limit always exceeds p for $p < 0.1611$. The dashed horizontal line in the graph marks the nominal coverage of $\alpha = 0.025$. The line shows that, with Wilson's method, the upper-tail coverage tends to be too small for $p < 0.5$ and too large for $p > 0.5$. The converse is true for lower tail coverage, of course. These tend to counteract each other when calculating the coverage of a confidence interval and the coverage of a Wilson confidence interval is generally good, sometimes much better than would be anticipated from the coverage in one tail.

Ties affect the coverage of intervals. In particular, when ties are present there are more possible outcomes as an outcome now includes the number of ties, as well as the numbers of successes and failures. Hence, there are more values that confidence limits can take. This means that there are more spikes in a plot of coverage against p, but the sizes of spikes tend to be smaller. To give an example involving ties, we must specify a joint distribution for S and T. For $p \geq 0.5$, we put

$$\Pr(S = s, T = t) = \lambda_s \binom{n}{t} \theta^t (1-\theta)^{n-t}. \binom{n-t}{s} p^s (1-p)^{n-t-s}, \tag{20}$$

for $s + t = 0, \ldots, n$, where s and t are integers, $0 \leq \theta \leq 1$, and $\lambda_0, \ldots, \lambda_s$ are positive constants. If $\lambda_s = 1$ for all s, then Equation (20) would imply that the marginal distribution of T is $\mathrm{bin}(n, \theta)$ and the conditional distribution $S|T = t$ would be $\mathrm{bin}(n - t, p)$. (The joint distribution of S, T and the number of failures would be multinomial.) Thus, Equation (20) with $\lambda_s \equiv 1$ seems a reasonable first-approximation to the joint distribution of S and T and is the motivation for defining their joint distribution by Equation (20). We choose θ to control the distribution of the number of ties. The λ_i must be chosen so that the distribution of X is binomial when ties are broken at random. If $X \sim \text{bin}(n, p)$, then from Equation (1),

$$\sum_{s=0}^{x} \sum_{t=x-s}^{n-s} \frac{1}{t+1} \Pr(S = s,\ T = t) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{21}$$

for $x = 0, \ldots, n$. This determines the λ_i. Setting $x = 0$ in Equation (20) gives λ_0; after $\lambda_0, \ldots, \lambda_i$ have been calculated ($i < n$), λ_{i+1} is determined by setting $x = i + 1$ in Equation (20). In the cases we consider, the λ_i are always positive provided $p > 0.5$. For $p < 0.5$ we interchange the roles of *success* and *failure*. That is, we substitute F for S, f for s, and $1 - p$ for p in Equations (20) and (21), where F is the number of failures. Calculation yields $\Pr(F = f,\ T = t)$ and we put $\Pr(S = s,\ T = t)\Pr(F = n - s - t,\ T = t)$. This gives a symmetric relationship between success and failure in their coverage probabilities, which is as it should be, as the labels *success* and *failure* are often assigned arbitrarily to the possible outcomes of a trial. We refer to this example (defined by Equations (20) and (21)) as the 'sampling-with-ties example'.

To examine the coverage of a method for this example, we determined the confidence (credible) interval that the method gives for each feasible combination of S, T. Then the upper-tail coverage for a given value of p is

$$\text{UC}_n(p) = \Pr(\{s, t\} \in \Omega),$$

where Ω is the set of $\{s, t\}$ combinations for which p is above the upper confidence limit. Lower-tail coverage has an equivalent definition and the coverage of a confidence interval is $1 -$ (lower-tail coverage + upper-tail coverage). The lower graph in Figure 1 gives the upper-tail coverage for Wilson's method when θ, the parameter in Equation (20) that controls the proportion of ties, is set equal to $0.2p(1 - p)$. (With this choice of θ, the number of ties does not swamp the number of successes or the number of failures.) Comparison of the two graphs in Figure 1 illustrates that spikes are smaller but more numerous when there are ties (lower graph) than when there are are no ties (upper graph).

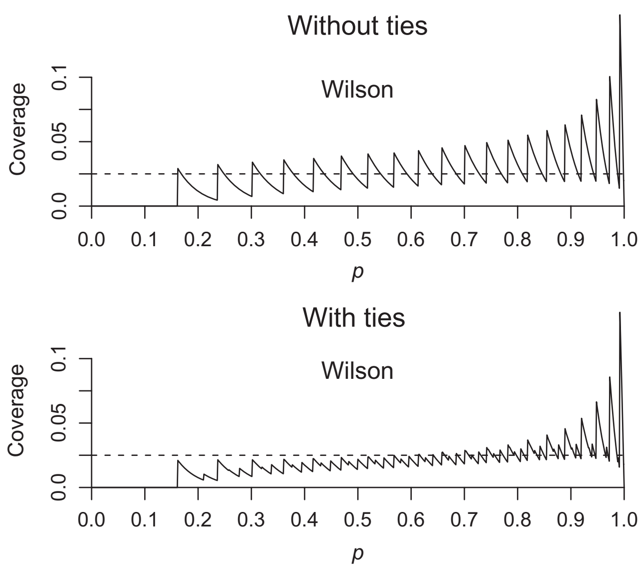

Figure 1. Upper-tail coverage for Wilson's method when there are no ties (upper graph) and for the sampling-with-ties example (lower graph).

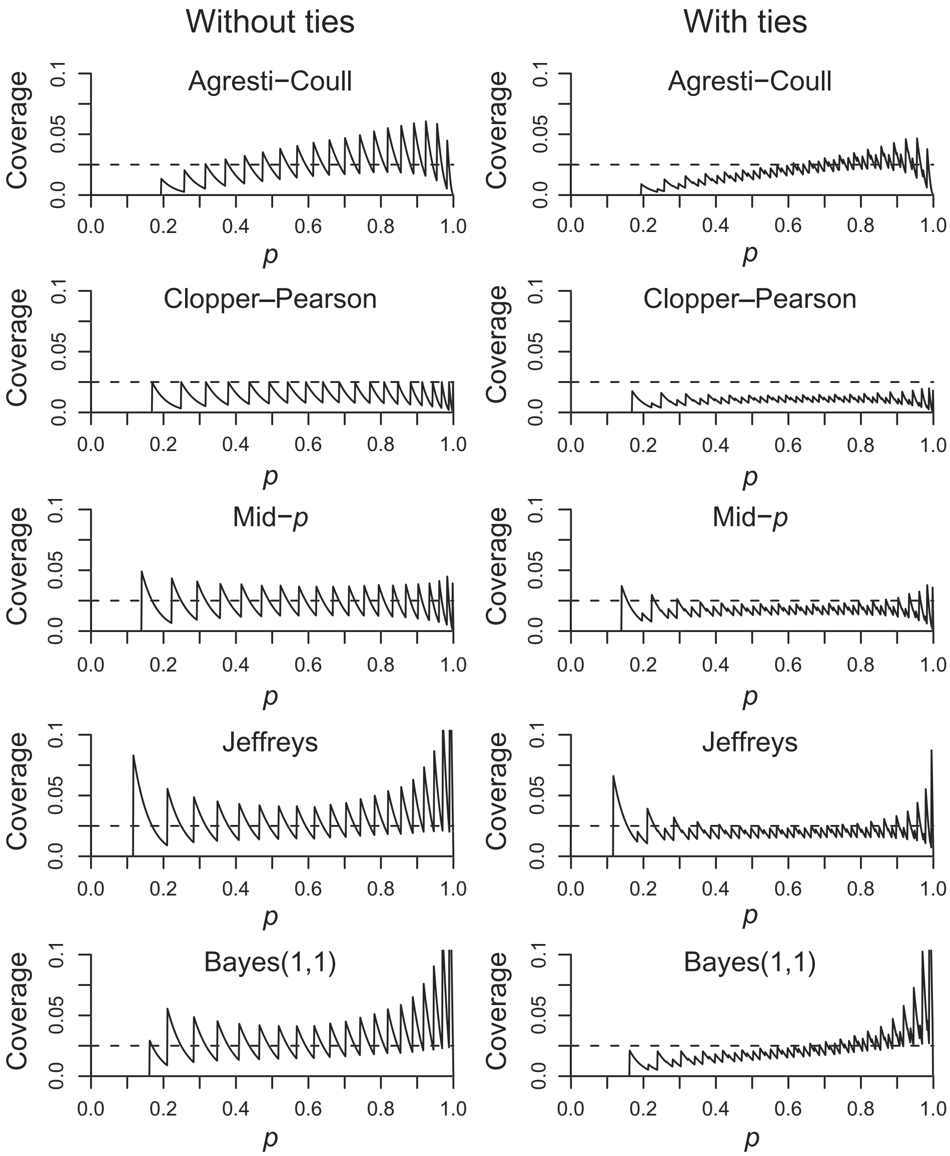

Figure 2. Upper-tail coverage for the Agresti–Coull, Clopper–Pearson, Mid-p, Jeffreys and Bayes(1,1) methods when there are no ties (left-hand graphs) and for the sampling-with-ties example (right-hand graphs).

Equivalent graphs for five other well-recommended methods are given in Figure 2. Each of them again shows that there are smaller, but more numerous spikes when there are ties (right-hand graphs) than when there are no ties (left-hand graphs). Figures 1 and 2 also both suggest that coverage tends to be more conservative in the presence of ties. However, there are differences in the coverage patterns of the methods. The Agresti–Coull method has a similar pattern to Wilson's method (Figure 1), with the coverage tending to be lower for $p < 0.5$ than for $p > 0.5$. This is also the case for the Bayes(1,1) method when ties are present but not when ties are absent. The other three methods have coverages that show almost no trend, with the mid-p and Jeffreys methods giving a similar level of coverage while the coverage for Clopper–Perason is noticeably

Table 1. Average interval coverage and average exceedance of coverage for 95% confidence intervals formed by various methods: results for 20 and 50 trials, both when ties are absent ($\theta = 0$) and when ties are present ($\theta > 0$).

Method	No ties ($\theta = 0$) Aver. cover.	No ties ($\theta = 0$) Aver. exceed.	$\theta = 0.1p(1-p)$ Aver. cover.	$\theta = 0.1p(1-p)$ Aver. exceed.	$\theta = 0.2p(1-p)$ Aver. cover.	$\theta = 0.2p(1-p)$ Aver. exceed.
$n = 20$						
Standard Wald	0.846	0.1042	0.856	0.0942	0.866	0.0843
Standard (cc)[a]	0.870	0.0844	0.878	0.0775	0.886	0.0728
Agresti–Coull	0.961	0.0015	0.965	0.0001	0.970	0.0000
Wilson	0.953	0.0054	0.958	0.0019	0.963	0.0010
Wilson (cc)[a]	0.976	0.0000	0.979	0.0000	0.982	0.0000
Clopper–Pearson	0.977	0.0000	0.980	0.0000	0.983	0.0000
Mid-p	0.961	0.0027	0.966	0.0003	0.970	0.0000
Bayes(0, 0)[b]	0.859	0.0931	0.868	0.0830	0.878	0.0759
Jeffreys(0.5, 0.5)[b]	0.951	0.0076	0.956	0.0031	0.962	0.0014
Bayes(1, 1)[b]	0.950	0.0083	0.955	0.0053	0.960	0.0044
$n = 50$						
Standard Wald	0.901	0.0497	0.910	0.0405	0.920	0.0345
Standard (cc)[a]	0.919	0.0358	0.927	0.0320	0.935	0.0299
Agresti–Coull	0.958	0.0013	0.963	0.0000	0.970	0.0000
Wilson	0.952	0.0034	0.958	0.0006	0.965	0.0004
Wilson (cc)[a]	0.969	0.0000	0.974	0.0000	0.979	0.0000
Clopper–Pearson	0.969	0.0000	0.974	0.0000	0.979	0.0000
Mid-p	0.955	0.0023	0.961	0.0002	0.968	0.0000
Bayes(0, 0)[b]	0.910	0.0414	0.919	0.0339	0.928	0.0308
Jeffreys(0.5, 0.5)[b]	0.950	0.0049	0.957	0.0011	0.964	0.0005
Bayes(1, 1)[b]	0.950	0.0048	0.956	0.0021	0.964	0.0018

Notes: [a]Standard Wald/Wilson method with a *continuity correction.*
[b]Bayesian credible interval where the numbers in brackets are the parameters of the beta prior.

more conservative, especially in the presence of ties. The Clopper–Pearson method, of course, is the only one of the methods that guarantees to give a coverage that is not liberal.

Table 1 gives the average coverage of ten methods as p varies from 0 to 1, including the methods shown in Figures 1 and 2. Methods were used to construct 95% confidence intervals. Three situations are considered: (i) no ties, and the sampling-with-ties-example with either (ii) $\theta = 0.01p(1-p)$ or (iii) $\theta = 0.02p(1-p)$. With (ii), when p is small just under 10% of the controls whose underlying score is less than the case are expected to get the same actual score as the case. With (iii) the proportion doubles to 20%.

Both for $n = 20$ and $n = 50$, the average coverage of intervals is greater for each method when there are ties than when there are no ties, and it is greatest for the larger proportion of ties. When there are no ties, the standard Wald method (with or without a continuity correction) gives poor average coverage that is well below the nominal level of 0.95, and the same is true of the Bayes credible interval when a beta(0, 0) distribution is taken as the prior. The average coverage of these methods improves in the presence of ties. Other methods in the table give a reasonable average coverage when there are no ties: the Bayesian credible interval from a uniform prior has an average coverage of exactly 0.95 while the average coverages of other methods are a little higher. Hence, as the probability of a tie increases, the average coverages of these methods worsen with our procedure for handling ties. However, the decrease in performance is not dramatic. Indeed, even for $\theta = 0.2p(1-p)$, the average coverages of the Wilson, mid-p, Jeffreys and Bayes(1, 1)

Table 2. Average interval coverage and average exceedance of coverage for 99% confidence intervals formed by various methods: results for 50 trials, both when ties are absent ($\theta = 0$) and when ties are present ($\theta > 0$).

	No ties ($\theta = 0$)		$\theta = 0.1p(1-p)$		$\theta = 0.2p(1-p)$	
Method	Aver. cover.	Aver. exceed.	Aver. cover.	Aver. exceed.	Aver. cover.	Aver. exceed.
Standard Wald	0.940	0.0500	0.946	0.0443	0.951	0.0396
Standard (cc)[a]	0.946	0.0429	0.952	0.0390	0.956	0.0362
Agresti–Coull	0.992	0.0002	0.993	0.0000	0.995	0.0000
Wilson	0.989	0.0017	0.991	0.0008	0.993	0.0006
Wilson (cc)[a]	0.994	0.0001	0.995	0.0001	0.996	0.0000
Clopper–Pearson	0.994	0.0000	0.996	0.0000	0.997	0.0000
Mid-p	0.992	0.0004	0.993	0.0000	0.995	0.0000
Bayes(0, 0)[b]	0.950	0.0399	0.954	0.0364	0.958	0.0345
Jeffreys(0.5, 0.5)[b]	0.990	0.0011	0.992	0.0002	0.994	0.0001
Bayes(1, 1)[b]	0.990	0.0012	0.992	0.0006	0.994	0.0005

Notes: [a]Standard Wald/Wilson method with a *continuity correction.*
[b]Bayesian credible interval where the numbers in brackets are the parameters of the beta prior.

intervals are better than the average coverage of the 'gold-standard' Clopper–Pearson method when there are no ties. In the main, Agresti–Coull intervals also have slightly better coverage in the presence of ties than Clopper–Pearson intervals when ties are absent.

By definition, confidence intervals are meant to be conservative rather than liberal, with a coverage that at least equals its nominal value for every value of p. Figures 1 and 2 show that many methods give a one-tail coverage that is sometimes liberal and they suggest that coverage is less liberal with our procedure when ties are present. To quantify the extent to which a method is liberal, we define the *exceedance* of an interval by

$$\text{exceedance} = \begin{cases} \text{nominal level} - \text{coverage} & \text{if nominal level} > \text{coverage} \\ 0 & \text{otherwise.} \end{cases}$$

The average exceedance as p varies from 0 to 1 was determined for each method and these values are also given in Table 1. The standard Wald methods and Bayes(0, 0) show average exceedances that are large regardless of whether or not ties are present. This is to be expected as these methods are markedly liberal when judged by average coverage. However, their average exceedances are noticeably better (i.e. smaller) when ties are present. Other methods also show clearly better average exceedances when ties are present, except when there was no room for improvement because, even with no ties, average exceedances were virtually zero.

The work reported in Table 1 was repeated, but with the nominal confidence level set at 0.99 rather than 0.95. Results for $n = 50$ are reported in Table 2. Patterns are similar to those found for a nominal coverage of 0.95. The coverage of the confidence intervals increases as the proportion of ties increases. The standard Wald methods and Bayes with a beta(0, 0) prior again give very liberal coverage while the other methods are conservative. As with 95% nominal coverage, the degrees of conservatism with the Agresti–Coull method, the Wilson methods, and the mid-p method were not excessive when there was a moderate proportion of ties, being similar to the level of the Clopper–Pearson method when there are no ties. The same is true of the Bayes methods with beta(0.5, 0.5) or beta(1, 1) as the prior. Taking both average coverage and average exceedance into consideration, these Bayes methods outperformed alternatives for forming 99% confidence intervals.

Table 3. Average length of 95% confidence intervals formed by various methods: results for 20 and 50 trials, both when ties are absent ($\theta = 0$) and when ties are present ($\theta > 0$).

	$n = 20$			$n = 50$		
Method	$\theta = 0$	$\theta = \theta_{0.1}$[a]	$\theta = \theta_{0.2}$[a]	$\theta = 0$	$\theta = \theta_{0.1}$[a]	$\theta = \theta_{0.2}$[a]
Standard Wald	0.316	0.321	0.327	0.211	0.215	0.220
Standard (cc)[b]	0.354	0.359	0.365	0.229	0.233	0.237
Agresti–Coull	0.337	0.341	0.345	0.218	0.221	0.225
Wilson	0.325	0.330	0.334	0.213	0.216	0.221
Wilson (cc)[b]	0.366	0.370	0.375	0.231	0.235	0.239
Clopper–Pearson	0.366	0.370	0.375	0.231	0.234	0.238
Mid-p	0.335	0.339	0.343	0.215	0.219	0.223
Bayes(0, 0)[c]	0.310	0.322	0.334	0.209	0.216	0.223
Jeffreys(0.5, 0.5)[c]	0.323	0.327	0.332	0.212	0.215	0.219
Bayes(1, 1)[c]	0.327	0.331	0.335	0.213	0.216	0.221

Notes: [a]$\theta_{0.1} = 0.1p(1 - p)$; $\theta_{0.2} = 0.2p(1 - p)$.
[b]Standard Wald/Wilson method with a *continuity correction.*
[c]Bayesian credible interval where the numbers in brackets are the parameters of the beta prior.

After coverage, the lengths of confidence intervals is the next most important feature of methods to construct them. The average length of confidence intervals was determined for each level of ties and each method. Results for 95% confidence/credible intervals are given in Table 3. The average width of intervals changed very little when ties were introduced or the proportion of ties increased. This is the case both for 20 and 50 trials. It suggests that our method does not usually increase the average size of intervals dramatically in order to make due allowance for the added uncertainty caused by a moderate number of ties.

5. A practical example

We have not constructed examples in which a large proportion of ties are present. In such cases, the assumptions that are made about the distribution of ties will affect results. Instead, we give an example based on a real database. For the Depression scale of the DASS-21 scale, the database of Crawford *et al.* [6] has 1421 controls with a score lower than 2, another 337 controls with a score of 2, and 1171 controls with a score greater than 2. Suppose a case scores 2 and a 95% confidence interval for the proportion of people with a lower underlying level of depression than the case is required.

Many controls have a score that ties with the case's. A naive approach is to treat half of them as having a lower level of depression than the case and half as having a higher level of depression. This incorrectly removes the uncertainty as to the precise number of controls that have a lower underlying depression level. It leads to a 95% confidence interval of (0.525, 0.561), using the mid-p method. In contrast, if our procedure is used in conjunction with the mid-p method, then a 95% confidence interval of (0.483, 0.602) is obtained. [This interval was calculated using the computer software described in the next section.] Hence, ignoring the uncertainty that arises from ties leads to an interval that is less than one-third of the width of an interval that takes this uncertainty into account. Moreover, our procedure (used in conjunction with the mid-p method) estimates that 32% is the appropriate confidence level to associate with the interval of (0.525, 0.561), rather than its nominal level of 95%. Almost identical differences are seen if the mid-p method is replaced by other reasonable methods of forming confidence intervals for a binomial proportion, such as Wilson's method or a Bayes method with a Jeffreys or beta(1,1) prior.

Another approach for handling ties is to simply discard them. This reduces the sample to 1421 controls with a score lower than the case and 1171 controls with a higher score. The mid-p method gives (0.529, 0.567) as the 95% confidence interval from this reduced sample. As with the other method that removes the uncertainty arising from ties, this interval is less than one-third of the width of the interval given by our procedure. Our procedure estimates its confidence level to be only 34%, rather than 95%. Obviously the uncertainty caused by ties should not be ignored, so using a procedure such as ours seems essential in this real example if a realistic confidence interval is to be formed.

6. Computer software

Calculating the confidence intervals that are given by our procedure would be a laborious task without a computer, even for only a modest number of ties. Hence a computer program has been written to accompany this paper. The program runs on PCs and can be freely accessed from the second author's web page at www.abdn.ac.uk/~psy086/dept/Binomial_Ties_CIs.htm. It can be downloaded (either as an uncompressed executable or as a zip file) or its executable form may be run without saving any files. The program implements all the methods described here for handling ties. It prompts the user to specify the number of successes, failures and ties that were observed and asks the user to state the confidence level that is required. The preferred method for forming confidence intervals must also be selected. Also, there are some 'user-tailored' methods that add flexibility by allowing the user to specify parameter values. The program outputs a point estimate of the binomial proportion and an equal-tailed confidence interval for that proportion.

To give more detail, the methods available are the standard Wald, Agresti–Coull and tailored Wald-form methods, Wilson's method and the tailored score-form method, the Clopper–Pearson, mid-p and tailored exact-form methods, and Bayesian credible interval using Jeffrey's prior, a beta(0,0) or beta(1,1) prior, or a tailored beta prior. With some methods, an optional continuity correction is offered and with the 'tailored' methods the user specifies parameter values. In particular, with the tailored Wald-form method, the user is prompted to specify the nominal sample size ($\tilde{n}$), the nominal number of successes ($\tilde{s}$) and the continuity correction c_1. The roles of ($\tilde{n}$) and c_1 are defined in Equation (12) and $\tilde{p} = \tilde{s}/\tilde{n}$ defines $\tilde{s}$. Usually the user should set c_1 equal to 0 (for no continuity correction) or $1/(2\tilde{n})$ (for the standard correction factor), but the choice is not restricted to these values. The tailored score-form method allows the user to specify the value of c_2 in Equation (15) and the tailored exact-form method allows the user to specify the value of c_3 in Equations (18) and (19). For Bayesian credible intervals, the form of the prior distribution is taken as a beta distribution and the user specifies its two parameters.

7. Concluding comments

The work in this paper shows that the proposed procedure is a viable means of extending methods of forming confidence intervals for binomial proportions so that the methods can handle ties. A variety of methods were considered in Section 3 and this showed that the procedure is widely applicable and generally straightforward to use. Formula for applying the procedure to the common methods of forming equal-tailed confidence intervals were given. Results presented in Table 3 suggest that the average width of confidence intervals will not increase greatly with our procedure when the proportion of ties is modest. However, the width must increase substantially if due allowance is to be made for a large proportion of ties.

As the binomial proportion increases from 0 to 1, the coverage of a method of forming confidence intervals repeatedly rises and falls. The changes are quite substantial when there are no ties. When ties are present, coverage is much more stable if our procedure is used – ties multiply the number of rises and falls but the drops from a peak to a trough are much smaller. This is illustrated very clearly in the figures in Section 4. These figures also suggest that a method which is sometimes quite liberal in the absence of ties will lead to a procedure that is appreciably less liberal when ties are present: in the figures, coverage crosses the nominal confidence level much less frequently when there are ties and, when the coverage does exceed its nominal level, it exceeds it by less.

In any numeric example, assumptions must be made about the distribution of the number of ties and only one form of distribution was considered in our examples. However, implications from the examples seem likely to generalize. In particular, the theoretical results given in Section 2 suggest that our procedure will generally be less liberal/more conservative than the method that it extends. This has relevance to the choice of method for forming a confidence interval for a proportion. In the absence of ties, the Clopper–Pearson method is often referred to as the gold-standard method because its coverage is always conservative and never liberal. Other less conservative methods have been proposed that aim to give a better average coverage without being too liberal too often. In the presence of ties, our procedure tends to make methods more conservative and reduce any tendency to be liberal. This increases the disadvantage of the Clopper–Pearson method while making several other methods (such as the mid-p, Jeffreys and Wilson methods) more attractive, with Jeffreys and Bayes(1,1) perhaps the methods of choice.

A computer program that implements the procedure proposed in this paper has been described in Section 5. The program is free and makes the procedure easy to use. A more targeted implementation of the procedure has also been produced that is useful to neuropsychologists and other medical workers [6]. That implementation has a database that contains scores for large samples of controls on five common mood scales, such as the *Depression and Anxiety Stress Scales* (*DASS*). The score of a case can be entered for one of the mood scales and, using the database, the computer program will estimate the proportion of controls with a smaller underlying value on that scale – typically there will be some controls whose scores tie with that of the case and it is assumed that these scores should be broken at random. A confidence interval for the proportion is determined using the procedure proposed in this paper. This application area is just one of many in medical, psychological and educational testing, where normative data contain a large number of multiple ties, so we believe that the methods developed in this paper have wide applicability.

Acknowledgements

We are grateful to Karen Vines and three referees whose constructive comments improved this paper significantly.

References

[1] A. Agresti and B.A. Coull, *Approximate is better than 'exact' for interval estimation of binomial proportions*, Amer. Statsit. 52 (1998), pp. 119–126.
[2] L.D. Brown, T.T. Cai, and A. DasGupta, *Interval estimation for a binomial proportion (with discussion)*, Statist. Sci. 16 (2001), pp. 101–133.
[3] B.P. Carlin and T.A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall, London, 1996.
[4] C.W. Coakley and M.A. Heise, *Versions of the sign test in the presence of ties*, Biometrics 52 (1996), pp. 1242–1251.
[5] W.J. Conover, *On methods of handling ties in the Wilcoxon signed-rank test*, J. Amer. Statist. Assoc. 68 (1973), pp. 985–988.
[6] J.R. Crawford, P.H. Garthwaite, C.J. Lawrie, J.D. Henry, M.A. MacDonaald, J. Sutherland, and P.A. Sinha, *A convenient method of obtaining percentile norms and accompanying interval estimates for self-report mood scales (DASS, DASS-21, HADS, PANAS, and sAD)*, Br. J. Clin. Psychol. 48 (2009), pp. 163–180.
[7] J.R. Crawford, P.H. Garthwaite, and D.L. Slick, *On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of scores*, Clin. Neuropsychol. 23 (2009), pp. 1173–1195.
[8] M.G. Hudgens and G.A. Satten, *Midrank unification of rank tests for the exact, tied, and censored data*, Nonparametr. Stat. 14 (2002), pp. 569–581.

[9] N.T. Longford, *Analysis of all-zero binomial outcomes with borderline and equilibrium priors*, J. Appl. Stat. 36 (2009), pp. 1259–1265.

[10] R.G. Newcombe, *Two-sided confidence intervals for the single proportion: comparison of seven methods*, Stat. Med. 17 (1998), pp. 857–872.

[11] J. Putter, *The treatment of ties in some nonparametric tests*, Ann. Math. Statist. 26 (1955), pp. 368–386.

[12] R.H. Randles, *On neutral responses (zeros) in the sign test and ties in the Wilcoxon-Mann-Whitney test*, Amer. Statist. 55 (2001), pp. 96–101.

[13] F. Tuyl, R. Gerlach, and K. Mengersen, *The rule of three, its variants and extensions*, Int. Stat. Rev. 77 (2009), pp. 266–275.

[14] P. Williamson, *A Bayesian alternative to the sign test in the presence of ties*, J. Stat. Comput. Simul. 68 (2001), pp. 135–152.

[15] E.B. Wilson, *Probable inference, the law of succession, and statistical inference*, J. Amer. Statist. Assoc. 22 (1927), pp. 209–212.

[16] R.L. Winkler, J.E. Smith, and D.G. Fryback, *The role of informative priors in zero-numerator problems: Being conservative vs. being candid*, Amer. Statist. 56 (2002), pp. 1–4.

[17] M.A. Woodbury, K.G. Manton, and L.A. Woodbury, *An extension of the sign test for replicated measurements*, Biometrics 33 (1977), pp. 453–461.

Appendix

Proof of Proposition 1 Given t, by assumption $\sum_{x=s}^{s+t} \gamma(x, p)/(t+1) > \alpha$ for $s = s^*(t) + 1, \ldots, n-t$. Hence, if $u < p$, then $\sum_{x=s}^{s+t} \gamma(x, u)/(t+1) > \alpha$ for $s = s^*(t) + 1, \ldots, n-t$. Equation (2) gives $\sum_{x=s}^{s+t} \gamma(x, u^*)/(t+1) = \alpha$ so, if $u^* < p$, then s must be smaller than $s^*(t+1)$. Consequently,

$$P(U^* < p|t) \leq P(S \leq s^*(t)|t). \tag{A1}$$

Now, $X \leq s^*(t)$ implies $S \leq s^*(t)$, so $P[\{S \leq s^*(t)\} \bigcap \{X \leq s^*(t)\}|t] = P[X \leq s^*(t)|t]$. Also, $P[\{S \leq s^*(t)\} \bigcap \{X > s^*(t)\}|t] = \sum_{j=1}^{t} P[S = s^*(t) - t + j|t].P[X > s^*(t)|S = s^*(t) - t + j, t] = \sum_{j=1}^{t} P[S = s^*(t) - t + j|t].(j/(t+1)) = \sum_{i=1}^{t} \sum_{j=1}^{i} P[S = s^*(t) - i + j|t]/(t+1)$. Hence,

$$P[S \leq s^*(t)|t] = P[X \leq s^*(t)|t] + \sum_{i=1}^{t} \sum_{j=1}^{i} P[S = s^*(t) - i + j|t]/(t+1). \tag{A2}$$

By assumption $P(S = s|t) \geq P(S = s-1|t)$ for $s \leq s^*(t) + t$, so $\sum_{j=1}^{i} P[S = s^*(t) - i + j|t] \leq \sum_{j=1}^{t+1} P[S = s^*(t) - i + j|t](i/(t+1)) = iP[X = s^*(t) - i + t + 1|t]$. Substituting in Equation (A2),

$$\begin{aligned} P[S \leq s^*(t)|t] &\leq P[X \leq s^*(t)|t] + \sum_{i=1}^{t} iP[X = s^*(t) - i + t + 1|t]\left(\frac{1}{t+1}\right) \\ &= \sum_{x=s^*(t)}^{s^*(t)+t} \frac{P(X \leq x|t)}{t+1}. \end{aligned}$$

Hence, from Equation (A1), $P(U^* < p|t) \leq \sum_{x=s^*(t)}^{s^*(t)+t} P(X \leq x|t)/(t+1)$, so from Equations (5) and (8)

$$\begin{aligned} P(U^* \leq p|t) &\leq \sum_{x=s^*(t)}^{s^*(t)+t} \frac{\gamma(x, p) - \psi(x, p) - h(x, t)}{t+1} \\ &\leq \alpha - \sum_{x=s^*(t)}^{s^*(t)+t} \frac{\psi(x, p) + h(x, t)}{t+1}. \end{aligned}$$

As $P(U^* \geq p) = 1 - P(U^* < p) = 1 - \sum_{t=0}^{n} P(T = t).P(U^* \leq p|t)$, Proposition 1 follows. ■

Proof of Propositions 2 *and* 3 Suppose that $\psi(x, p) \geq \lambda_1 - \lambda_2\gamma(x, p)$ for all x and p, with $\lambda_2 \geq 0$. Then $\sum_{t=0}^{n} \sum_{x=s^*(t)}^{s^*(t)+t} P(T = t).\psi(x, p)/(t + 1) \geq \lambda_1 \sum_{t=0}^{n} P(T = t) - \lambda_2 \sum_{t=0}^{n} \{P(T = t) \sum_{x=s^*(t)}^{s^*(t)+t} \gamma(x, p)/(t + 1)\} \geq \lambda_1 - \lambda_2\alpha$, from Equation (6). Hence, if Equation (11) holds,

$$P(U^* \geq p) \geq (1 - \alpha) + \lambda_1 - \lambda_2\alpha. \tag{A3}$$

This proves Proposition 3. When the assumptions underlying Proposition 2 apply, then $\psi(x, p) \geq 0$ for all x and p, so we may put $\lambda_1 = \lambda_2 = 0$. Then Equation (A3) gives $P(U^* \geq p) \geq 1 - \alpha$, which is the result given in Proposition 2. ■