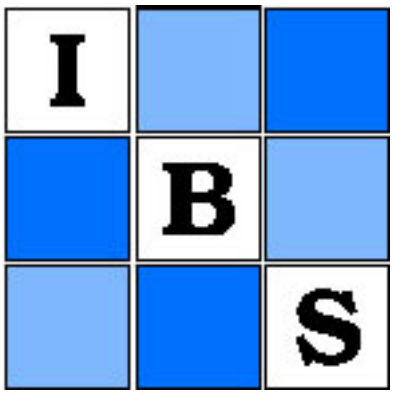

BIOMETRICS 57, 963–971
September 2001

CONSULTANT'S FORUM

On Small-Sample Confidence Intervals for Parameters in Discrete Distributions

Alan Agresti* **and Yongyi Min**

Department of Statistics, University of Florida,
Gainesville, Florida 32611-8545, U.S.A.
**email:* aa@stat.ufl.edu

SUMMARY. The traditional definition of a confidence interval requires the coverage probability at any value of the parameter to be at least the nominal confidence level. In constructing such intervals for parameters in discrete distributions, less conservative behavior results from inverting a single two-sided test than inverting two separate one-sided tests of half the nominal level each. We illustrate for a variety of discrete problems, including interval estimation of a binomial parameter, the difference and the ratio of two binomial parameters for independent samples, and the odds ratio.

KEY WORDS: Binomial distribution; Clopper–Pearson interval; Difference of proportions; Exact test; Odds ratio; Proportion; Relative risk; Tail method.

1. Introduction

Let T be a discrete statistic with probability mass function $f(t;\theta)$ and cumulative distribution function $F(t;\theta)$ indexed by a parameter θ. Some applications, especially in legal or regulatory environments, require interval estimators for θ to guarantee coverage probability of at least $1-\alpha$, for some fixed α, for all θ. Such methodology is also useful with small samples when one is unwilling to trust the uncertain performance of a large-sample approximation.

The usual approach inverts a family of tests having size at most α. For such a test, for each value θ_0 of θ, let $A(\theta_0)$ denote the acceptance region for testing H_0: $\theta = \theta_0$. Then for each value t of T, let $C(t) = \{\theta_0 : t \in A(\theta_0)\}$. This is a confidence region with the desired property. For a typical θ_0, $A(\theta_0)$ does not achieve probability of Type I error exactly equal to α because of discreteness. Hence, such confidence intervals are conservative. The actual coverage probability varies for different values of θ but exceeds $1-\alpha$ (Neyman, 1935) unless one artificially transforms T to a continuous variable using supplementary randomization (e.g., Anscombe, 1948; Stevens, 1950). These confidence intervals and the related significance tests are often referred to as exact because they use the true null distribution of T rather than an approximation based on large-sample normality. However, the actual coverage probability is not exact but only guaranteed to be bounded below by the nominal confidence level.

The approach to constructing such an interval commonly presented in theory of statistics texts inverts two separate one-sided tests each having size at most $\alpha/2$. For instance, if $F(t;\theta)$ is a decreasing function of θ for each t, the interval (θ_L, θ_U) is defined by the equations

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

where t_o is the observed value of T. This method 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. The latter behavior results from the distribution of $F(T;\theta)$ being stochastically larger than uniform when T is discrete (Casella and Berger, 1990, p. 421).

This article shows that, for constructing confidence intervals with discrete distributions, it is usually better to invert a single two-sided test than to invert two separate one-sided tests. Here, "better" means that intervals tend to be shorter and coverage probabilities tend to be closer to the nominal level. We first discuss potential disadvantages of the tail method. Using particular examples of coverage probability graphs, we then illustrate the advantages of basing a confidence interval on inversion of a two-sided test. We use such tests with two-sided P-values that order the sample space (1) by null probabilities or (2) by null tail probabilities or (3) by a criterion measuring distance from the null, such as a score statistic. We show examples with the binomial parameter, the difference and the ratio of two binomial parameters for independent samples, and the odds ratio.

2. Inverting Two One-Sided Tests Versus Inverting a Two-Sided Test

In constructing a confidence interval by inverting a test, forming acceptance regions such that

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

for all θ_0 guarantees that the confidence level is at least the nominal level. Inverting the 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 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 is then 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 an overall P-value defined as $P = 2 \times \min[P_{\theta_0}(T \geq t_o), P_{\theta_0}(T \leq t_o)]$ exceeds α (taking $P = 1.0$ if this exceeds 1.0).

A 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, e.g., the interval can never exclude θ by falling below it. To illustrate, suppose T has the geometric distribution $f(t;\theta) = (1-\theta)\theta^t, t = 0, 1, 2, \ldots$. Then $F(t;\theta) = 1 - \theta^{t+1}$, and using (1) yields the tail interval $((\alpha/2)^{1/t_o}, (1-\alpha/2)^{1/(t_o+1)})$. All θ between 0 and $1 - \alpha/2$ never fall above a confidence interval, and the coverage probability exceeds $1 - \alpha/2$ over this region.

To construct a confidence region using a single two-sided test, one approach enters 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 $1-\alpha$, i.e., $A(\theta_0)$ contains the smallest possible number of most likely outcomes (under $\theta = \theta_0$). This leads to optimality in terms of minimizing total length (Sterne, 1954; Crow, 1956). The intervals also satisfy a nestedness property, an interval with larger nominal confidence level necessarily containing one with a smaller nominal level. A slight complication is the lack of a unique way of forming $A(\theta_0)$ in many cases. In its crudest partitioning of the sample space, it corresponds to using the P-value

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

the sum of null probabilities of outcomes no more likely than the observed outcome. The confidence interval is the set of θ_0 for which

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

An endpoint θ_U (or θ_L) of this interval then satisfies

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

In a related approach, Blaker (2000) 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

$$P_{\theta_0}[\gamma(T,\theta_0) \leq \gamma(t_o,\theta_0)] > \alpha. \tag{4}$$

This corresponds to using a P-value that equals $\min[P_{\theta_0}(T \geq t_o), P_{\theta_0}(T \leq t_o)]$ plus an attainable probability in the other tail that is as close as possible to, but not greater than, that one-tailed probability. Blaker showed that such intervals also have the nestedness property but may not have length optimality. The P-value for this approach is necessarily no greater than the P-value for the tail method, which uses $2\min[P_{\theta_0}(T \geq t_o), P_{\theta_0}(T \leq t_o)]$. Thus, the intervals based on (4) have the important advantage of necessarily being contained in intervals obtained with the tail method.

A third two-sided approach orders the sample space according to the distance of t_0 from H_0. One forms P-values according to a statistic that describes this distance, such as a score or likelihood-ratio statistic. To reduce conservativeness, it is preferable to use a statistic that tends to be less discrete, as discussed in Sections 4 and 5.

These two-sided approaches usually provide sensible results. The confidence regions 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 two-sided tests is not guaranteed to be an interval because the endpoints of the acceptance region need not be monotone in θ_0. Casella and Berger (1990, p. 417) and Santner and Duffy (1989, p. 37) discussed this in the context of the binomial parameter. For the two-sided method (3) based on ordered null probabilities, an endpoint from inverting a two-sided test with nominal confidence level of $1 - \alpha$ can be identical to an endpoint from the tail method with nominal confidence level of $1-2\alpha$. In the geometric case, e.g., because of the monotone decrease in the probabilities, this method yields $[\alpha^{1/t_o}, 1]$. More generally, suppose θ_L is such that

$$P_{\theta_L}[f(T;\theta_L) \leq f(t_o;\theta_L)] = \sum_{t \geq t_o} f(t;\theta_L) = \alpha.$$

Then θ_L is the lower endpoint from two-sided approach (3) with nominal confidence level $1 - \alpha$ and the lower endpoint using one-sided approach (1) with nominal level $1 - 2\alpha$. This happens when $f(t;\theta_L)$ is monotone decreasing in t, the geometric distribution being an extreme example in which this occurs for all θ.

Unfortunately, no single method for constructing confidence regions with discrete distributions can have optimality simultaneously in length, necessarily yielding an interval, and nestedness (Blaker, 2000). For the cases discussed below, similar results occurred from inverting an exact test using method (3) with P-value based on ordered null probabilities, method (4) with P-value based on two-tail probabilities, or the method based on P-value for the score statistic. The latter two methods may yield slightly wider intervals than method (3) based on ordered null probabilities, but in our experience, they have fewer anomalous cases.

3. Confidence Intervals for a Binomial Parameter

Let T be 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, the Clopper–Pearson interval (Clopper and Pearson, 1934). The endpoints satisfy

$$\sum_{k=t_o}^{n} \binom{n}{k} \pi_L^k (1-\pi_L)^{n-k} = \alpha/2$$

and

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

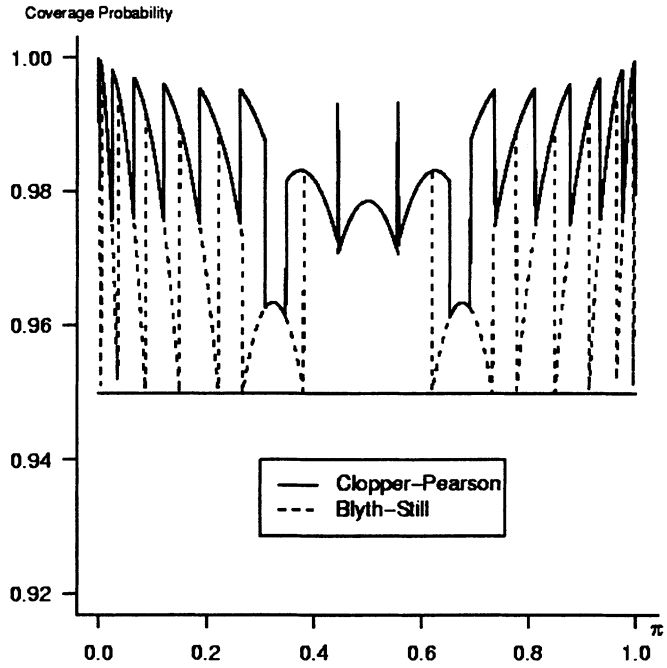

Figure 1. Coverage probabilities for 95% confidence intervals for a binomial parameter π with $n = 10$.

except that $\pi_L = 0$ when $t_o = 0$ and $\pi_U = 1$ when $t_o = n$. When $t_o = 0$, this confidence interval is $[0, 1-(\alpha/2)^{1/n}]$. 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 $n \leq \log(\alpha/2)/\log(0.5)$, e.g., $n \leq 5$ for $\alpha = .05$.

Sterne (1954) proposed method (3) of inverting a single test with outcomes ordered by their null probabilities. Blyth and Still (1983) and Casella (1986) amended this method slightly so that the confidence region cannot contain unconnected intervals and so natural symmetry and invariance properties are satisfied. The Blaker (2000) two-tailed probability approach (4) yields similar intervals that, unlike the Blyth–Still–Casella intervals, necessarily have the nestedness property with confidence levels, are contained within the Clopper–Pearson intervals, and are relatively simple to compute (Blaker's article contains short S-plus functions for doing this). The Blyth–Still interval is available in StatXact 4 (Cytel, 1999), the software having greatest scope for small-sample inference in discrete problems.

Figure 1 illustrates the superiority of forming the confidence interval by inverting a single two-sided test. It shows the actual coverage probabilities of the Clopper–Pearson and Blyth–Still intervals for nominal 95% confidence intervals, plotted as a function of π, when $n = 10$. Table 1 shows the 11 confidence intervals for each method. When $n = 10$, the ratio of expected lengths of the Blyth–Still and Clopper–Pearson intervals varies between 0.865 and .954, with a mean of 0.935. For comparison, Table 1 also shows the intervals using Blaker's (2000) two-sided approach (4) and by inverting the exact test using the score statistic. These are similar to the Blyth–Still intervals.

4. Confidence Intervals for Odds Ratio

When there are nuisance parameters, construction of an interval is more complicated. For exact inference with contingency tables, historically, the most popular approach is the conditional one that eliminates nuisance parameters by conditioning on their sufficient statistics.

For instance, consider inference for the odds ratio θ in a 2×2 contingency table. 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 the hypergeometric distribution

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

For this problem, Cornfield (1956) suggested the tail approach (1). This is the most commonly used exact method in practice (in fact, it is the only option in StatXact).

In forming a confidence interval for θ, Baptista and Pike (1977) adapted the approach (3) of inverting a single test based on ordered null probabilities. Table 2 shows this and the tail interval when $n = 20$ and each marginal count is 10. Figure 2 plots coverage probabilities for $\log(\theta)$ for the two approaches. Again, inverting a single test gives much better

Table 1

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

	Clopper–Pearson interval		Blyth–Still interval		Blaker interval		Score-test interval	
t	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
0	0.000	0.308	0.000	0.267	0.000	0.283	0.000	0.300
1	0.002	0.445	0.005	0.444	0.005	0.444	0.005	0.450
2	0.025	0.556	0.037	0.556	0.037	0.556	0.037	0.550
3	0.067	0.652	0.087	0.619	0.087	0.619	0.087	0.619
4	0.122	0.738	0.150	0.733	0.150	0.717	0.150	0.700
5	0.187	0.813	0.222	0.778	0.222	0.778	0.222	0.778

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

Table 2
Nominal 95% confidence intervals for odds ratio with count n_{11} when each row and column marginal total is 10

	Cornfield conditional interval		Invert two-sided conditional test		Invert two-sided conditional score test		Mid-P adapted Cornfield	
t	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
0	0.000	0.090	0.000	0.069	0.000	0.066	0.000	0.064
1	0.0003	0.309	0.0005	0.296	0.0005	0.290	0.0005	0.235
2	0.004	0.764	0.006	0.676	0.006	0.669	0.006	0.600
3	0.018	1.683	0.025	1.480	0.025	1.494	0.024	1.340
4	0.052	3.605	0.069	3.380	0.063	3.449	0.068	2.870
5	0.126	7.942	0.158	6.350	0.157	6.350	0.160	6.253

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

results. For $\log(\theta)$ between zero and four, we computed the expected lengths for the two methods, conditional on $0 < n_{11} < 10$ (with $n_{1+} = n_{+1} = 10$) so the interval width is finite. On the log scale, their ratio varies between 0.894 and 0.905, with a mean of 0.899. Similar results occur by inverting the test using the exact conditional distribution but with the score statistic (Cornfield, 1956) or inverting the test using Blaker's two-tailed P-value. Table 2 also shows the score-based intervals.

A related problem is constructing a confidence interval for an odds ratio that is assumed constant in a set of 2×2 tables. Gart (1970) described the tail interval of form (1). For computing and software, see Mehta, Patel, and Gray (1985), Vollset, Hirji, and Elashoff (1991), and StatXact. For examples of the advantage of instead inverting a single two-sided test, see Kim and Agresti (1995), who used approach (3) with ordered null probabilities. When possibly many points in the sample space have the same value of the test statistic, they showed one can reduce the conservativeness by using the null probability to form a finer partitioning within fixed values of the test statistic. For instance, to illustrate the tail method, Gart 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.

5. Confidence Intervals for Difference of Proportions

Next consider the difference of proportions for two independent binomial samples, where X_1 is bin(n_1, π_1), X_2 is bin(n_2, π_2), and $\hat{\pi}_i = X_i/n_i$. The joint probability mass function can be expressed in terms of $\theta = \pi_1 - \pi_2$ and a nuisance parameter such as π_1 or π_2 or $(\pi_1 + \pi_2)/2$, e.g.,

$$f(x_1, x_2; n_1, n_2, \theta, \pi_2) = \binom{n_1}{x_1}(\theta + \pi_2)^{x_1}(1 - \theta - \pi_2)^{n_1 - x_1} \times \binom{n_2}{x_2}\pi_2^{x_2}(1 - \pi_2)^{n_2 - x_2}.$$

The conditional approach for eliminating the nuisance parameter π_2 does not apply here since the canonical parameter is the difference of logits rather than the difference of proportions. One can eliminate π_2 using the unconditional approach of maximizing the P-value over its possible values. For instance, with a direction-sensitive statistic T for testing H_0: $\theta = \theta_0$, one-sided P-values for H_a: $\theta > \theta_0$ and H_a: $\theta < \theta_0$ are

$$P_U(\theta_0) = \sup_{\pi_2} P[T \geq t_0; \theta_0, \pi_2],$$
$$P_L(\theta_0) = \sup_{\pi_2} P[T \leq t_0; \theta_0, \pi_2], \quad (5)$$

where the supremum is taken over the permissible π_2 for the fixed θ_0.

For the tail method, the confidence interval satisfies $\theta_L = \sup\{\theta_0 : P_U(\theta_0) > \alpha/2\}$ and $\theta_U = \inf\{\theta_0 : P_L(\theta_0) > \alpha/2\}$. Santner and Snell (1980) proposed this interval using $T = \hat{\pi}_1 - \hat{\pi}_2$. Soms (1989a,b) discussed and programmed their interval and one inverting the Wald statistic. StatXact 4 provides the Santner and Snell interval. Santner and Snell (1980) also discussed the approach of inverting tests based on ordered null probabilities. They noted that it usually gave shorter in-

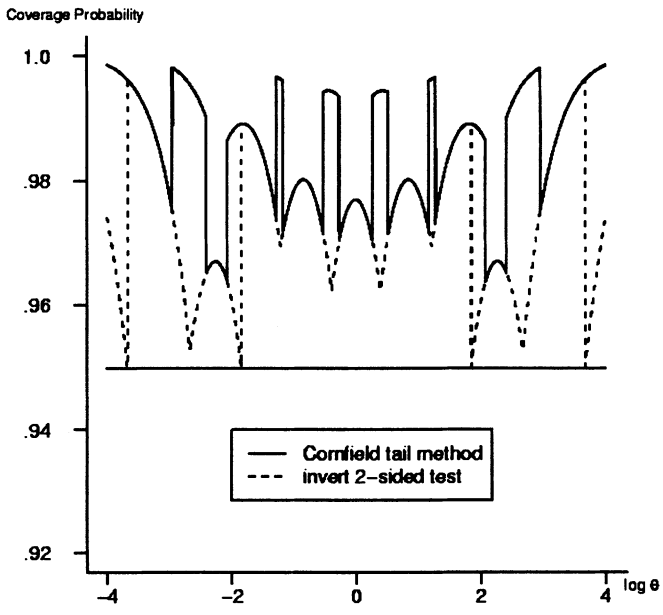

Figure 2. Coverage probabilities for 95% confidence intervals for $\pi_1 - \pi_2$ based on independent binomials, with $n_1 = n_2 = 10$.

Table 3
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 Interval		Invert two-sided Score test	
x_1	x_2	Lower	Upper	Lower	Upper	Lower	Upper
5	0	0.014	0.829	0.118	0.813	0.132	0.778
5	1	−0.089	0.764	−0.020	0.741	−0.001	0.700
5	2	−0.188	0.695	−0.146	0.671	−0.142	0.646
5	3	−0.283	0.620	−0.256	0.601	−0.249	0.560
5	4	−0.373	0.542	−0.369	0.539	−0.349	0.507
5	5	−0.459	0.459	−0.456	0.456	−0.419	0.419
2	0	−0.283	0.620	−0.129	0.556	−0.132	0.525
2	1	−0.373	0.542	−0.280	0.464	−0.265	0.441
2	2	−0.459	0.459	−0.386	0.386	−0.377	0.377
2	3	−0.542	0.373	−0.490	0.309	−0.455	0.296
2	4	−0.620	0.283	−0.585	0.229	−0.551	0.224

tervals and was preferable but was computationally infeasible (in 1980) except for very small n_i.

Chan and Zhang (1999) 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). Chan and Zhang showed that better performance results from using a less discrete statistic, such as the score statistic (Mee, 1984; Miettinen and Nurminen, 1985)

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

where $\tilde{\pi}_1$ and $\tilde{\pi}_2$ denote the maximum likelihood estimates of π_1 and π_2 subject to $\pi_1 - \pi_2 = \theta_0$. However, Chan and Zhang (1999) used only the tail method for this statistic. Better performance yet results from using the score statistic with a single two-sided test, in which the P-value compares $|T|$ to $|t_o|$.

Table 3 shows some intervals for the Santner–Snell and Chan–Zhang tail methods and for the two-sided inversion using the score statistic for various (x_1, x_2) values with $n_1 = n_2 = 10$. Similar improvements occurred from inverting a single likelihood-ratio test. In implementing the two-sided inversion, first we adapt the bisection method used in a FORTRAN program by Chan and Zhang (1999) to search for the lower and upper limits. Because of the theoretical possibility that disjoint intervals may exist, we supplement results from the bisection method with an intensive random search over the parameter space complement to that part represented by the bisection-based interval. This random search applies relatively greater weight to parameter points near those at the boundary obtained with the bisection method in order to enhance the chance of finding additionally needed points. In the rare cases

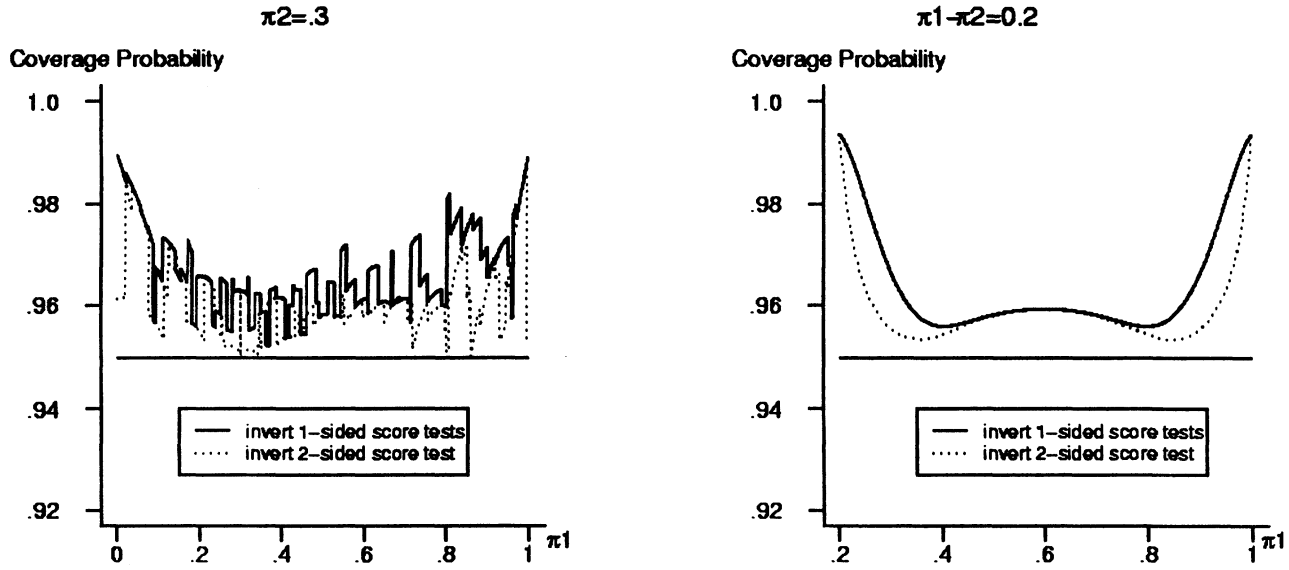

Figure 3. Coverage probabilities for 95% confidence intervals for $\pi_1 - \pi_2$ based on independent binomials, with $n_1 = n_2 = 10$.

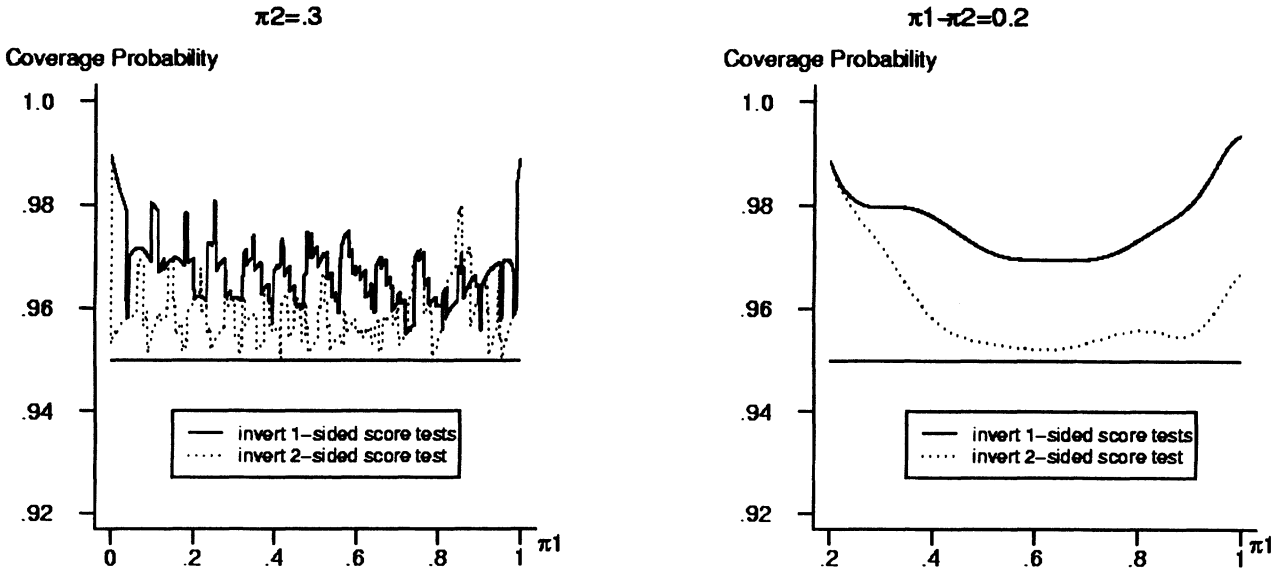

Figure 4. Coverage probabilities for 95% confidence intervals for $\pi_1 - \pi_2$ based on independent binomials, with $n_1 = 20$, $n_2 = 10$.

that this process determines the existence of disjoint intervals, we use the interval for $\pi_1 - \pi_2$ from the lowest lower bound to the highest upper bound. Testing of our algorithm showed accuracy to the fourth decimal place. As a check, in cases studied, we obtained confidence limits for all possible sample outcomes with the given sample sizes and evaluated and plotted coverage probability curves to analyze whether the nominal confidence level uniformly provided a lower bound; we did not observe any violations.

Figure 3 illustrates performance, plotting the coverage probability for the Chan–Zhang tail method and the two-sided score test approach 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. When $\pi_1 - \pi_2 = 0.2$, the ratio of expected length varies between 0.937 and 0.948, with a mean of 0.945, as π_1 varies between 0.2 and 1. Even greater differences in coverage probability curves can occur with unbalanced sample sizes. Figure 4 illustrates this, making the same comparisons but with $n_1 = 20$ and $n_2 = 10$. Both methods tend to be very conservative when both parameters are near zero or one.

An alternative way to invert unconditional tests uses the Berger and Boos (1994) method of eliminating the nuisance parameter. That method takes the supremum in (5) over a high confidence region (e.g., 99.9%) for the nuisance parameter and adjusts the P-value (e.g., by adding 0.001) so that the overall nominal size is not exceeded. StatXact has the option of adapting the Santner–Snell interval in this manner. We considered it for the two-sided score interval, but this did not provide improved performance over the score interval based on taking the supremum over the entire space for the nuisance parameter. Likewise, Chan and Zhang (1999) noted that using this sort of restricted search over values of the nuisance parameter did not improve performance of the interval based on the tail method with score tests.

Interestingly, Coe and Tamhane (1993) and Santner and Yamagami (1993) also dealt with the problem of interval estimation of $\pi_1 - \pi_2$ with a generalized version of approach (3) based on ordered null probabilities. Surprisingly, their methods 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 (1980) or Chan and Zhang (1999) tail-method intervals. The Coe–Tamhane and Santner–Yamagami methods used different approaches in constructing the acceptance regions. The result is that the Coe–Tamhane intervals tend to be shorter for small to moderate $|\hat{p}_1 - \hat{p}_2|$ whereas the Santner–Yamagami intervals tend to be shorter for large $|\hat{p}_1 - \hat{p}_2|$. Lee, Serachitopol, and Brown (1997) evaluated these and other intervals for $\pi_1 - \pi_2$ and showed that, over a broad range of cases, the Coe–Tamhane interval had the best performance. Coe (1998) provided a SAS macro for the Coe–Tamhane approach.

6. Confidence Intervals for Other Parameters

Similar results occur for other parameters in discrete data problems. For instance, the discussion of the previous section on the difference of two binomial parameters also applies to their ratio, the relative risk $\theta = \pi_1/\pi_2$. Again, an unconditional approach can eliminate the nuisance parameter in the test to be inverted. We illustrate by inverting tests using the score statistic (Koopman, 1984; Miettinen and Nurminen, 1985), which has good performance for large-sample confidence intervals (Gart and Nam, 1988). Figure 5 compares coverage probabilities of 95% confidence intervals based on the tail method and based on inverting a single two-sided score test when $n_1 = n_2 = 10$. One panel refers to $\pi_2 = 0.3$ and the other to $\theta = 2$. When $\theta = 2$, the ratio of expected widths on the log scale, conditional on $x_1 > 0$ and $x_2 > 0$, varies between 0.801 and 0.864, with a mean of 0.833, as π_1 varies between zero and one.

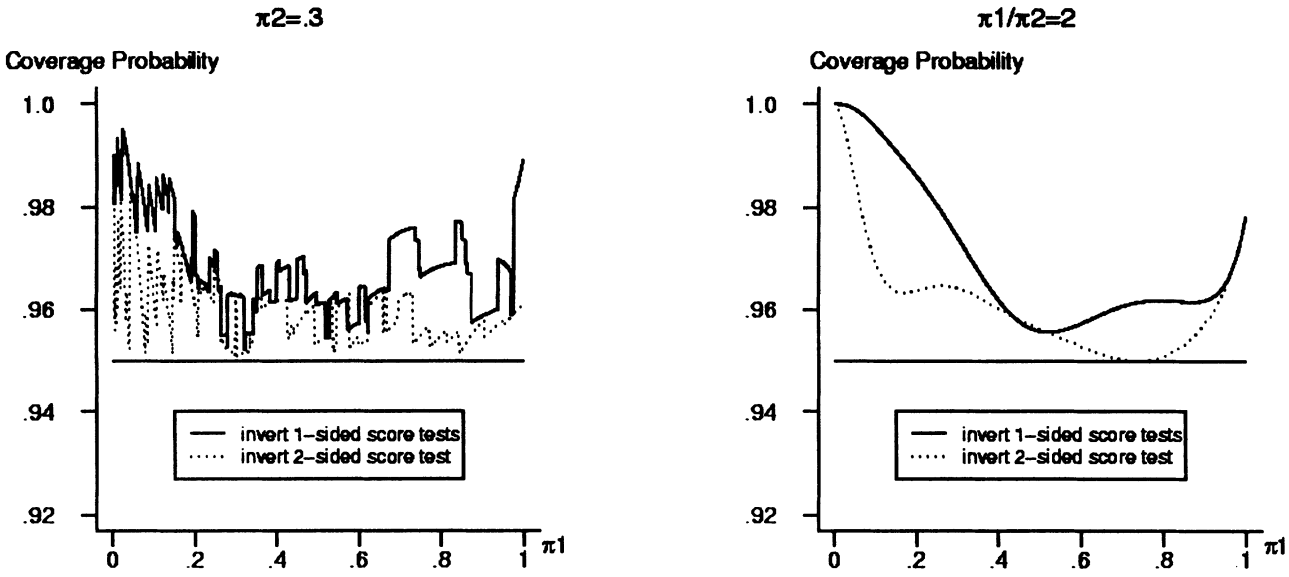

Figure 5. Coverage probabilities for 95% confidence intervals for π_1/π_2 based on independent binomials, with $n_1 = n_2 = 10$.

Another case analyzed with the tail method in most theory of statistics textbooks is the Poisson parameter. The tail method yields an interval specified with chi-squared percentage points (Garwood, 1936), but a plot in Casella and Berger (1990, p. 422) illustrates its conservativeness. Crow and Gardner (1959), Walton (1970), Casella and Robert (1989), Blaker (2000), and Kabaila and Byrne (2001) discussed improved intervals based on two-sided tests.

A class of parameters that includes the odds ratio is the set of parameters for logistic regression models. Cox (1970, p. 48) applied the tail method, using the conditional distribution to eliminate other parameters. Inverting a two-sided test has the benefits illustrated above for the odds ratio.

For independent binomial samples, we used the unconditional approach to obtain confidence intervals for the difference or ratio of proportions. In principle, one can also apply the unconditional approach to the odds ratio and more generally to logistic regression parameters even though the conditional approach is available. An open question is whether the unconditional two-sided approach may provide improvement over the conditional two-sided approach in some cases. The potential for improvement exists because of a reduction in discreteness. This is the case for testing equality of two independent binomials (e.g., Suissa and Shuster, 1985). However, there is also the potential for increased conservativism because of the approach of eliminating the nuisance parameter by taking a supremum of P-values (with respect to the nuisance parameter) instead of conditioning it out. Our preliminary study of this, for the odds ratio, suggests that the unconditional approach tends to provide intervals with coverage probabilities nearer the nominal levels. This will be addressed in detail in a future article.

7. A Limitation of the Two-Sided Approach

The past four sections showed examples of advantages of basing confidence intervals on two-sided tests. While the main theme of this article is recommending this over the tail method, a referee has suggested a qualification. In many studies, the goal is to show that a new treatment is better than a standard one. Such studies often use one-sided tests. Interval estimation is consistent with the test when the interval is based on inverting two one-sided tests (e.g., a 95% interval is then consistent with the result of the one-sided test with nominal size 0.025). Also, in some noninferiority trials, it is a regulatory requirement to use a confidence interval for the difference of proportions and compare one bound to a prespecified value. Again, using an interval based on inverting one-sided tests guarantees that the size of an implicit one-sided test does not exceed the nominal size.

A related comment is that users often are particularly interested in one of the bounds (say, the lower one) and interpret 95% intervals as imparting 97.5% confidence that the parameter falls above that bound. This inference is not appropriate with intervals based on inverting a two-sided test. Of course, in discrete cases, 97.5% is a lower bound and the actual confidence may be considerably higher than one prefers.

For such goals, one can argue in favor of simply calculating a one-sided confidence bound instead of a confidence interval. This may be a psychological barrier for many statisticians because most statistical texts discuss one-sided tests but few discuss one-sided confidence bounds.

8. Summary and Recommendations

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, apart from the caveat of the previous section, invert a two-sided test rather than two one-sided tests (the tail method). Second, in that test, use a test statistic that alleviates the discreteness (e.g., for comparing two proportions, use the score statistic rather than $\hat{\pi}_1 - \hat{\pi}_2$). Third, when appropriate, use an unconditional rather than conditional method of eliminating nuisance parameters.

A fourth recommendation is relevant if one is willing to relax slightly the requirement that the actual coverage probability have the nominal level as a lower bound at every parameter value. We then recommend inverting a two-sided exact test but using the mid P-value. The mid P-value has form $P(T > t_o) + (1/2)P(T = t_o)$, which is the expected value of the Stevens (1950) randomized P-value $P(T > t_o) + U \times P(T = t_o)$, where U is a uniform$(0, 1)$ random variable, which achieves the nominal size. The mid P-value has null expected value of 0.5, and the sum of the two one-sided mid P-values equals 1.0. Evaluations for a variety of problems (e.g., Mehta and Walsh, 1992; Newcombe, 1998a) have shown that, although this method no longer guarantees coverage probabilities of at least the nominal level, it still tends to be somewhat conservative, although 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 size 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. Table 2 shows the confidence interval for the odds ratio that StatXact reports using the mid P adjustment of the Cornfield interval, which is a conditional one. Comparing this to the ordinary Cornfield interval in this table illustrates the shortening of intervals that can occur with the mid P method.

Related to this last point, we emphasize in closing that, except for the last paragraph, this article has discussed only 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 of consulting environment, for which one may be willing to sacrifice quality of performance somewhat in favor of 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 a good choice (e.g., Gart and Nam, 1988; Newcombe, 1998a,b). This tends to have an 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), basing conservative intervals on inverting the exact score test or the test using Blaker's (2000) two-tailed P-value 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 result in much improved performance, e.g., see Agresti and Caffo (2000).

ACKNOWLEDGEMENTS

This research was partially supported by grants from NIH and NSF. The authors thank George Casella, Cyrus Mehta, Thomas Santner, Roger Berger, and a referee for helpful comments.

RÉSUMÉ

La définition traditionnelle d'un intervalle de confiance requiert de la probabilité de recouvrement de toute valeur du paramètre d'être au moins égale au niveau de confiance nominal. Pour des paramètres de distributions discrètes, on adopte un comportement moins conservatif en construisant de tels intervalles à partir d'une famille de tests bilatéraux plutôt qu'à partir de deux familles de tests unilatéraux séparés dont le niveau est la moitié du niveau nominal. Nous illustrons cela avec un certain nombre de problèmes discrets incluant l'estimation par intervalle d'un paramètre de binomiale, la différence et le rapport de deux paramètres de binomiales à partir d'échantillons indépendants, ainsi que le rapport de chances.

REFERENCES

Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures. *The American Statistician* **54,** 280–288.

Anscombe, F. J. (1948). The validity of comparative experiments (with discussion). *Journal of the Royal Statistical Society, Series A* **111,** 181–211.

Baptista, J. and Pike, M. C. (1977). Exact two-sided confidence limits for the odds ratio in a 2 × 2 table. *Journal of the Royal Statistical Society, Series C* **26,** 214–220.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89,** 1012–1016.

Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics* **28,** 783–798.

Blyth, C. R. and Still, H. A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association* **78,** 108–116.

Casella, G. (1986). Refining binomial confidence intervals. *Canadian Journal of Statistics* **14,** 113–129.

Casella, G. and Berger, R. (1990). *Statistical Inference.* Pacific Grove, California: Wadsworth.

Casella, G. and Robert, C. (1989). Refining Poisson confidence intervals. *Canadian Journal of Statistics* **17,** 45–57.

Chan, I. S. F. and Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* **55,** 1202–1209.

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26,** 404–413.

Coe, P. R. (1998). A SAS macro to calculate exact confidence intervals for the difference of two proportions. *Proceedings of the 23rd Annual SAS Users Group International Conference,* 1400–1405.

Coe, P. R. and Tamhane, A. C. (1993). Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. *Communications in Statistics—Simulation and Computation* **22,** 925–938.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,* **4,** 135–148.

Cox, D. R. (1970). *Analysis of Binary Data.* London: Chapman and Hall.

Crow, E. L. (1956). Confidence intervals for a proportion. *Biometrika* **43,** 423–435.

Crow, E. L. and Gardner, R. S. (1959). Confidence intervals for the expectation of a Poisson variable. *Biometrika* **46,** 441–447.

Cytel. (1999). *StatXact 4 for Windows.* Cytel Software: Cambridge, Massachusetts.

Gart, J. J. (1970). Point and interval estimation of the common odds ratio in the combination of 2×2 tables with fixed marginals. *Biometrika* **57,** 471–475.

Gart, J. J. and Nam, J. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* **44,** 323–338.

Garwood, F. (1936). Fiducial limits for the Poisson distribution. *Biometrika* **28,** 437–442.

Kabaila, P. and Byrne, J. (2001). Exact short Poisson confidence intervals. *Canadian Journal of Statistics* **29,** in press.

Kim, D. and Agresti, A. (1995). Improved exact inference about conditional association in three-way contingency tables. *Journal of the American Statistical Association* **90,** 632–639.

Koopman, P. A. R. (1984). Confidence intervals for the ratio of two binomial proportions. *Biometrics* **40,** 513–517.

Lee, J. J., Serachitopol, D. M., and Brown, B. W. (1997). Likelihood-weighted confidence intervals for the difference of two binomial proportions. *Biometrical Journal* **39,** 387–407.

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

Mehta, C. R. and Walsh, S. J. (1992). Comparison of exact, mid-p, and Mantel–Haenszel confidence intervals for the common odds ratio across several 2×2 contingency tables. *The American Statistician* **46,** 146–150.

Mehta, C. R., Patel, N. R., and Gray, R. (1985). Computing an exact confidence interval for the common odds ratio in several 2 by 2 contingency tables. *Journal of the American Statistical Association* **80,** 969–973.

Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine* **4,** 213–226.

Newcombe, R. (1998a). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine* **17,** 857–872.

Newcombe, R. (1998b). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine* **17,** 873–890.

Neyman, J. (1935). On the problem of confidence limits. *Annals of Mathematical Statistics* **6,** 111–116.

Santner, T. J. and Duffy, D. E. (1989). *The Statistical Analysis of Discrete Data.* Berlin: Springer-Verlag.

Santner, T. J. and Snell, M. K. (1980). 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* **75,** 386–394.

Santner, T. J. and Yamagami, S. (1993). Invariant small sample confidence-intervals for the difference of 2 success probabilities. *Communications in Statistics—Simulation and Computation* **22,** 33–59.

Soms, A. P. (1989a). Exact confidence intervals, based on the Z statistic, for the difference between two proportions. *Communications in Statistics—Simulation and Computation* **18,** 1325–1341.

Soms, A. P. (1989b). Some recent results for exact confidence intervals for the difference between two proportions. *Communications in Statistics—Simulation and Computation* **18,** 1343–1357.

Sterne, T. E. (1954). Some remarks on confidence or fiducial limits. *Biometrika* **41,** 275–278.

Stevens, W. L. (1950). Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* **37,** 117–129.

Suissa, S. and Shuster, J. J. (1985). Exact unconditional sample sizes for the 2 by 2 binomial trial. *Journal of the Royal Statistical Society, Series A* **148,** 317–327.

Vollset, S. E., Hirji, K. F., and Elashoff, R. M. (1991). Fast computation of exact confidence limits for the common odds ratio in a series of 2×2 tables. *Journal of the American Statistical Association* **86,** 404–409.

Walton, G. S. (1970). A note on nonrandomized Neyman-shortest unbiased confidence intervals for the binomial and Poisson parameters. *Biometrika* **57,** 223–224.

Received October 2000. *Revised January* 2001.
Accepted January 2001.