On matching confidence intervals and tests for some discrete distributions: methodological and computational aspects

Jan Klaschka¹ · Jenő Reiczgel²

Received: 5 December 2019 / Accepted: 7 April 2020 / Published online: 18 April 2020
© The Author(s) 2020

Abstract

Exact two-tailed tests and two-sided confidence intervals (CIs) for a binomial proportion or Poisson parameter by Sterne (Biometrika 41:117–129, 1954) or Blaker (Can J Stat 28(4):783–798, 2000) are successful in reducing conservatism of the Clopper–Pearson method. However, the methods suffer from an inconsistency between the tests and the corresponding CIs: In some cases, a parameter value is rejected by the test, though it lies in the CI. The problem results from non-unimodality of the test p value functions. We propose a slight modification of the tests that avoids the inconsistency, while preserving nestedness and exactness. Fast and accurate algorithms for both the test modification and calculation of confidence bounds are presented together with their theoretical background.

Keywords Exact test · Sterne’s and Blaker’s exact confidence interval · Binomial model · Poisson model · Computational algorithm

1 Introduction

We follow up, in a sense, Fay (2010a, b) who pointed to some problems in categorical data analysis that may occur when a statistical test and a confidence interval (CI) is applied to the same data. It may happen that the test rejects a parameter value that lies in the CI or the other way round. He recommended, in order to avoid such inconsistencies,

J. Klaschka: First author’s work was supported by institutional fund RVO:67985807.

✉ Jenő Reiczgel
reiczgel.jeno@univet.hu
Jan Klaschka
klaschka@cs.cas.cz

¹ Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic

² University of Veterinary Medicine Budapest, Budapest, Hungary

using tests and CIs in matching pairs and offered (unlike some prestigious statistical programs, at least their versions of that time) software tools (R packages) making it possible to follow the recommendation.

However, even following this recommendation leaves some discrepancies between CIs and test results unresolved. Fay (2010b) calls them *unavoidable inconsistencies*. They stem from the fact that inverting some tests yields non-connected confidence sets whose gaps have to be “filled” in order to obtain confidence *intervals*. When a parameter value lies in such a gap, it is rejected by the test, yet it belongs to the CI.

We show that even such “unavoidable” inconsistencies can be avoided—it only is necessary to slightly modify the test. We propose an appropriate modification, as well as algorithms and software tools that make it applicable in some important cases, such as Sterne’s and Blaker’s methods for the binomial and Poisson distributions.

2 Desirable properties of confidence intervals and tests

2.1 General considerations

Throughout the paper, we will deal with count data and their probability model—a nonnegative integer-valued variable T whose distribution has a single scalar parameter θ to be estimated and/or tested. (Possible additional parameters are supposed to be fixed and known.)

There are many types of tests and CIs for such data. That stems perhaps from the typical impossibility of constructing the uniformly most powerful tests, or CIs with uniform coverage probability (unless the tests and CIs are randomized) as well as from the lack of consensus about the principles and criteria worth following. Moreover, different principles, each of them looking appealing, may prove to be mutually incompatible, and it remains up to an individual choice which one should be preferred.

The binomial distribution is a archetypal example to illustrate the multiplicity of methods. It is known (see, e.g., Brown et al. 2001) that CIs for the binomial proportion cannot have constant coverage probability over the whole parameter space $[0, 1]$, or equivalently, a nonrandomized test of hypothesis $\theta = \theta_0$ cannot have a constant type I error rate over all θ_0 .

This explains the diversity why some authors require coverage probabilities being never below the nominal confidence level and test sizes never exceeding the nominal α (Schilling and Doi 2014; Andres and Tejedor 2004; Reiczigel et al. 2008) whereas others advocate approximate methods and require that actual coverage or type one error rate is close to the nominal on average (Agresti and Coull 1998).

The approximate CI referred to nowadays as standard or Wald CI dates back to Laplace (1812) and there is a quite long list of alternative proposals (see Brown et al. (2001), for a review) trying to overcome some drawbacks of the original method (primarily its too low coverage probability near the extremes of the parameter space).

The first exact interval for the binomial proportion was proposed by Clopper and Pearson (1934), but it was later strongly criticized for being far too conservative, or equivalently, much longer than necessary, in case of two-sided intervals. Later on other authors developed less conservative exact intervals for the two-sided case.

Sterne (1954) proposed a principle leading to the shortest possible confidence sets but, unfortunately, it was shown by Crow (1956) that the *sets may not be connected*. The natural remedy is to transform them into intervals by filling the gaps. Reiczigel (2003) demonstrated that the resulting intervals are still fairly short. The length optimality, nevertheless, is lost.

Blyth and Still (1983) and Casella (1986) proposed a method of obtaining exact CIs of minimum length, but these exhibit a problem of *not being nested*: a CI at a lower level may not be a subset of a higher level interval. In terms of testing this means that the same hypothesis may be rejected at significance level α_1 , but remains not rejected at a level of $\alpha_2 > \alpha_1$. Blaker (2000) proved that exact binomial CIs *cannot* be length-optimal and nested at the same time. His intervals are subsets of the Clopper–Pearson intervals, nested, and nearly length-optimal.

The above review of binomial CIs is by no means complete [see, e.g. (Schilling and Doi 2014; Lecoutre and Poitevineau 2016)].

2.2 Eliminating inconsistencies: methodological proposal

Our aim is to set up a framework for constructing exact two-tailed tests and exact two-sided CIs for a real-valued parameter θ of the distribution of a variable T on $\{0, 1, 2, \dots\}$. Here are three simple and natural criteria that we regard as important.

- (I) The confidence sets are intervals.
- (II) Confidence sets computed from the same sample at different confidence levels are nested: a confidence set at a higher confidence level contains another one at a lower level.
- (III) The test and the confidence set invert each other, that is, the test rejects the hypothesis $\theta = \theta_0$ at significance level α if and only if θ_0 is not in the $1 - \alpha$ confidence set.

When (II) and (III) hold, *p value functions* (also referred to as evidence functions, CI functions, etc.—see Hirji (2006)) $f_k(\theta)$, one for each possible value k of T , may be defined as

$$f_k(\theta) = \sup\{\alpha; \theta \text{ is in the } 1 - \alpha \text{ confidence set for } T = k\} \quad (1)$$

or, equivalently,

$$f_k(\theta_0) = \inf\{\alpha; \theta = \theta_0 \text{ rejected at significance level } \alpha \text{ when } T = k\}.$$

Equivalence of the above two definitions follows from nestedness (criterion II).

On the other hand, assume that some functions $f_k(\theta)$ with values in $[0, 1]$ are given. Let us define a test based on these functions as

$$\text{For } T = k, \text{ reject } \theta = \theta_0 \text{ at significance level } \alpha \text{ if and only if } f_k(\theta_0) \leq \alpha \quad (2)$$

and also a level $1 - \alpha$ confidence set as $\{\theta; f_k(\theta) > \alpha\}$. Then the test and confidence sets satisfy (II) and (III).

Fig. 1 A portion of Sterne's p value function for a binomial proportion given $k = 8$ successes out of $n = 100$ trials. Testing for $H_0 : \theta = 0.15$ results in rejection, whereas a 95% CI contains the value of 0.15

However, (II) and (III) cannot guarantee that also (I) holds. In order to explain the fact, we need the concept of function unimodality, and since the definitions in the literature slightly vary, we will define it.

Definition 1 Function $f(x)$ defined on a nonempty set of reals I is *unimodal* if for every such triplet x_1, x_2, x_3 of arguments that $x_1 < x_2 < x_3$, the inequality $f(x_2) \geq \min(f(x_1), f(x_3))$ holds, and *strictly unimodal*, if the inequality is strict. Function $f(x)$ is *reversely unimodal* (*strictly reversely unimodal*) if $-f(x)$ is unimodal (strictly unimodal).

(The property of being unimodal is referred to elsewhere as bimonotonicity, see Hirji (2006), or quasiconcaveness, see Klaschka (2010).)

It is easy to show that a p value function $f_k(x)$ defined on an interval is unimodal if and only if the sets $\{x; f_k(x) > a\}$ for all $a \in \mathbb{R}$ are intervals. So, when $f_k(x)$ is *not unimodal*, it yields, in some cases (for some α values), non-connected confidence sets $\{\theta; f_k(\theta) > \alpha\}$ and it is natural to substitute them with their convex hulls. After that, (I) holds, and (II) as well (since the operation of convex hull preserves the set inclusion). However, (III) is violated: When θ_0 lies in the convex hull of $\{\theta; f_k(\theta) > \alpha\}$ but not in the set itself, hypothesis $\theta = \theta_0$ is rejected for $T = k$. That is what Fay (2010b) calls *unavoidable inconsistencies* (in contrast with other inconsistencies between test results and CIs that result from applying side by side two different types of test and CI).

This happens for example in the binomial case if testing for $H_0 : \theta = 0.15$ at level $\alpha = 0.05$ by Sterne's test and observing $k = 8$ successes in a sample of $n = 100$. The resulting p value is $p = 0.0496$, whereas the 95% Sterne CI is $(0.0375, 0.1534)$. Figure 1 illustrates the situation.

A closer look at the problem of "unavoidable" inconsistencies reveals, however, that they *can be avoided*. It only requires that the p value functions are derived from the convex hulls of the (potentially non-connected) confidence sets by (1). Since derived now from nested intervals, the modified p value functions are unimodal. Of course, the test should be modified accordingly, that is, derived from the modified p value

functions via (2). That is a way—and, in our opinion, the most reasonable one—to fulfil all three conditions (I), (II), and (III).

More explicitly, the modified p value function $f_k^*(\theta)$ (the uniformly least unimodal function satisfying $f_k^*(\theta) \geq f_k(\theta)$) may be written as

$$f_k^*(\theta) = \min \left\{ \sup_{\xi \leq \theta} f_k(\xi), \sup_{\xi \geq \theta} f_k(\xi) \right\} \quad (3)$$

Formula (3) still cannot be utilized directly for calculation. In Sect. 3 we will give practical recipes how to calculate the modified p values in some important cases.

3 Computational algorithms

In this section we describe the principles that the algorithms are based on, and then specify all details of the modified Sterne and Blaker methods with respect to the binomial and Poisson distributions.

3.1 Methods

Let T be a nonnegative integer-valued random variable whose distribution belongs to a family with a single parameter $\theta \in \Theta$, the parameter space Θ being an interval of reals.

Sterne's method (the probability based method by Hirji (2006)) is based on p value functions

$$f_k^S(\theta) = \sum_{P_\theta(T=i) \leq P_\theta(T=k)} P_\theta(T=i). \quad (4)$$

Another method of our interest will be Blaker's method (by Hirji (2006), the combined tails method). Its p value function for $T = k$ (referred to as *acceptability* in Blaker 2000) is defined as

$$f_k^B(\theta) = \begin{cases} P_\theta(T \geq k) + P_\theta(T \leq k_\theta^*) & \text{if } P_\theta(T \geq k) < P_\theta(T \leq k), \\ 1 & \text{if } P_\theta(T \geq k) = P_\theta(T \leq k), \\ P_\theta(T \leq k) + P_\theta(T \geq k_\theta^{**}) & \text{if } P_\theta(T \geq k) > P_\theta(T \leq k), \end{cases}$$

where $k_\theta^* = \max\{i; P_\theta(T \leq i) \leq P_\theta(T \geq k)\}$, and $k_\theta^{**} = \min\{i; P_\theta(T \geq i) \leq P_\theta(T \leq k)\}$.

Both Sterne's and Blaker's methods guarantee (Blaker 2000) that confidence sets $\{\theta; f_k(\theta) > \alpha\}$ are nested and exact, and all the more so their convex hulls (CIs).

Moreover, Blaker's p value function $f_k^B(\theta)$ is always less or equal to the Clopper–Pearson p value function

$$f_k^{CP}(\theta) = \min\{2P_\theta(T \geq k), 2P_\theta(T \leq k), 1\},$$

so that the Blaker CI must be a subset of the Clopper–Pearson one. Note that an analogous statement does not hold for the Sterne p value function and the corresponding CIs.

We will specifically apply the Sterne and Blaker methods to the binomial $\text{Bin}(n, \theta)$ (n fixed) and Poisson $\text{Poi}(\theta)$ distributions.

3.2 Theoretical background of the algorithms

The following two Lemmas hold for the binomial and Poisson distributions.

Lemma 1 *Let T have either $\text{Bin}(n, \theta)$, or $\text{Poi}(\theta)$ distribution. Then for i, j from the support of T , $i < j$, there is a unique solution $\theta = \sigma_{ij}^S$ to*

$$P_\theta(T = i) = P_\theta(T = j) > 0. \quad (5)$$

The difference $P_\theta(T = j) - P_\theta(T = i)$ changes its sign from negative to positive at σ_{ij}^S , and $\sigma_{ij}^S < \sigma_{rs}^S$, whenever $i \leq r$, $j \leq s$ and $\{i, j\} \neq \{r, s\}$.

Proof Simple algebra. $\square$

Lemma 2 *Let T have either $\text{Bin}(n, \theta)$, or $\text{Poi}(\theta)$ distribution. Then for i, j from the support of T , $i < j$, there is a unique solution $\theta = \sigma_{ij}^B$ to*

$$P_\theta(T \leq i) = P_\theta(T \geq j) > 0.$$

The difference $P_\theta(T \geq j) - P_\theta(T \leq i)$ changes its sign from negative to positive at σ_{ij}^B , and $\sigma_{ij}^B < \sigma_{rs}^B$, whenever $i \leq r$, $j \leq s$ and $\{i, j\} \neq \{r, s\}$.

Proof The Lemma is a trivial consequence of the facts that the functions $P_\theta(T = i)$ of θ are continuous, and both distribution families are stochastically increasing, i.e. $P_\theta(T \leq i)$ is a strictly decreasing function of θ , unless i is the maximum of the support of T .

Due to Lemmas 1 and 2, both Sterne's and Blaker's p value functions may be described in a unified manner. Though dealing in detail just with the binomial and Poisson distributions, we present the description so that "the door is left open" for a broader class of distributions. Thus, we will write the minimum of the support of variable T as $k_{\min}$ (instead of 0), and maximum, if one exists, as $k_{\max}$ (instead of n in case of $\text{Bin}(n, \theta)$). Symbols f_k and σ stand either for f_k^S and σ^S , or f_k^B and σ^B . So, in case of a finite support of T the p value function may be described as

$$f_k(\theta) = \begin{cases} P_\theta(T \geq k) & \text{if } \theta < \sigma_{k_{\min}, k}, k \neq k_{\min}, \\ P_\theta(T \geq k) + P_\theta(T \leq i) & \text{if } \sigma_{ik} \leq \theta < \sigma_{i+1, k}, k_{\min} < i < k - 1, \\ & \text{if } \sigma_{k-1, k} \leq \theta \leq \sigma_{k, k+1}, k_{\min} \neq k \neq k_{\max}, \\ & \text{or } \theta \leq \sigma_{k, k+1}, k = k_{\min}, \\ & \text{or } \theta \geq \sigma_{k-1, k}, k = k_{\max}, \\ P_\theta(T \leq k) + P_\theta(T \geq i) & \text{if } \sigma_{k, i-1} < \theta \leq \sigma_{ki}, k+1 < i < k_{\max}, \\ P_\theta(T \leq k) & \text{if } \theta > \sigma_{k, k_{\max}}, k \neq k_{\max}, \end{cases} \quad (6)$$

Fig. 2 Sterne p value function for a binomial proportion given $k = 2$ successes out of $n = 15$ trials. The locations of jumps $\sigma_{i,k}$ are parameter values where $P_{\sigma_{i,k}}(T = i) = P_{\sigma_{i,k}}(T = k)$ for some $i \neq k$, thus the critical region of the test $\{i : P_{\theta}(T = i) \leq P_{\theta}(T = k)\}$ changes

which, provided that the support of T goes to infinity, simplifies to

$$f_k(\theta) = \begin{cases} P_{\theta}(T \geq k) & \text{if } \theta < \sigma_{k_{\min},k}, k \neq k_{\min}, \\ P_{\theta}(T \geq k) + P_{\theta}(T \leq i) & \text{if } \sigma_{ik} \leq \theta < \sigma_{i+1,k}, k_{\min} < i < k-1, \\ 1 & \text{if } \sigma_{k-1,k} \leq \theta \leq \sigma_{k,k+1}, k \neq k_{\min}, \\ & \text{or } \theta \leq \sigma_{k,k+1}, k = k_{\min}, \\ P_{\theta}(T \leq k) + P_{\theta}(T \geq i) & \text{if } \sigma_{k,i-1} < \theta \leq \sigma_{ki}. \end{cases} \quad (7)$$

We can see that for the binomial and Poisson distribution, and, more generally, for any distribution given that (6) or (7) holds and the likelihood functions $P_{\theta}(T = k)$ are continuous, both Sterne's and Blaker's p value functions $f_k(\theta)$ are piecewise continuous, and the only discontinuities are jumps upwards at $\{\sigma_{ik}; i < k\}$ and downwards at $\{\sigma_{ki}; i > k\}$. At each discontinuity point each of the p value functions has both one-sided limits, and takes the higher of the limits as its value. An example is presented in Fig. 2.

Note that in case of the Blaker method, formulas (6) or (7) apply to all stochastically increasing families, and analogous ("mirror-reflected") ones to the stochastically decreasing families. However, the same is not true for the Sterne method—for example in case of the geometric distribution (5) has no solution, and (4) leads to one-sided tests and CIs.

The following pair of Lemmas on the shape and maxima of the continuous parts of the p value functions is important for numerical calculations.

Lemma 3 *Let T be distributed as either $\text{Bin}(n, \theta)$, or $\text{Poi}(\theta)$, and let $f_k(\theta)$ stand for $f_k^S(\theta)$, or $f_k^B(\theta)$. When $f_k(\theta)$ is continuous and different from the unit constant on interval I , then it is strictly reversely unimodal on I .*

Proof From (6) and (7) it follows that under the assumptions of the Lemma, $f_k(\theta) = 1 - \sum_{i \in A} P_{\theta}(T = i)$ on I , where A is a nonempty finite proper subset of the support of T , and the elements of A are consecutive integers. It is easy to show—see details in the "Appendix"—that $g(\theta) = \sum_{i \in A} P_{\theta}(T = i)$ is then strictly unimodal in the whole parameter space, and thus also on I . Since $f_k(\theta) = 1 - g(\theta)$ on I , $f_k(\theta)$ is strictly reversely unimodal on I . $\square$

As a consequence of the above Lemma, continuous parts of the p value functions (for the given methods and distributions) are—roughly speaking—maximized on the

interval boundaries. However, since not only closed intervals should be considered, the following Lemma states the fact in a more careful way.

Lemma 4 *Let T be distributed as either $\text{Bin}(n, \theta)$, or $\text{Poi}(\theta)$, and let $f_k(\theta)$ stand for $f_k^S(\theta)$, or $f_k^B(\theta)$. When $f_k(\theta)$ is continuous on interval I with bounds $a < b$, then*

$$\sup_{\theta \in I} f_k(\theta) = \max\left\{\lim_{\theta \rightarrow a^+} f_k(\theta), \lim_{\theta \rightarrow b^-} f_k(\theta)\right\}.$$

Unless $f_k(\theta)$ equals the unit constant on I , $f_k(\theta)$ is smaller than the supremum over I at every point of the interior of I .

Proof A trivial consequence of Lemma 3. $\square$

3.3 p value function modification

Formula (3) defines a modified p value $f_k^*(\theta)$ at θ through suprema of $f_k(\cdot)$ over potentially infinite sets $\{\xi; \xi \leq \theta\}$ and $\{\xi; \xi \geq \theta\}$, and so it cannot be used directly as a calculation recipe. However, for the distributions and methods of our interest, the modified p value calculation may be, as shown below, reduced to taking maximum of two numbers.

Proposition 1 *Let T be distributed as either $\text{Bin}(n, \theta)$, or $\text{Poi}(\theta)$, and let $f_k(\theta)$ and σ_{ij} stand either for $f_k^S(\theta)$ and σ_{ij}^S , or for $f_k^B(\theta)$ and σ_{ij}^B . Then the modified p value function defined by (3) can be written equivalently as*

$$f_k^*(\theta) = \begin{cases} \max\{f_k(\theta), f_k(\sigma_{ik})\} & \text{if } \sigma_{ik} \leq \theta < \sigma_{i+1,k}, i < k-1, \\ \max\{f_k(\theta), f_k(\sigma_{ki})\} & \text{if } \sigma_{k,i-1} < \theta \leq \sigma_{ki}, i > k+1, \\ f_k(\theta) & \text{otherwise.} \end{cases}$$

The proof of the Proposition is elementary in case of the Blaker method, but in case of the Sterne method it is harder and based on the following Lemma.

Lemma 5 *Let T be distributed as either $\text{Bin}(n, \theta)$, or $\text{Poi}(\theta)$. Let $f_k^{S-}(\theta)$ denote the smaller one of the one-sided limits of f_k^S at discontinuity point θ . Let θ_1, θ_2 be a pair of discontinuity points of $f_k^S(\theta)$ so that either $\theta_1 = \sigma_{ik}^S, \theta_2 = \sigma_{jk}^S$, where $k > i > j$, or $\theta_1 = \sigma_{ki}^S, \theta_2 = \sigma_{kj}^S$, where $k < i < j$. Then*

$$f_k^S(\theta_1) > f_k^S(\theta_2), \quad (8)$$

and

$$f_k^{S-}(\theta_1) > f_k^{S-}(\theta_2). \quad (9)$$

Proof See the ‘‘Appendix’’. $\square$

Proof of Proposition 1 We may assume that $\sigma = \sigma_{ki}, i > k$ —case $\sigma = \sigma_{ik}, i < k$ is analogous.

Fig. 3 Original (solid line) and modified (dotted line) Sterne p value function in the example shown in Fig. 2 ($k = 2$ successes out of $n = 15$ trials). The displayed section is that from $\sigma_{2,8}$ to $\sigma_{2,9}$

Due to Lemma 4, $\max\{f_k(\theta), f_k(\sigma_{ki})\}$ maximizes f_k over $[\theta, \sigma]$, and what only remains to prove is that $f_k(\xi) < f_k(\sigma)$ for $\xi > \sigma$. In case of the Sterne method, that follows from Lemmas 4 and 5: Due to Lemma 4, supremum of $f_k^S(\xi)$ over $\{\xi; \xi > \sigma\}$ equals one of the one-sided limits of $f_k^S(\xi)$ at the boundaries of the continuous segments to the right of σ . Due to Lemma 5 (and the fact that there is a jump downwards at σ), however, all these are smaller than $f_k^S(\sigma)$. In case of the Blaker method, obviously $f_k^B(\sigma) = 2P_\sigma(T \leq k)$, and $f_k^B(\xi) \leq 2P_\xi(T \leq k)$ for all ξ , which together yields $f_k^B(\xi) < f_k^B(\sigma)$ for $\xi > \sigma$.

Thus, the proof of the Proposition is complete. $\square$

Due to Proposition 1 a common framework of the modified p value calculation for all four cases of interest (combinations of method and distribution) can be outlined as follows:

1. Input data: θ_0 (hypothetical parameter value), k (number of successes), in case of binomial distribution n (number of trials).
2. Identify the continuous segment of $f_k(\theta)$ where θ_0 belongs.
3. Calculate $f_k(\theta_0)$.
4. If $f_k(\theta_0) = 1$ or θ_0 belongs to the leftmost or rightmost segment, output $f_k(\theta_0)$ and finish. Otherwise continue.
5. Put $\sigma := \sigma_{ik}$ if $\sigma_{ik} \leq \theta_0 < \sigma_{i+1,k}$, and $\sigma := \sigma_{ki}$ if $\sigma_{k,i-1} < \theta_0 \leq \sigma_{ki}$.
6. Calculate $f_k(\sigma)$, output $\max\{f_k(\theta_0), f_k(\sigma)\}$, and finish.

Remarks:

- In case of the Sterne method, the discontinuity points σ_{ik} and σ_{ki} can be calculated algebraically. Thus the easiest (though not the only) way of performing Step 2 is to calculate these points and compare θ_0 with them.
- In case of the Blaker method, point σ_{ij} can only be calculated numerically as the root of equation $P_\theta(T \leq i) = P_\theta(T \geq j)$ in θ . Thus, it appears better to perform Step 2 by determining the order number of the continuous segment first, and only then to calculate numerically point σ from Step 5. For instance, when $\sigma_{ik} \leq \theta_0 < \sigma_{i+1,k}$, index i can be found as $y - 1$ where y is α -quantile of the corresponding distribution ($\text{Bin}(n, \theta_0)$, or $\text{Poi}(\theta_0)$), and $\alpha = P_{\theta_0}(T \geq k)$.

Figure 3 illustrates the relationship between the original and modified Sterne p value function using the same example as was shown in Fig. 2.

Programs in R for the Sterne method and both distributions may be downloaded from www2.univet.hu/users/jreiczig/SterneTestAndCI. Optional modifications of the Blaker p value functions for both binomial and Poisson distributions are implemented (and referred to as *unimodalization*) in R package `BlakerCI`.

3.4 Confidence interval construction

Computational algorithms for confidence bounds presented in this section rely on the same properties of p value functions $f_k(\theta)$ as the p value modifying algorithms of Sect. 3.3. Note that the lower and upper $1 - \alpha$ confidence bounds for $T = k$ are the infimum and supremum, resp., of $\{\theta; f_k(\theta) > \alpha\}$ or, equivalently, of $\{\theta; f_k^*(\theta) > \alpha\}$ where functions $f_k^*(\theta)$ are derived from $f_k(\theta)$ by modification (3).

While the algorithms modifying the Sterne and Blaker p value functions were outlined within a common framework in Sect. 3.3, we find it better to treat now each method separately.

The search for the *lower* Sterne's $1 - \alpha$ confidence bound for the binomial or Poisson parameter θ may be outlined as follows.

1. Input parameters: α ($1 -$ confidence level), k (number of successes), in case of binomial distribution n (number of trials), ε (numerical tolerance).
2. If $k = 0$, return 0 and finish. Otherwise continue.
3. Find smallest i such that for $\sigma = \sigma_{ik}^S$ inequality $f_k^S(\sigma) > \alpha$ holds.
Calculate $f^- := f_k^S(\sigma) - P_\sigma(T = k)$.
Remark: Right- and left-sided limits at discontinuity point σ .
4. If $f^- \leq \alpha$, return σ and finish. Otherwise continue.
Remark: If condition is fulfilled, $f_k^S(\theta) \leq \alpha$ for all $\theta < \sigma$ due to Lemmas 4 and 5.
5. If $i = 0$, set $\theta_{low} := 0$, otherwise $\theta_{low} := \sigma_{i-1,k}^S$. Set $\theta_{upp} := \sigma$.
Remark: Level α is exceeded within $(\sigma_{i-1,k}^S, \sigma)$ (or $(0, \sigma)$ when $i = 0$), and due to Lemmas 4 and 5 nowhere to the left of the interval.
6. Set $\theta_{mid} := (\theta_{low} + \theta_{upp})/2$.
If $f_k^S(\theta_{mid}) > \alpha$, put $\theta_{upp} := \theta_{mid}$, otherwise put $\theta_{low} := \theta_{mid}$.
Remark: This iterative interval halving works safely due to Lemma 3.
7. Iterate previous step, until $\theta_{upp} - \theta_{low} < \varepsilon$. Then output θ_{low} and finish.

The *upper* Sterne's confidence bound calculation is analogous (with obvious omissions in the Poisson case).

The above algorithm is novel, and corresponding programs in R for the binomial and Poisson distributions can be downloaded from www2.univet.hu/users/jreiczig/SterneTestAndCI.

The algorithm for the Blaker confidence bounds calculation makes use of the following Lemma.

Lemma 6 *Let T be distributed as $\text{Bin}(n, \theta)$, or $\text{Poi}(\theta)$. The lower (upper) Blaker's $1 - \alpha$ confidence bound ($0 < \alpha < 1$) for θ lies in $[\theta_1, \theta_2]$ ($[\theta_2, \theta_1]$) where θ_1 is the lower (upper) Clopper–Pearson $1 - \alpha$ confidence bound, and θ_2 is the closest to θ_1 discontinuity point of $f_k^B(\theta)$ such that $\theta_2 > \theta_1$ ($\theta_2 < \theta_1$).*

Proof Let us consider the case of the lower confidence limit (the other case is analogous), and assume that $k > 0$ (case $k = 0$ is trivial). Clearly, $f_k^B(\theta) \leq 2P_\theta(T \geq k)$ for all θ . Equality holds at the discontinuity points σ_{ik}^B , $i < k$ (and so for $\theta = \theta_2$). The right-hand side is a strictly increasing function of θ , and equals α for $\theta = \theta_1$ (since θ_1 is a Clopper–Pearson confidence bound). Then, $f_k^B(\theta) \leq 2P_{\theta_1}(T \geq k) = \alpha$ for $\theta \leq \theta_1$, and $f_k^B(\theta_2) = 2P_{\theta_2}(T \geq k) > \alpha$. Thus, $\theta_1 \leq \inf\{\theta; f_k^B(\theta) > \alpha\} \leq \theta_2$. $\square$

The search for the *lower* confidence bound proceeds as follows.

1. Input parameters: α ($1 - \text{confidence level}$), k (number of successes), in case of binomial distribution n (number of trials), ε (numerical tolerance).
2. If $k = 0$, return 0 and finish. Otherwise continue.
3. Calculate θ_1 as the Clopper–Pearson lower $1 - \alpha$ confidence bound.
Remark: A single call of the beta distribution quantile function (binomial case), or of the chi-square quantile function (Poisson case).
4. Find i as the smallest j such that $\theta_1 < \sigma_{jk}^B$.
Remark: Calculated as α_1 -quantile of $\text{Bin}(n, \theta_1)$ (binomial case) or $\text{Poi}(\theta_1)$ (Poisson case), where $\alpha_1 = P_{\theta_1}(T \geq k)$. (Note that σ_{ik}^B solves $P_\sigma(T \geq k) = P_\sigma(T \leq i)$ in σ .)
5. Initialize θ_{low} as θ_1 , and θ_{upp} as θ_{ML} , the maximum likelihood estimate of θ (i.e. k/n for the binomial, and k for the Poisson distribution).
Remark: By Lemma 6, the confidence bound must lie in $[\theta_1, \sigma_{ik}^B]$. Though θ_{upp} is initialized outside the interval, convergence beyond σ_{ik}^B is not allowed in further Steps.
6. Set $\theta_{mid} := (\theta_{low} + \theta_{upp})/2$.
If $f_k^B(\theta_{mid}) > \alpha$ or $\theta_{mid} \geq \sigma_{ik}^B$, set $\theta_{upp} := \theta_{mid}$.
Otherwise put $\theta_{low} := \theta_{mid}$.
Remark: Interval halving, slightly modified by $\theta_{mid} \geq \sigma_{ik}^B$ checking. The condition is checked as $P_{\theta_{mid}}(T \geq k) \leq P_{\theta_{mid}}(T \leq i)$.
7. Iterate Step 6 until $\theta_{upp} - \theta_{low} < \varepsilon$. Then return θ_{low} , and finish.

The *upper* confidence bound calculation is analogous. (Of course, for the Poisson distribution, Step 2 of the above algorithm has no counterpart.)

It should be remarked that setting $\theta_{low} = \theta_{ML}$ in Step 5 is justified by the fact that $f_k^B(\theta_{ML}) = 1$. This follows from inequalities $P_{\theta_{ML}}(T \geq k) > 1/2$ and $P_{\theta_{ML}}(T \leq k) > 1/2$ —see e.g. Neumann (1966, 1970) for a proof.

The above algorithms for the binomial and Poisson distributions are implemented in R package `BlakerCI`.

4 Discussion

Beyond test modification to eliminate inconsistencies between tests and CIs, we presented algorithms for the calculation of modified p values and confidence bounds. Our algorithms exhibit considerable improvements of speed and precision compared to those by Blaker (2000) and Reiczigel (2003), where the p value function is evaluated on an equidistant grid in the parameter space to find the smallest and greatest

values not rejected by the test at level α . The fixed-grid algorithms clearly have a poor precision – speed tradeoff and, moreover, the resolution of the grid does not imply the same precision for the CI endpoints (see Klaschka (2010) for details.)

Fay (2010a, Online supplementary material) was aware of the deficiencies of the fixed grid algorithms, and designed for R package `exactci` a more sophisticated one. Each of the confidence limits is localized in an interval of length ε (input parameter) by stepwise partitioning of the parameter space.

The algorithm has much better precision—speed tradeoff than the fixed-grid algorithms, but still it is much slower than the algorithms of Sect. 3.4 (for speed comparisons in the Blaker case see Klaschka 2010). The reason lies in ignoring facts stated by Lemmas 3, 4, and 6. Furthermore, the algorithm suffers from precision problems in case of the Blaker method. Namely, although discontinuity points of the p value function are calculated numerically with given precision ε , an error less than ε in the discontinuity point may lead to a much larger error in the confidence limit.

An algorithm for Blaker’s confidence limits similar to that of Sect. 3.4 was proposed by Lecoutre and Poitevineau (2016). It is based on the properties of p value functions analogous to those mentioned in Lemmas 3 and 6, and uses interval halving within the interval bounded by the Clopper–Pearson confidence limit from one side, and a p value function discontinuity point from the other side. The algorithm applies to more distributions than the one in this paper. On the other hand, the Sterne case is not covered.

The reason why inconsistencies between tests and CIs persist is that a non-connected confidence set is incontestably annoying, whereas when testing a hypothesis like $\theta = \theta_0$, one doesn’t feel necessary to test other values in the neighbourhood of θ_0 at the same time. Thus, rejecting a parameter value which lies between two non-rejected ones is not at all conspicuous, so it remains in most cases unnoticed. Moreover, the original p value functions are given by a easy-to-calculate comprehensible formula, and modifying it (which may require, as shown in Sect. 3.3, application of numerical methods), represents some inconvenience.

Of course, a more efficient way of enforcing our proposal would consist in convincing software developers rather than users. In this respect, however, we do not expect an immediate victory. Even Fay (2010a, b), who addressed a more frequent and much less subtle kind of inference inconsistencies, does not seem to have been generally heard out so far. For instance, `poisson.test(5, r = 1.8)` in R still results, nowadays, in a (Sterne test) p value of 0.036 but the 95% (Clopper–Pearson) CI of (1.6, 11.7), as in Fay (2010b).

We have to note that there exists yet another important kind of inconsistencies pertinent to the methods that we are dealing with in this paper. Vos and Hudson (2008) show that, when dealing with binomially distributed data, some methods, including the Sterne’s and Blaker’s one, generate conflicts between inferences from samples of different sizes. For instance, when testing $H_0 : \theta = 0.2$ and observing 2 successes out of 33, Sterne’s test results in a p value of 0.0495. Then one would intuitively expect that with 2 successes out of 34 trials, the same test should reject the same hypothesis “even more resolutely”. The test, however, yields 0.0504 as the p value. Nonetheless, since our opinions on the problem and its possible solutions diverge, we do not address it here, and it is left to be the topic of another paper.

Acknowledgements Open access funding provided by University of Veterinary Medicine (ATE).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix

We will use notation

$$b(i, n, \theta) = \binom{n}{i} \theta^i (1 - \theta)^{n-i}, \quad n \geq 1, \quad 0 \leq i \leq n, \quad 0 \leq \theta \leq 1$$

(defining $0^0 = 1$ so that $b(0, n, 0) = b(n, n, 1) = 1$). Further we will denote the Poisson probabilities

$$\pi(i, \theta) = \frac{\theta^i}{i!} e^{-\theta}, \quad i \geq 0, \quad \theta \geq 0$$

(defining $0^0 = 1$, so that $\pi(0, 0) = 1$).

Proof (Details of the Proof of Lemma 3) Binomial case: Let $g(\theta) = \sum_{i=x}^y b(i, n, \theta)$ where $0 \leq x \leq y \leq n$, and $\{x, y\} \neq \{0, n\}$. We need to prove that $g(\theta)$ is strictly unimodal on $[0, 1]$. Case $n = 1$ is trivial, and for $x = 0$ or $y = n$ function $g(\theta)$ is evidently strictly monotone, and hence strictly unimodal on $[0, 1]$. For $n \geq 2$, $1 \leq i < n$ clearly

$$\frac{d}{d\theta} b(i, n, \theta) = n[b(i-1, n-1, \theta) - b(i, n-1, \theta)],$$

and thus for $1 \leq x \leq y < n$

$$\frac{d}{d\theta} g(\theta) = n[b(x-1, n-1, \theta) - b(y, n-1, \theta)].$$

Since

$$\frac{b(y, n-1, \theta)}{b(x-1, n-1, \theta)} = \left(\frac{\theta}{1-\theta}\right)^{y-x+1} \prod_{i=x}^y \frac{n-i}{i},$$

a continuous function strictly increasing from 0 to $+\infty$, the derivative of $g(\theta)$ vanishes at a single $\theta^* \in (0, 1)$, where

$$\theta^* = \frac{a}{1+a}, \quad a = \left(\prod_{i=x}^y \frac{i}{n-i} \right)^{\frac{1}{y-x+1}}.$$

Moreover, the derivative is positive or negative on $[0, \theta^*]$ or $[\theta^*, 1]$, resp. Function $g(\theta)$ is thus strictly increasing for $\theta \leq \theta^*$ and strictly decreasing for $\theta \geq \theta^*$, which is evidently sufficient for strict unimodality. This completes the proof for the binomial case.

Poisson case is analogous: Let $g(\theta) = \sum_{i=x}^y \pi(i, \theta)$ where $0 \leq x \leq y < \infty$. We have to prove that $g(\theta)$ is strictly unimodal on $[0, +\infty)$. For $x = 0$, function $g(\theta)$ is clearly strictly decreasing in the whole domain. Otherwise, since for $i \geq 1$

$$\frac{d}{d\theta} \pi(i, \theta) = \pi(i-1, \theta) - \pi(i, \theta),$$

we have for $1 \leq x \leq y < +\infty$

$$\frac{d}{d\theta} g(\theta) = \pi(x-1, \theta) - \pi(y, \theta).$$

Since

$$\frac{\pi(y, \theta)}{\pi(x-1, \theta)} = \frac{\theta^{y-x+1}}{\prod_{i=x}^y i},$$

a continuous function strictly increasing from 0 to $+\infty$, the derivative of $g(\theta)$ vanishes and changes its sign from positive to negative at a single $\theta^* \in (0, +\infty)$, where

$$\theta^* = \left(\prod_{i=x}^y i \right)^{\frac{1}{y-x+1}}.$$

This completes the Poisson case, and the whole proof. $\square$

Proof (Proof of Lemma 5) Clearly, it is sufficient to deal with the case $k < i < j$ (the other case is analogous), and assume $j = i + 1$.

Function $P_\theta(T = k)$ of θ is decreasing to the right of the maximum likelihood estimate of θ , and therefore on $[\theta_1, \theta_2]$. Thus, since

$$f_k^S(\theta_1) = f_k^{S-}(\theta_1) + P_{\theta_1}(T = i) = f_k^{S-}(\theta_1) + P_{\theta_1}(T = k),$$

and

$$f_k^S(\theta_2) = f_k^{S-}(\theta_2) + P_{\theta_2}(T = i + 1) = f_k^{S-}(\theta_2) + P_{\theta_2}(T = k),$$

(8) follows from (9), and it is sufficient to prove (9).

So we have to prove

$$P_{\theta_1}(T \leq k) + P_{\theta_1}(T > i) > P_{\theta_2}(T \leq k) + P_{\theta_1}(T > i + 1),$$

or equivalently

$$P_{\theta_1}(k < T \leq i) < P_{\theta_2}(k < T \leq i + 1),$$

which can also be written as

$$P_{\theta_1}(T = i + 1) > P_{\theta_1}(k + 1 \leq T \leq i + 1) - P_{\theta_2}(k + 1 \leq T \leq i + 1). \quad (10)$$

We will conclude the proof by showing that function $P_{\theta}(k + 1 \leq T \leq i + 1)$ of θ is increasing on (θ_1, θ_2) , and so the right-hand side of (10) is negative.

The rest of the proof is split into the binomial and Poisson cases.

Binomial case: Let T be distributed as $\text{Bin}(n, \theta)$, $n > 1$. From

$$\frac{d}{d\theta} b(r, n, \theta) = n(b(r - 1, n - 1, \theta) - b(r, n - 1, \theta)),$$

(putting formally $b(n, n - 1, \theta) = 0$), we have

$$\frac{d}{d\theta} P_{\theta}(k + 1 \leq T \leq i + 1) = n[b(k, n - 1, \theta) - b(i + 1, n - 1, \theta)].$$

Case $i + 1 = n$ is clear. Otherwise, since $b(i + 1, n - 1, \theta)/b(k, n - 1, \theta)$ grows monotonically from zero to infinity on $(0, 1)$, the difference $b(k, n - 1, \theta) - b(i + 1, n - 1, \theta)$ is positive on $(0, \theta^*)$ and negative on $(\theta^*, 1)$ for some $\theta^* \in (0, 1)$. The binomial case will be complete, if $\theta_2 < \theta^*$.

However, by simple algebra we get $\theta^* = c/(1 + c)$ where

$$c^{i-k+1} = \prod_{r=k+1}^{i+1} \frac{r}{n-r}$$

and $\theta_2 = a/(1 + a)$ where

$$a^{i-k+1} = \prod_{r=k+1}^{i+1} \frac{r}{n-r+1},$$

and so $a < c$ and $\theta_2 < \theta^*$ indeed, which concludes the binomial branch of the proof.

Poisson case: When T is distributed as $\text{Poi}(\theta)$,

$$\frac{d}{d\theta} P_{\theta}(k + 1 \leq T \leq i + 1) = \pi(k, \theta) - \pi(i + 1, \theta),$$

and it remains to prove that the right-hand side is positive on (θ_1, θ_2) . The ratio $\pi(i+1, \theta)/\pi(k, \theta)$ is less than 1 and thus $\pi(k, \theta) - \pi(i+1, \theta)$ positive on $(0, \theta^*)$ where θ^* solves $\pi(i+1, \theta) - \pi(k, \theta)$. This, however, shows that $\theta^* = \theta_2$. That concludes the Poisson branch, and the whole proof. $\square$

References

- Agresti A, Coull A (1998) Approximate is better than “exact” for interval estimation of binomial proportions. *Am Stat* 52:119–126
- Andres AM, Tejedor IH (2004) Exact unconditional non-classical tests on the difference of two proportions. *Comput Stat Data Anal* 45(2):373–388
- Blaker H (2000) Confidence curves and improved exact confidence intervals for discrete distributions. *Can J Stat* 28(4):783–798 (**Corrigenda: Canad J Statist** 29(4):681)
- Blyth CR, Still HA (1983) Binomial confidence intervals. *JASA* 78:108–116
- Brown LD, Cai TT, DasGupta A (2001) Interval estimation for a binomial proportion. *Stat Sci* 16:101–133
- Casella G (1986) Refining binomial confidence intervals. *Can J Stat* 14:113–129
- Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26:404–413
- Crow EL (1956) Confidence intervals for a proportion. *Biometrika* 43:423–435
- Fay MP (2010a) Confidence intervals that match Fisher’s exact and Blaker’s exact tests. *Biostatistics* 11:373–374, online supplementary material: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2852239/bin/kxp050_index.html
- Fay MP (2010b) Two-sided exact tests and matching confidence intervals for discrete data. *R J* 2(1):53–58
- Hirji KF (2006) Exact analysis of discrete data. CRC Press, Boca Raton
- Klaschka J (2010) BlakerCI: An algorithm and R package for the Blaker’s binomial confidence limits calculation. Technical Report 1099, Institute of Computer Science, Czech Academy of Sciences
- Laplace PS (1812) *Théorie Analytique des Probabilités*. Courcier, Paris
- Lecoutre B, Poitevineau J (2016) New results for computing Blaker’s exact confidence interval for one parameter discrete distributions. *Commun Stat Simul Comput* 45(3):1041–1053. <https://doi.org/10.1080/03610918.2014.911900>
- Neumann P (1966) Über den Median der Binomial- und Poissonverteilung. *Wissenschaftliche Zeitschrift der Technischen Universität Dresden* 15:223–226
- Neumann P (1970) Über den Median einiger diskreter Verteilungen und eine damit zusammenhängende monotone Konvergenz. *Wissenschaftliche Zeitschrift der Technischen Universität Dresden* 19:29–33
- Reiczigel J (2003) Confidence intervals for the binomial parameter: some new considerations. *Stat Med* 22(4):611–621. <https://doi.org/10.1002/sim.1320>
- Reiczigel J, Abonyi-Toth Z, Singer J (2008) An exact confidence set for two binomial proportions and exact unconditional confidence intervals for the difference and ratio of proportions. *Comput Stat Data Anal* 52:5046–5053
- Schilling M, Doi AJ (2014) A coverage probability approach to finding an optimal binomial confidence procedure. *Am Stat* 68(3):133–145. <https://doi.org/10.1080/00031305.2014.899274>
- Sterne TE (1954) Some remarks on confidence or fiducial limits. *Biometrika* 41:117–129
- Vos PW, Hudson S (2008) Problems with binomial two-sided tests and the associated confidence intervals. *Aust N Z J Stat* 50(1):81–89. <https://doi.org/10.1111/j.1467-842X.2007.00501.x>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.