Statistical Methods in Medical Research 2003; **12**: 3–21

Dealing with discreteness: making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact

A Agresti Department of Statistics, University of Florida, Gainesville, USA

'Exact' methods for categorical data are exact in terms of using probability distributions that do not depend on unknown parameters. However, they are conservative inferentially. The actual error probabilities for tests and confidence intervals are bounded above by the nominal level. This article examines the conservatism for interval estimation and describes ways of reducing it. We illustrate for confidence intervals for several basic parameters, including the binomial parameter, the difference between two binomial parameters for independent samples, and the odds ratio and relative risk. Less conservative behavior results from devices such as (1) inverting tests using statistics that are 'less discrete', (2) inverting a single two-sided test rather than two separate one-sided tests each having size at least half the nominal level, (3) using unconditional rather than conditional methods (where appropriate) and (4) inverting tests using alternative p-values. The article concludes with recommendations for selecting an interval in three situations—when one needs to guarantee a lower bound on a coverage probability, when it is sufficient to have actual coverage probability near the nominal level, and when teaching in a classroom or consulting environment.

1 Introduction: Discreteness and conservatism

Recent years have seen considerable development and extensions of 'exact' small-sample methods for contingency tables. This methodology is useful when one is unwilling to trust the uncertain performance of an inferential method based on a large-sample approximation. See Mehta[1] and Agresti[2] for recent reviews, and *StatXact*[3] for software having the greatest scope for small-sample inference in discrete problems.

The word 'exact' in quotes in the article title and the title of this section refers to methods that use distributions determined exactly rather than as approximations; that is, those distributions do not depend on unknown parameters. However, they are not exact in the sense that inferences based on them have error probabilities exactly equal to the nominal values. Rather, the nominal values are upper bounds for the true error probabilities.

We illustrate with a simple example. For a binomial random variable X with $n = 5$ trials and parameter π, consider a test of $H_0: \pi = 0.5$ against $H_a: \pi \neq 0.5$. Under H_0, the exact distribution of X is binomial with $n = 5$ and parameter 0.5. Now, suppose the outcome is $x = 5$, with a sample proportion $\hat{\pi} = x/n = 1.0$, the maximum likelihood (ML) estimate of π. With 'exact' inference the p-value is the binomial two-tailed

Address for correspondence: Alan Agresti, Distinguished Professor, Department of Statistics, University of Florida, Gainesville, Florida, 32611-8545, USA. E-mail AA@STAT.UFL.EDU

 10.1191/0962280203sm311ra

probability of 0 or 5 outcomes in 5 trials, which is $2(1/2)^5 = 0.0625$. Now let us consider a scientist who believes in the sacredness of a 0.05 significance level, rejecting H_0 only if the p-value of the test is no greater than 0.05. Using that nominal significance level, the scientist cannot reject H_0. However, the actual size (probability of type I error) of the test is not 0.05. Rather, the probability of falsely rejecting H_0 is 0, since when $n = 5$ no possible x provides a p-value below 0.05.

For a large-sample normal approximate test, a test statistic is

$$z = (\hat{\pi} - \pi_0)/\sqrt{\pi_0(1 - \pi_0)/n} = (1.0 - 0.5)/\sqrt{0.5(0.5)/5} = 2.24,$$

which has a two-sided p-value from the standard normal distribution of 0.025. Thus, the test rejects H_0 at the 0.05 level. The p-value for the z test is less than 0.05 only when $x = 0$ or 5; thus, its actual probability of type I error is the binomial probability of these outcomes when $\pi = 0.5$, which is 0.0625. Generally, the 'exact' binomial test has the nominal size as an upper bound for the actual size. The large-sample normal test may have actual size below or above the nominal level. In some cases that size may be closer than the size of the exact test to the nominal level. However, with large-sample approximation the potential also exists of having actual size much above the nominal level, and whether this may happen is more difficult to predict with more complex problems with nuisance parameters. Thus, such approximations are sometimes unacceptable in practice.

Confidence intervals correspond to inverting a family of tests. For instance, a 95% confidence interval for a parameter consists of the set of values not rejected at the 0.05 significance level in a corresponding test. Inverting a family of tests that has actual size no greater than 0.05 for each possible parameter value results in a confidence interval having coverage probability at least equal to 0.95. Thus, conservatism of 'exact' tests propagates to 'exact' confidence intervals, and possibly poor behavior of large-sample tests propagates to large-sample confidence intervals.

For instance, with the best known 'exact' method for interval estimation of a binomial parameter (the 'Clopper–Pearson' method), the 95% confidence interval when $x = 5$ in $n = 5$ trials is (0.478, 1.000). We will see this means that π_0 must be below 0.478 in order for the binomial right-tail probability in testing $H_0: \pi = \pi_0$ against $H_a: \pi > \pi_0$ to fall below 0.025. In fact, when $n = 5$ this 'exact' 95% confidence interval contains 0.5 for *every* value of x. Thus, the actual coverage probability of this 'exact' interval when $\pi = 0.5$ is 1.0, not 0.95. By contrast, inverting the large-sample normal approximate test described above (but with $H_0: \pi = \pi_0$ rather than $H_0: \pi = 0.5$) yields an interval having coverage probability 0.9375 when $\pi = 0.5$, as the interval contains 0.5 when $x = 0$ or $x = 5$.

Conservatism is mainly problematic with small samples. As n increases with individual probabilities approaching 0, actual error probabilities approach nominal levels. The focus of this article is studying ways to reduce the conservatism in 'exact' small-sample interval estimation for some important parameters in categorical data analysis. Section 2 summarizes some remedies for reducing the conservative effects of discreteness. The remainder of the article shows particular cases. Section 3 presents confidence intervals for a binomial proportion. Section 4 presents confidence intervals

for the difference between two binomial proportions with independent samples. Section 5 discusses confidence intervals for the odds ratio in 2×2 tables. Section 6 briefly discusses other cases, including the relative risk. Section 7 summarizes and makes recommendations. We will see that with small samples, substantial improvement can result from reducing the conservatism of 'exact' confidence intervals.

2 The tail method, and remedies for reducing its conservatism

'Exact' inference requires the actual error probability to be no greater than the nominal level, which we denote by α. For a test, the actual size is no greater than α. For a confidence interval, the actual coverage probability is at least $1 - \alpha$ for all possible values of θ. The usual approach to 'exact' interval estimation inverts a family of 'exact' tests having size at most α.

Let T be a discrete test statistic with probability mass function $f(t; \theta)$ and cumulative distribution function $F(t; \theta)$ indexed by a parameter θ. For an 'exact' test, for each value θ_0 of θ let $A(\theta_0)$ denote the acceptance region for testing $H_0: \theta = \theta_0$. This is the set of values t of T for which the p-value exceeds α. Then, for each t, let $C(t) = \{\theta_0: t \in A(\theta_0)\}$. The set of $\{C(t)\}$ for various t are the confidence regions with the desired property. In other words, having acceptance regions such that

$$P_{\theta_0}[T \in A(\theta_0)] \geq 1 - \alpha$$

for all θ_0 guarantees that the confidence level for $\{C(t)\}$ is at least $1 - \alpha$. For a typical θ_0, one cannot form $A(\theta_0)$ to achieve probability of type I error exactly equal to a, because of discreteness. Hence, such confidence intervals are conservative. The actual coverage probability of $C(T)$ varies for different values of θ but is bounded below by $1 - \alpha$ (Neyman[4]) unless one makes an artificial transformation of T to a continuous variable using supplementary randomization.[5]

A common way to construct an 'exact' interval inverts two separate one-sided tests that each have size at most $\alpha/2$. For test statistic T for $H_0: \theta = \theta_0$, let t_0 denote the observed value. Suppose relatively large values of T provide evidence in favor of $H_a: \theta > \theta_0$ and relatively small values provide evidence in favor of $H_a: \theta < \theta_0$. If $F(t; \theta)$ is a strictly decreasing function of θ for each t, the confidence interval (θ_L, θ_U) is defined by solutions to the equations

$$P(T \leq t_o; \theta_U) = \alpha/2, \qquad P(T \geq t_o; \theta_L) = \alpha/2. \tag{1}$$

This method of forming a confidence interval is often called the *tail method*. When T is continuous, method (1) yields coverage probability $1 - \alpha$ at all θ, but when T is discrete $1 - \alpha$ is a lower bound. In technical terms, the bound results from the distribution of $F(T; \theta)$ being stochastically larger than uniform when T is discrete (Casella and Berger,[6] pp. 77, 434).

Inverting a family of tests corresponds to forming the confidence region from the set of θ_0 for which the test's p-value exceeds α. The tail method (1) requires the stronger condition that the probability be no greater than $\alpha/2$ that T falls below $A(\theta_0)$ and no

greater than $\alpha/2$ that T falls above $A(\theta_0)$. The interval for this method is the set of θ_0 for which each one-sided p-value exceeds $\alpha/2$. Equivalently, it corresponds to forming the confidence region from the set of θ_0 for which a single p-value defined as $2 \times \min[P_{\theta_0}(T \geq t_o),\ P_{\theta_0}(T \leq t_o)]$ (but with $p = 1.0$ if this doubling exceeds 1.0) exceeds α.

One disadvantage of the tail method is that for sufficiently small and sufficiently large θ, the lower bound on the coverage probability is actually $1 - \alpha/2$ rather than $1 - \alpha$. For sufficiently small θ, for instance, the interval can never exclude θ by falling below it.

Alternatives to the tail method exist for constructing confidence intervals that are better—the intervals tend to be shorter and coverage probabilities tend to be closer to the nominal level. We now summarize a few of these.

2.1 Confidence intervals based on two-sided tests

One approach to improving interval estimation of θ inverts a single two-sided test instead of two equal-tail one-sided tests. For instance, a possible two-sided p-value is $\min[P_{\theta_0}(T \geq t_o), P_{\theta_0}(T \leq t_o)]$ plus an attainable probability in the other tail that is as close as possible to, but no greater than, that one-tailed probability.[7] This p-value is no greater than that for the tail method. Hence the confidence intervals based on inverting such a test necessarily are contained in confidence intervals obtained with the tail method.

Another two-sided approach forms the acceptance region $A(\theta_0)$ by entering the test statistic values t in $A(\theta_0)$ in order of their null probabilities, starting with the highest, stopping when the total probability is at least $1 - \alpha$; that is, $A(\theta_0)$ contains the smallest possible number of most likely outcomes (under $\theta = \theta_0$). When inverted to form confidence intervals, this approach satisfies the optimality criterion of minimizing total length.[8] A slight complication is the lack of a unique way of forming $A(\theta_0)$. In its crudest partitioning of the sample space it corresponds to testing using the p-value

$$P_{\theta_0}[f(T; \theta_0) \leq f(t_o; \theta_0)], \tag{2}$$

the sum of null probabilities that are no greater than the probability of the observed result. The confidence interval is the set of θ_0 for which

$$P_{\theta_0}[f(T; \theta_0) \leq f(t_o; \theta_0)] > \alpha.$$

In a related approach, Blaker[7] defined $\gamma(t, \theta) = \min[P_\theta(T \geq t), P_\theta(T \leq t)]$ and suggested forming the confidence interval as the set of θ_0 for which

$$P_{\theta_0}[\gamma(T, \theta_0) \leq \gamma(t_o, \theta_0)] > \alpha. \tag{3}$$

This corresponds to a test based on the p-value mentioned above that equals the minimum one-tail probability plus an attainable probability in the other tail that is as close as possible to, but not greater than, that one-tail probability. Blaker showed that, although such intervals may not have length optimality, they necessarily are contained within intervals formed using the tail method. These intervals and the ones based on

length optimality satisfy a nestedness property, in which an interval with larger nominal confidence level necessarily contains one with a smaller nominal level.

An alternative way to invert a two-sided test orders points for the acceptance region and forms p-values according to a statistic that describes the distance of the observed data from H_0. one could use a statistic T based on a standard criterion, such as the likelihood-ratio statistic, the Wald statistic (based on dividing the ML estimate by its standard error) or the score statistic (based on dividing the derivative of the log-likelihood at θ_0 by its standard error). These are the three statistics commonly used for large-sample inference.

These various two-sided approaches do not have the tail method disadvantage of a lower bound of $1 - \alpha/2$ for the coverage probability over part of the parameter space. However, anomalies can occur. For instance, a confidence region based on these two-sided p-values is not necessarily an interval, because the endpoints of the acceptance region need not be monotone in θ_0. See Casella and Berger[6] (p. 431) and Santner and Duffy[9] (p. 37) for discussion of this for the binomial parameter. Unfortunately, no single method for constructing confidence regions with discrete distributions can have optimality simultaneously in the criteria of length, necessarily being an interval, and nestedness.

In some studies a disadvantage of inverting a single two-sided test is non-equivalence with results of one-sided tests, such as tests for whether a new treatment is better than a standard one. For such studies one can argue in favor of simply calculating a one-sided confidence bound instead of a confidence interval.

2.2 Confidence intervals based on less discrete statistics

In constructing a test or a confidence interval based on a test, the test statistic should not be any more discrete than necessary. For instance, consider the binomial parameter π. In testing $H_0: \pi = \pi_0$, one possible criterion for summarizing evidence about H_0 is $\hat{\pi}$. However, this statistic is severely discrete for small samples. It is better to base tests and subsequent confidence intervals on a standardization, such as by dividing it by its null standard error, or the relative likelihood values. Then, in testing $H_0: \pi_1 = 0.4$, for instance, $\hat{\pi} = 0.5$ gives less evidence than $\hat{\pi} = 0.3$ against H_0.

2.3 Confidence intervals based on alternative *p*-values

It is sometimes possible to reduce conservativeness by using a less discrete form of p-value. For instance, Cohen and Sackrowitz[10] and Kim and Agresti[11] based p-values on a finer partition of the sample space than provided by a test statistic T alone, to generate a less discrete sampling distribution for the p-value. A simple way to do this supplements T by the probabilities of the various samples for which T equals the observed value t_o. Instead of including the probabilities of all relevant samples having $T = t_o$ in the p-value, one includes only probabilities of those samples that are no more likely to occur than the observed one. This modified p-value is legitimate, since it satisfies the usual definition of a p-value (Casella and Berger,[6] p. 397)

$$p_{H_0}(p\text{-value} \leq \alpha) \leq \alpha \text{ for } 0 < \alpha < 1. \tag{4}$$

The modified p-value cannot exceed the usual one, whether based on a one-sided or two-sided approach, so a test and confidence interval based on it is less conservative.

2.4 Confidence intervals based on an unconditional approach with nuisance parameters

For comparisons of parameters from two discrete distributions, the joint distribution of the data involves the parameter of interest (e.g., a difference between parameter values for two samples) plus some other parameter(s). These other parameters are nuisance parameters, usually not being of primary interest. When nuisance parameters exist, construction of a confidence interval is more complicated. For 'exact' inference with contingency tables, a popular approach is a conditional one that eliminates nuisance parameters by conditioning on their sufficient statistics. This is the basis of Fisher's exact test and a method that section 5 discusses for constructing a confidence interval for an odds ratio. The conditional approach increases the degree of discreteness, however. In some cases this can result in unacceptable conservatism. It is even possible for a conditional distribution to be degenerate, when only one sample can have the required values of the sufficient statistics. More importantly, the conditional approach is limited to certain parameters (the 'natural parameter' for exponential family distributions). For comparing two binomial distributions, for instance, it is limited to the difference of logits, which is the log odds ratio.

An alternative approach to eliminating the nuisance parameter is unconditional. For a nuisance parameter ψ, let $p(\theta_0; \psi)$ denote the p-value for testing $H_0: \theta = \theta_0$ for a given value of ψ. The unconditional approach takes the p-value to be $\sup_\psi p(\theta_0; \psi)$, the largest over all possible values for the nuisance parameter. This is a legitimate p-value (Casella and Berger,[6] p. 397). As usual, the confidence interval consists of values of θ_0 for which this p-value exceeds α. This approach is also conservative. However, if $p(\theta_0; \psi)$ is relatively stable in ψ, this method has the potential to improve on conditional methods. See, for instance, Suissa and Shuster,[12] who showed improvement in power over Fisher's exact test for testing equality of two independent binomials.

2.5 Almost 'exact' approaches

Our focus in this article is on 'exact' methods for which the nominal confidence level is necessarily a lower bound on the actual level. In practice, it is often reasonable to relax this requirement slightly. Conservativeness can be reduced somewhat if the coverage probability for a confidence interval is allowed to go slightly below $1 - \alpha$ for some θ values.

An increasingly popular way to do this inverts a test using an exact distribution but with the *mid p-value*. This replaces $p(T = t_o)$ in the p-value by $(1/2)p(T = t_o)$. For instance, a one-sided p-value has form $p(T > t_o) + (1/2)P(T = t_o)$. This depends only on the data, unlike the Stevens[5] randomized p-value of form $p(T > t_o) + U \times p(T = t_o)$ where U is a uniform(0, 1) random variable. The randomized p-value achieves the nominal size, and the mid p-value replaces U in it by its expected value. Then, it is possible to exceed the nominal size, but usually not by much. Note that the sum of the one-tailed mid p-values equals 1, whereas for discrete data the sum of the two one-tailed ordinary p-values exceeds 1.

The mid p-value does not necessarily satisfy (4). Intervals based on inverting tests using the mid p-value cannot guarantee coverage probabilities of at least the nominal level. However, evaluations for a variety of problems[3,14] have shown that it still tends to be somewhat conservative, though necessarily less so than using the ordinary p-value. An advantage over ordinary asymptotic methods is that it uses the exact distribution and provides an essentially exact method for moderate sample sizes, since the difference between the mid p-value and ordinary 'exact' p-value diminishes as the sample size increases and the discreteness in the tails diminishes. This recommendation is particularly relevant for the conditional approach, which has greater discreteness than the unconditional approach.

3 Analyses for a binomial proportion

Let T denote a binomial variate for n trials with parameter π, denoted bin(n, π). The tail method (1) gives the most commonly cited 'exact' confidence interval for π, the Clopper–Pearson interval. The endpoints satisfy

$$\sum_{k=t_o}^{n} \binom{n}{k} \pi_L^k (1-\pi_L)^{n-k} = \alpha/2 \text{ and } \sum_{k=0}^{t_o} \binom{n}{k} \pi_U^k (1-\pi_U)^{n-k} = \alpha/2,$$

except that $\pi_L = 0$ when $t_0 = 0$ and $\pi_U = 1$ when $t_o = n$. Various evaluations have shown that this interval tends to be extremely conservative for small to moderate n.[14,15,17] When $t_o = 0$, it equals $[0, 1 - (\alpha/2)^{1/n}]$. The actual coverage probability necessarily exceeds $1 - \alpha/2$ for π below $1 - (\alpha/2)^{1/n}$ and above $(\alpha/2)^{1/n}$. This is the entire parameter space when $n \leq \log(\alpha/2)/\log(0.5)$, for instance $n \leq 5$ for $\alpha = 0.05$.

Sterne[18] proposed inverting a single test by forming the acceptance region with outcomes ordered by their probabilities (i.e., p-value (2)). Blyth and Still[19] and Casella[20] amended this method slightly so that the confidence region cannot contain unconnected intervals and so natural symmetry and invariance properties are satisfied. Blaker[7] discussed intervals based on inverting the test having p-value equal to the minimum tail probability plus the probability no greater than that in the other tail. This yields intervals similar to the Blyth–Still–Casella intervals that are contained within the Clopper–Pearson intervals and are simpler to compute (the Blaker article contains short S-Plus functions for doing this). Unlike the Blyth–Still–Casella intervals, these intervals necessarily have the nestedness property. The Blyth–Still interval is available in *StatXact*.[3]

For any method, the actual coverage probability at a fixed value of π is the sum of the binomial probabilities of all those outcomes t_0 for which the resulting interval covers π. Figure 1 shows the actual coverage probabilities of the Clopper–Pearson and Blaker intervals for nominal 95% confidence intervals, plotted as a function of π, when $n = 10$. This figure illustrates the superiority of forming the confidence interval by inverting a single two-sided test. Table 1 shows the 11 confidence intervals for each method. For comparison, Table 1 also shows the Blyth–Still intervals. These are similar

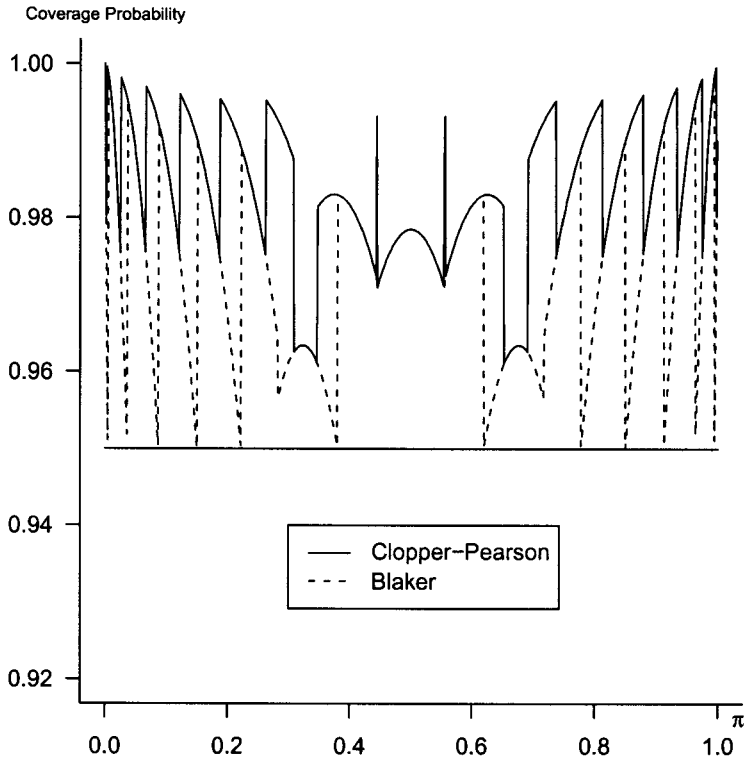

Figure 1 Coverage probabilities for 95% confidence intervals for a binomial parameter π with $n = 10$

Table 1 Nominal 95% confidence intervals for a binomial proportion with t successes in $n = 10$ trials

	Clopper–Pearson interval		Blaker interval		Blyth–Still interval	
t	Lower	Upper	Lower	Upper	Lower	Upper
0	0.000	0.308	0.000	0.283	0.000	0.267
1	0.002	0.445	0.005	0.444	0.005	0.444
2	0.025	0.556	0.037	0.555	0.037	0.556
3	0.067	0.652	0.087	0.619	0.087	0.619
4	0.122	0.738	0.150	0.717	0.150	0.733
5	0.187	0.813	0.222	0.778	0.222	0.778

Note: Blyth–Still intervals were obtained using *StatXact*. For count $6 \leq t \leq 10$, limits equal $(1 - \theta_U, 1 - \theta_L)$ for limits given for $10 - t$.

to the Blaker intervals. As n increases, the conservatism of the Clopper–Pearson interval dies out rather slowly.[16]

Some 'exact' methods (such as Clopper–Pearson) are so conservative that, for applications in which maintaining at least the desired level is not crucial, it may even be preferable to use a good large-sample method rather than that 'exact' method. For estimating a proportion, the most popular large-sample 95% confidence interval is $\hat{\pi} \pm 2\sqrt{\hat{\pi}(1-\hat{\pi})/n}$. This interval is based on inverting results of the Wald test using test statistic $z = (\hat{\pi} - \pi_0)/\sqrt{\hat{\pi}(1-\hat{\pi})/n}$; that is, it is the set of π_0 for which $|z| \leq 2$. Unfortunately, this 'Wald inerval' behaves very poorly; for instance, it yields the degenerate interval (1.0, 1.0) for the example of $x = 5$ in $n = 5$ trials discussed at the beginning of this article, and generally the coverage probabilities tend to be too low even for quite large samples.[17] However, other large-sample intervals behave quite well. The interval based on inverting the test using test statistic $z = (\hat{\pi} - \pi_0)/\sqrt{\pi_0(1-\pi_0)/n}$ (i.e., the score test) has coverage probability that tends to fluctuate around 0.95 except for a couple of low probabilities for π values close to 0 and 1.[16] The adjustment of the Wald interval that first adds two outcomes of each type before computing $\hat{\pi} \pm 2\sqrt{\hat{\pi}(1-\hat{\pi})/n}$ has the same center as the score interval but is slightly wider and tends to be somewhat conservative, but not as much so as the Clopper–Pearson interval.[16] Figure 2 illustrates, showing coverage probabilities when $n = 10$ for the

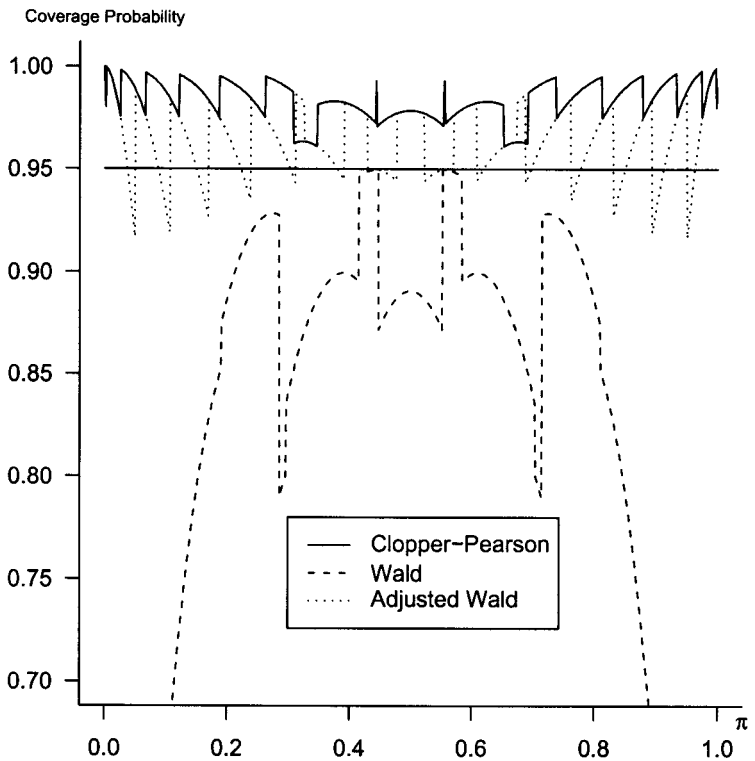

Figure 2 Coverage probabilities for 95% confidence intervals for a binomial parameter π with $n = 10$

very conservative 'exact' Clopper–Pearson interval, the very liberal Wald interval, and the adjusted Wald large-sample interval using an extra two outcomes of each type. Forming an interval by inverting the likelihood-ratio test also works better than the Wald interval.

4 Difference between two binomial parameters

Next consider the difference of proportions for two independent binomial samples, where X_i is bin(n_i, π_i) and $\hat{\pi}_i = X_i/n_i$, $i = 1, 2$. The joint probability mass function is the product of the binomial mass functions for X_1 and X_2. This can be expressed in terms of $\theta = \pi_1 - \pi_2$ and a nuisance parameter such as π_1 or π_2 or $(\pi_1 + \pi_2)/2$; for example

$$\begin{aligned} f(x_1, x_2; n_1, n_2, \pi_1, \pi_2) &= \binom{n_1}{x_1}(\pi_1)^{x_1}(1-\pi_1)^{n_1-x_1}\binom{n_2}{x_2}\pi_2^{x_2}(1-\pi_2)^{n_2-x_2} \\ &= \binom{n_1}{x_1}(\theta+\pi_2)^{x_1}(1-\theta-\pi_2)^{n_1-x_1}\binom{n_2}{x_2}\pi_2^{x_2}(1-\pi_2)^{n_2-x_2} \\ &= f(x_1, x_2; n_1, n_2, \theta, \pi_2). \end{aligned}$$

For binary data, the conditional approach for eliminating the nuisance parameter π_2 applies only with the logit of the probability, so it applies for the odds ratio or its log rather than the difference of proportions.

One way to eliminate π_2 uses the unconditional product mass function to obtain a p-value as if π_2 were known and then maximizes this p-value over the possible values of π_2. With a statistic T such that large t_o contradicts H_0, the p-value for $H_0: \theta = \theta_0$ is

$$p(\theta_0) = \sup_{\pi_2} p[T \geq t_0; \theta_0, \pi_2],$$

where the *sup* is taken over the permissible π_2 for the fixed θ_0. Santner and Snell[21] proposed an unconditional approach by inverting two one-sided tests using $T = \hat{\pi}_1 - \hat{\pi}_2$.

For interval estimation of $\pi_1 - \pi_2$, Santner and Snell[21] actually stated a preference for the Sterne[18] approach, noting that it usually gives shorter intervals. However, that approach was then computationally infeasible except for very small $\{n_i\}$. Chan and Zhang[22] showed that conservativeness of the Santner and Snell tail method was exacerbated by the severe discreteness of $T = \hat{\pi}_1 - \hat{\pi}_2$ for small samples. For that application of the tail method, each sample with the same value of $\hat{\pi}_1 - \hat{\pi}_2$ has the same interval (for the given sample sizes). As discussed in section 2.2, improved performance results from inverting a test with a less discrete statistic. Chan and Zhang used the score

statistic[23,24] but with its exact distribution. For testing $H_0: \pi_1 - \pi_2 = \theta_0$, the version of that statistic with large-sample standard normal null distribution is

$$T = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \theta_0}{\sqrt{\tilde{\pi}_1(1-\tilde{\pi}_1)/n_1 + \tilde{\pi}_2(1-\tilde{\pi}_2)/n_2}}, \tag{5}$$

where $\tilde{\pi}_1$ and $\tilde{\pi}_2$ denote the ML estimates of π_1 and π_2 subject to $\pi_1 - \tau_2 = \theta_0$. Chan and Zhang[22] used the tail method with this statistic. Better performance yet tends to result from inverting the score test as a single two-sided test, in which the p-value compares the chi-squared form T^2 of the score statistic to $t^2_{0.25}$ *StatXact*, as of Version 5, provides these intervals and the Santner–Snell interval.

Table 2 shows some intervals for the Santner and Snell tail method, the Chan and Zhang tail method (i.e., inverting two one-sided score tests) and for the Agresti and Min two-sided adaptation, for various (x_1, x_2) values with $n_1 = n_2 = 10$. Figure 3 illustrates performance, plotting the coverage probability for these three methods as a function of π_1. The first panel in Figure 3 holds $\pi_2 = 0.3$ fixed and the second panel holds $\pi_1 - \pi_2 = 0.2$ fixed. Greater differences in coverage probability curves can occur with unbalanced sample sizes.

Interestingly, Coe and Tamhane[26] and Santner and Yamagami[27] also dealt with interval estimation of $\pi_1 - \pi_2$ with a generalized Sterne-type approach, but have not received much attention in the subsequent literature or in statistical practice. These methods also provide intervals with better coverage properties than the Santner and Snell[21] or Chan and Zhang[22] tail-method intervals. These two articles used different adaptations of the Sterne method in constructing the acceptance regions. The result is that the Coe and Tamhane intervals tend to be shorter for small to moderate $|\hat{\pi}_1 - \hat{\pi}_2|$ whereas the Santner and Yamagami intervals tend to be shorter for large $|\hat{\pi}_1 - \hat{\pi}_2|$. Coe[28] provided a SAS macro for the Coe and Tamhane approach.

As in the single-sample case, when the guarantee of maintaining at least the desired coverage probability is not crucial, some 'exact' methods can be so conservative as to be

Table 2 Nominal 95% confidence intervals for difference of proportions with binomial outcomes x_1 and x_2 in $n_1 = n_2 = 10$ independent trials

		Santner–Snell interval		Chan–Zhang score interval		Agresti–Min score interval		Agresti–Caffo adj. Wald interval	
x_1	x_2	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
5	0	0.014	0.829	0.118	0.813	0.132	0.778	0.093	0.740
5	1	−0.089	0.764	−0.020	0.741	−0.001	0.700	−0.020	0.686
5	2	−0.188	0.695	−0.146	0.671	−0.142	0.646	−0.124	0.624
5	3	−0.282	0.620	−0.260	0.601	−0.249	0.560	−0.222	0.556
5	4	−0.373	0.542	−0.369	0.539	−0.349	0.507	−0.314	0.481
5	5	−0.459	0.459	−0.456	0.456	−0.419	0.419	−0.400	0.400
2	0	−0.272	0.620	−0.129	0.556	−0.132	0.525	−0.124	0.457
2	1	−0.372	0.542	−0.280	0.464	−0.265	0.441	−0.240	0.407
2	2	−0.459	0.459	−0.386	0.386	−0.377	0.377	−0.346	0.346
2	3	−0.542	0.373	−0.490	0.309	−0.455	0.296	−0.446	0.279
2	4	−0.620	0.282	−0.585	0.229	−0.551	0.224	−0.538	0.205

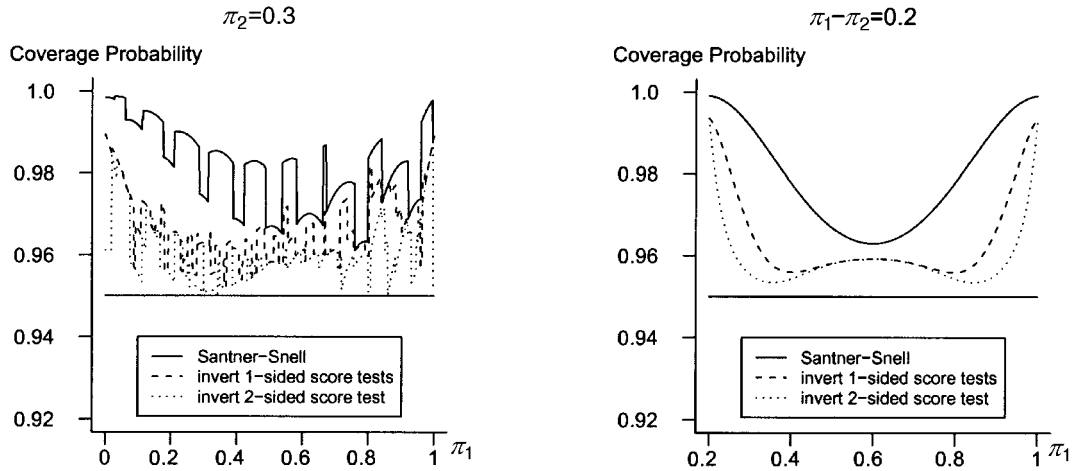

Figure 3 Coverage probabilities of 95% confidence intervals for $\pi_1 - \pi_2$ based on independent binomials with $n_1 = n_2 = 10$

less useful than an approximate large-sample method. However, the most popular large-sample 95% confidence interval,

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm 2\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}},$$

which inverts the Wald test, behaves poorly with small samples. It tends to have coverage probabilities much below the nominal values, especially when both π_i are near 0 or near 1. Agresti and Caffo[29] showed that the simple adaptation of adding two observations to each sample, one of each type, before computing the Wald interval improves it dramatically. Table 2 also shows this adjusted Wald interval, and Figure 4 compares its coverage probabilities when $n_1 = n_2 = 10$ to those for the Santner and Snell method and the ordinary Wald interval. Among methods that are more computationally intensive, inverting the large-sample score test by treating (5) as standard normal also works quite well. (See Nurminen[30] for its implementation.)

5 Confidence intervals for the odds ratio in 2 × 2 tables

Next we consider confidence intervals for the odds ratio θ in a 2 × 2 contingency table. Here, the standard 'exact' approach is the conditional one. Assuming a multinomial distribution for the cell counts $\{n_{ij}\}$, or assuming $\{n_{ij}\}$ are independent Poisson, or assuming the rows or the columns are independent binomials, conditioning on row and column marginal totals yields a distribution depending only on θ. For testing $H_0: \theta = 1$, this distribution is the hypergeometric. Constructing a confidence interval requires

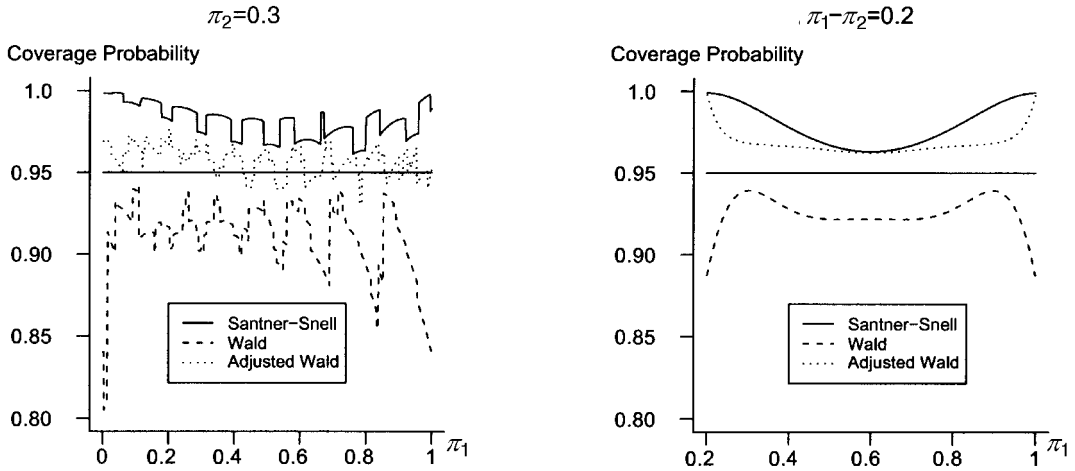

Figure 4 Coverage probabilities of 95% confidence intervals for $\pi_1 - \pi_2$ based on independent binomials with $n_1 = n_2 = 10$

inverting the family of tests for various non-null values θ_0. This leads to a noncentral version of the hypergeometric distribution,

$$P(n_{11} = t | \{n_{i+}\}, \{n_{+j}\}; \theta) = \frac{\binom{n_{1+}}{t}\binom{n - n_{1+}}{n_{+1} - t}\theta^t}{\sum_s \binom{n_{1+}}{s}\binom{n - n_{1+}}{n_{+1} - s}\theta^s}.$$

For 'exact' interval estimation of this parameter, Cornfield[31] suggested the tail method (1). This is the most common 'exact' method in practice, and it is the only option in *StatXact*.

In forming a confidence interval for θ, Baptista and Pike[32] adapted the Sterne[18] approach of inverting a single two-sided test with acceptance region based on ordered null probabilities. Alternatively, one could invert a two-sided test using a standard test statistic. The score statistic for testing $H_0: \theta = \theta_0$ with two independent binomials[24] is proportional to

$$T = n_1(\hat{\pi}_1 - \hat{\pi}_1)^2\left[\frac{1}{n_1\tilde{\pi}_1(1 - \tilde{\pi}_1)} + \frac{1}{n_2\tilde{n}_2(1 - \tilde{\pi}_2)}\right],$$

where $\tilde{\pi}_1$ and $\tilde{\pi}_1$ are the ML estimates of π_1 and π_2 subject to $\theta = \theta_0$. Agresti and Min[25] inverted 'exact' conditional tests using this statistic. The left side of Table 3 shows the Cornfield and Agresti–Min intervals when $n = 20$ and each marginal count is 10. Figure 5 plots coverage probabilities for $\log(\theta)$ for the two approaches, conditional on these margins. Inverting a single two-sided test gives better results. Similar results occur by inverting the test using the exact conditional distribution but with Blaker's[7] p-value.

Table 3 Nominal 95% confidence intervals for odds ratio with count n_{11} when each row and column marginal total is 10

	Cornfield conditional interval		Invert 2-sided conditional test		Invert 2-sided unconditional test		Mid-p adapted Cornfield	
n_{11}	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
0	0.000	0.09	0.000	0.07	0.000	0.05	0.000	0.06
1	0.0003	0.31	0.0005	0.30	0.0007	0.23	0.0005	0.24
2	0.004	0.76	0.006	0.68	0.006	0.56	0.006	0.60
3	0.018	1.68	0.025	1.48	0.018	1.29	0.024	1.34
4	0.052	3.60	0.069	3.38	0.052	2.81	0.068	2.87
5	0.126	7.94	0.158	6.35	0.130	7.70	0.160	6.25

Note: For count $6 \leq n_{11} \leq 10$, limits equal $(1/\theta_U, 1/\theta_L)$ for limits given for $10 - n_{11}$.

Another way to construct intervals that are shorter than with Cornfield's 'exact' method is to invert a test using the mid-p-value. Table 3 shows the resulting adaptation of the Cornfield intervals. They also tend to be a bit shorter than those obtained by inverting the single two-sided 'exact' conditional score test. However, they do not have the guarantee that the coverage probability is at least the nominal level.

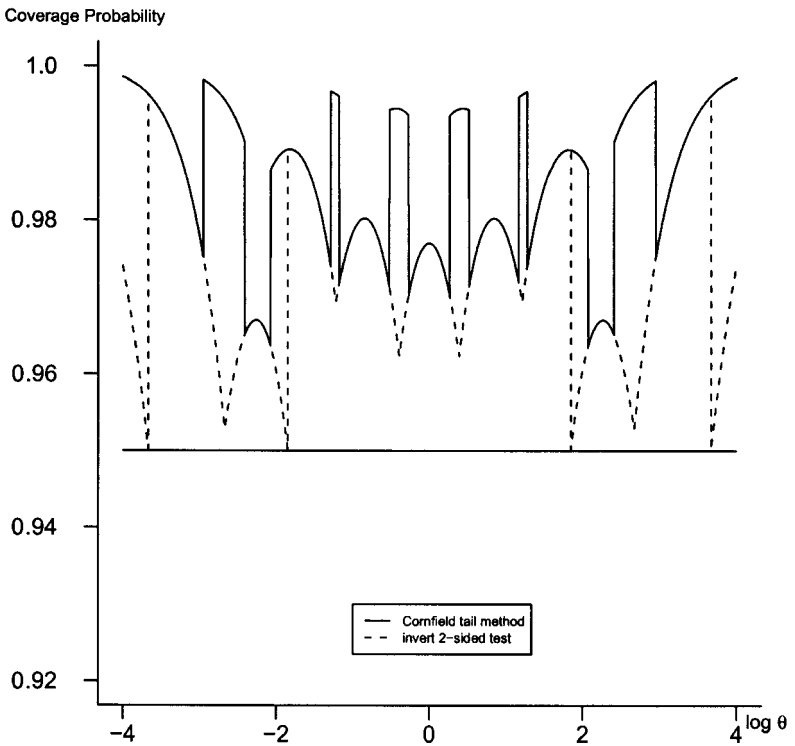

Figure 5 Coverage probabilities for 95% confidence intervals for the log odds ratio, with $n_1 = n_2 = 10$ and outcome margins $= 10$

As section 2.4 mentioned, the conditioning argument used in the exact conditional approach exacerbates the discreteness. This can cause severe conservativeness problems. Perhaps surprisingly, except for Troendle and Frank[33] and Agresti and Min,[34] the unconditional approach described in the previous section for $\pi_1 - \pi_2$ does not seem to have been used for the odds ratio. This approach is possible when the contingency table arises from two independent binomial samples, in which case $\theta = [\pi_1/(1-\pi_1)]/[\pi_2/(1-\pi_2)]$. It also applies for a single multinomial sample over the four cells, after conditioning on the row totals. Because the total number of outcomes of the two types (i.e., the two column totals) is not fixed, the relevant product binomial distribution is much less discrete. This gives the potential to reduce conservatism because of this, yet there is also the potential of increasing conservatism by forming the p-value using the worst-case scenario for the nuisance parameter.

We used the unconditional approach with the score test statistic to construct confidence intervals for the odds ratio. Table 3 shows some examples. For the case $n_1 = n_2 = 10$, Figure 6 compares coverage probabilities for the Cornfield 'exact' conditional interval, the conditional interval based on inverting an 'exact' two-sided score test, and the 'exact' unconditional interval using the score statistic. Here, we generated the two binomial samples without any restriction on the response margins. Plots are shown as a function of π_1 when π_2 is fixed at 0.3 and when θ is fixed at 2.0. See Agresti and Min[34] for further details.

Again, for some purposes it is better to use a good large-sample method than an overly conservative 'exact' one such as Cornfield's. The delta method yields the simple large-sample 95% interval for the log odds ratio,

$$\log(\hat{\theta}) \pm 2\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}, \tag{6}$$

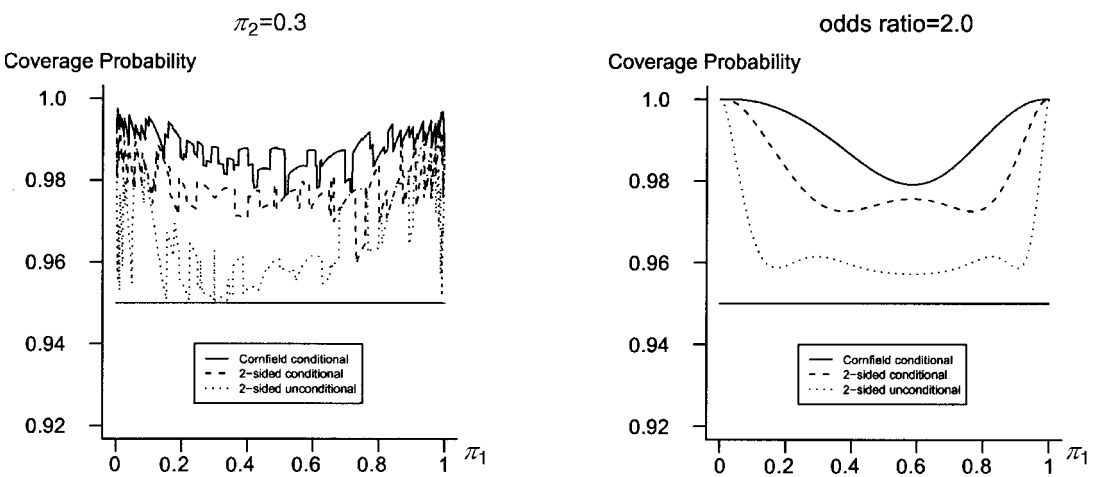

Figure 6 Coverage probabilities for 95% confidence intervals for odds ratio, when $n_1 = n_2 = 10$

which works quite well, usually being somewhat conservative. When any $n_{ij} = 0$, it is possible to improve the delta method formula by using the sample $\hat{\theta}$ value (0 or ∞) as one end point but adding a constant to the cells in using (6) to obtain the other endpoint.[35]

In closing this section, we mention that considerable debate has occurred over the years about the conditional versus unconditional approach to testing whether $\theta = 1$. See Sprott[36] (Section 6.4.4) for a recent cogent support of arguments originally voiced by Fisher against the unconditional approach. The same arguments apply to interval estimation.

6 Confidence intervals for other parameters

Similar results occur for other parameters of interest in discrete data problems. For instance, the discussion of section 4 on the difference between two binomial parameters applies also to their ratio, the relative risk $\theta = \pi_1/\pi_2$. Again, an unconditional approach eliminates the nuisance parameter in the test to be inverted. We illustrate by inverting 'exact' tests using the score statistic that is used for large-sample inference.[24,37] The score test statistic for $H_0: \theta = \theta_0$ is

$$T = \frac{n_1(\hat{\pi}_1 - \tilde{\pi}_1)^2}{\tilde{\pi}_1(1 - \tilde{\pi}_2)} + \frac{n_2(\hat{\pi}_2 - \tilde{\pi}_2)^2}{\tilde{\pi}_2(1 - \tilde{\pi}_2)},$$

where $\tilde{\pi}_1$ and $\tilde{\pi}_1$ are the ML estimates of π_1 and π_2 subject to $\pi_1/\pi_2 = \theta_0$. (For $\theta_0 = 1$, this and the score statistics for the odds ratio and the difference of proportions all simplify to the ordinary Pearson chi-squared statistic.) Figure 7 compares coverage

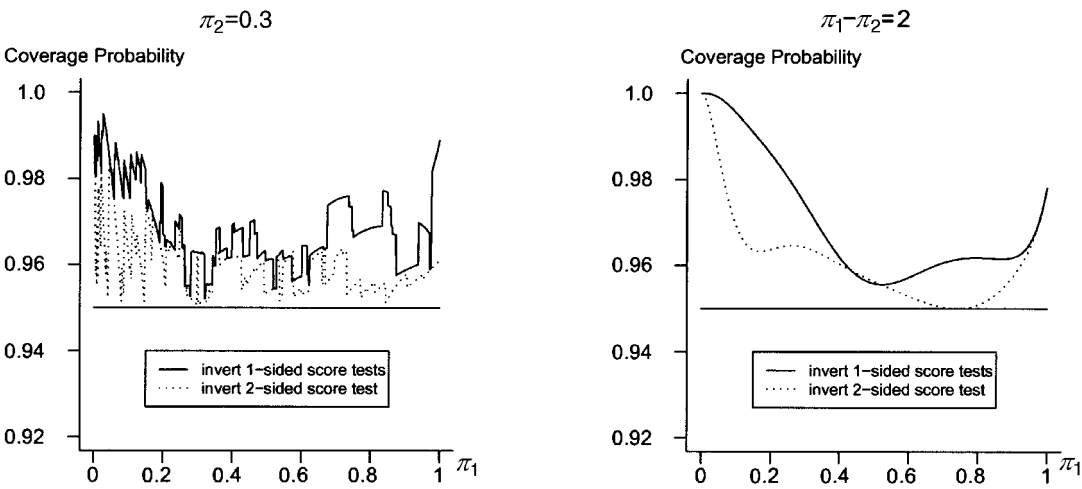

Figure 7 Coverage probabilities of 95% confidence intervals for $\pi_1 - \pi_2$ based on independent binomials with $n_1 = n_2 = 10$

probabilities of 95% confidence intervals based on the tail method and based on inverting the single two-sided score test using T with its exact distribution, when $n_1 = n_2 = 10$. One panel refers to $\pi_2 = 0.3$ and the other to $\theta = 2$. Large-sample approximate intervals based on inverting the chi-squared test using T also have good performance.[30,38]

A case related to the previous section is construction of a confidence interval for an odds ratio that is assumed constant in a set of 2×2 tables. Cox[39] (p. 48) and Gart[40] described the tail interval of form (1). For computing and software, see Mehta *et al.*,[41] Vollset *et al.*,[42] and *StatXact*. For examples of the advantage of instead inverting a single two-sided test, see Kim and Agresti,[11] who used a Sterne-type approach. When many points in the sample space can have the same value of the test statistic, they showed how one can reduce the conservativeness further by using the null probability of the observed table to form a finer partitioning within fixed values of the test statistic (Section 2.3). For instance, to illustrate the tail method, Gart[40] gave a 95% confidence interval of (0.05, 1.16) for a $2 \times 2 \times 18$ table. Inverting the two-sided test, the Kim and Agresti interval yields (0.06, 1.14), and it reduces further to (0.09, 0.99) with a more finely partitioned p-value.

A class of parameters that includes the odds ratio is the set of parameters for logistic regression models. Cox[39] (p. 48) suggested the tail method, using the conditional distribution to eliminate other parameters. Inverting a two-sided test using that distribution would tend to give shorter intervals. An open question is whether an unconditional approach may provide further improvement in some cases, because of a reduction in discreteness. This is the case for small samples with the odds ratio for a single 2×2 table. However, it is unclear how the conservativeness may increase by taking the supremum for the p-value using several tables, and the procedure would be highly computationally intensive. Implementing the Berger and Boos[43] approach of maximizing over a confidence interval for the nuisance parameters and adjusting the p-value appropriately may be helpful.

7 Summary: Recommendations on dealing with discreteness

In summary, discreteness has the effect of making 'exact' confidence intervals more conservative than desired, We make the following recommendations for reducing the effects of that discreteness. First, as emphasized throughout this article, invert two-sided tests rather than two one-sided tests (the tail method). Secondly, in that test use a test statistic that alleviates the discreteness (e.g., for comparing two proportions, use the score statistic rather than $\hat{\pi}_1 - \hat{\pi}_2$). Thirdly, when appropriate use an unconditional rather than conditional method of eliminating nuisance parameters.

This article has primarily discussed confidence interval methods that attain at least the nominal confidence level. More generally, for three types of situations in which a statistician might select a method, we believe the preferred method differs. One situation is that dealt with in this article, in which one needs to guarantee a lower bound on the coverage probability. A second situation, more important for most statistical practice, is when one wants the actual coverage probability to be close to the nominal level but not necessarily to have it as a lower bound. A third situation is that of

teaching basic statistical methods in a classroom or consulting environment, for which one may be willing to sacrifice quality of performance somewhat in favor of greater simplicity.

For most statistical practice (i.e., situation two), for interval estimation of a proportion or a difference or ratio of proportions, the inversion of the asymptotic score test seems to be a good choice.[14,38,44] This tends to have actual level fluctuating around the nominal level. If one prefers that level to be a bit more conservative, mid-p adaptations of 'exact' methods work well. For situations that require a bound on the error (i.e., situation one), it appears that basing conservative intervals on inverting the 'exact' score test has reasonable performance. For teaching (i.e., situation three), the Wald-type interval of point estimate plus and minus a normal-score multiple of a standard error is simplest. Unfortunately, this can perform poorly, but simple adjustments sometimes provide much improved performance.

Acknowledgements

This research was partially supported by grants from NIH and NSF. Thanks to Yongyi Min for constructing the figures and obtaining many of the tabulated values.

References

1 Mehta CR. The exact analysis of contingency tables in medical research. *Statistical Methods in Medical Research* 1994; **3**: 135–56.
2 Agresti A. Exact inference for categorical data: Recent advances and continuing controversies. *Statistics in Medicine* 2001; **20**: 2709–22.
3 *StatXact 5 for Windows*. Cambridge, MA: Cytel Software, 2001.
4 Neyman J. On the problem of confidence limits. *Annals of Mathemtical Statistics* 1935; **6**: 111–6.
5 Stevens WL. Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* 1950; **37**: 117–29.
6 Casella G, Berger RL. *Statistical Inference*, 2nd ed. Pacific Grove, CA: Wadsworth, 2001.
7 Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics* 2000; **28**: 783–98.
8 Crow EL. Confidence intervals for a proportion. *Biometrika* 1956; **43**: 423–35.
9 Santner TJ, Duffy DE. *The Statistical Analysis of Discrete Data.* Berlin: Springer-Verlag, 1989.
10 Cohen A, Sackrowitz HB. An evaluation of some tests of trend in contingency tables. *Journal of the American Statistical Association* 1992; **87**: 470–75.
11 Kim D, Agresti A. Improved exact inference about conditional association in three-way contingency tables. *Journal of the American Statistical Association* 1995; **90**: 632–39.
12 Suissa S, Shuster JJ. Exact unconditional sample sizes for the 2 by 2 binomial trial. *Journal of the Royal Statistical Society* 1985; **A148**: 317–27.
13 Mehta CR, Walsh SJ. Comparison of exact, mid-p, and Mantel–Haenszel confidence intervals for the common odds ratio across several 2×2 contingency tables. *The American Statistician* 1992; **46**: 146–50.
14 Newcombe R. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998; **17**: 857–72.
15 Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26**: 404–13.
16 Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician* 1998; **52**: 119–26.
17 Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science* 2001; **16**: 101–17.
18 Sterne TE. Some remarks on confidence or fiducial limits. *Biometrika* 1954; **41**: 275–78.

19 Blyth CR, Still HA. Binomial confidence intervals. *Journal of the American Statistical Association* 1983; **78**: 108–16.

20 Casella G. Refining binomial confidence intervals. *Canadian Journal of Statistics* 1986; **14**: 113–29.

21 Santer TJ, Snell MK. Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables. *Journal of the American Statistical Association* 1980; **75**: 386–94.

22 Chan ISF, Zhang Z. Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* 1999; **55**: 1202–9.

23 Mee RW. Confidence bounds for the difference between two probabilities (letter). *Biometrics* 1984; **40**: 1175–76.

24 Miettinen O, Nurminen M. Comparative analysis of two rates. *Statistics in Medicine* 1985; **4**: 213–26.

25 Agresti A, Min Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* 2001; **57**: 963–71.

26 Coe PR, Tamhane AC. Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. *Communications in Statistics, Part B—Simulation and Computation* 1993; **22**: 925–38.

27 Santner TJ, Yamagami S. Invariant small sample confidence-intervals for the difference of 2 success probabilities. *Communications in Statistics, Part B—Simulation and Computation* 1993; **22**: 33–59.

28 Coe PR. A SAS macro to calculate exact confidence intervals for the difference of two proportions. *Proceedings of the Twenty-Third Annual SAS Users Group International Conference*, 1998; pp. 1400–1405.

29 Agresti A, Caffo B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician* 2000; **54**: 280–88.

30 Nurminen M. Confidence intervals for the ratio and difference of two binomial proportions. *Biometrics* 1986; **42**: 675–76.

31 Cornfield J. A statistical problem arising from retrospective studies. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, J Neyman ed. 1956; **4**: 135–48.

32 Baptista J, Pike MC. Exact two-sided confidence limits for the odds ratio in a 2×2 table. *Journal of the Royal Statistical Society Series C* 1977; **26**: 214–20.

33 Troendle JF, Frank J. Unbiased confidence intervals for the odds ratio of two independent binomial samples with application to case–control data. *Biometrics* 2001; **57**: 484–89.

34 Agresti A, Min Y. Unconditional small-sample confidence intervals for the odds ratio. To appear in *Biostatistics*, 2002; **3**: 379–386.

35 Agresti A. On logit confidence intervals for the odds ratio with small samples. *Biometrics* 1999; **55**: 597–602.

36 Sprott DA. *Statistical inference in science*. New York: Springer, 2000.

37 Koopman PAR. Confidence intervals for the ratio of two binomial proportions. *Biometrics* 1984; **40**: 513–17.

38 Gart JJ, Nam J. Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* 1988; **44**: 323–38.

39 Cox DR. *Analysis of binary data*. London: Chapman and Hall, 1970.

40 Gart JJ. Point and interval estimation of the common odds ratio in the combination of 2×2 tables with fixed marginals. *Biometrika* 1970; **57**: 471–75.

41 Mehta CR, Patel NR, Gray R. Computing an exact confidence interval for the common odds ratio in several 2 by 2 contingency tables. *Journal of the American Statistical Association* 1985; **80**: 969–73.

42 Vollset SE, Hirji KF, Elashoff RM. Fast computation of exact confidence limits for the common odds ratio in a series of 2×2 tables. *Journal of the American Statistical Association* 1991; **86**: 404–9.

43 Berger RL, Boos DD. *P* values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 1994; **89**: 1012–16.

44 Newcombe R. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* 1998; **17**: 873–90.