Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions

Alan Agresti; Brent A. Coull

The American Statistician, Vol. 52, No. 2. (May, 1998), pp. 119-126.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28199805%2952%3A2%3C119%3AAIBT%22F%3E2.0.CO%3B2-S

The American Statistician is currently published by American Statistical Association.

Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions

Alan AGRESTI and Brent A. COULL

For interval estimation of a proportion, coverage probabilities tend to be too large for "exact" confidence intervals based on inverting the binomial test and too small for the interval based on inverting the Wald large-sample normal test (i.e., sample proportion ± z-score × estimated standard error). Wilson's suggestion of inverting the related score test with null rather than estimated standard error yields coverage probabilities close to nominal confidence levels, even for very small sample sizes. The 95% score interval has similar behavior as the adjusted Wald interval obtained after adding two "successes" and two "failures" to the sample. In elementary courses, with the score and adjusted Wald methods it is unnecessary to provide students with awkward sample size guidelines.

KEY WORDS: Confidence interval; Discrete distribution; Exact inference; Poisson distribution; Small sample; Score test.

1. INTRODUCTION

One of the most basic analyses in statistical inference is forming a confidence interval for a binomial parameter p. Let X denote a binomial variate for sample size n, and let $\hat{p} = X/n$ denote the sample proportion. Most introductory statistics textbooks present the confidence interval based on the asymptotic normality of the sample proportion and estimating the standard error. This $100(1-\alpha)\%$ confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \tag{1}$$

where z_c denotes the $1-c$ quantile of the standard normal distribution. This is called the *Wald confidence interval* for p, since it results from inverting the Wald test for p; that is, the interval is the set of p_0 values having P value exceeding α in testing $H_0 : p = p_0$ against $H_a : p \neq p_0$ using the test statistic $z = (\hat{p} - p_0)/\sqrt{\hat{p}(1-\hat{p})/n}$. Historically, this is surely one of the first confidence intervals proposed for any parameter (see, e.g., Laplace 1812, p. 283).

To avoid approximation, most advanced statistics textbooks recommend the Clopper–Pearson (1934) "exact" confidence interval for p, based on inverting equal-tailed binomial tests of $H_0 : p = p_0$. It has endpoints that are the solutions in p_0 to the equations

$$\sum_{k=x}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \alpha/2$$

and

$$\sum_{k=0}^{x} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \alpha/2,$$

except that the lower bound is 0 when $x = 0$ and the upper bound is 1 when $x = n$. This interval estimator is guaranteed to have coverage probability of *at least* $1-\alpha$ for every possible value of p. When $x = 1, 2, \ldots, n-1$, the confidence interval equals

$$\left[1 + \frac{n-x+1}{xF_{2x,2(n-x+1),1-\alpha/2}}\right]^{-1} < p < \left[1 + \frac{n-x}{(x+1)F_{2(x+1),2(n-x),\alpha/2}}\right]^{-1},$$

where $F_{a,b,c}$ denotes the $1-c$ quantile from the F distribution with degrees of freedom a and b. Equivalently, the lower endpoint is the $\alpha/2$ quantile of a beta distribution with parameters x and $n-x+1$, and the upper endpoint is the $1-\alpha/2$ quantile of a beta distribution with parameters $x+1$ and $n-x$. Letters to the editor from J. Klotz and from L. Leemis and K. S. Trivedi in the November 1996 issue of this journal (p. 389) showed how simple it is to calculate this interval using Minitab or S-Plus.

A considerable literature exists about these and other, less common, methods of forming confidence intervals for p. Santner and Duffy (1989, pp. 33-43) and Vollset (1993) reviewed a variety of methods. It has been known for some time that the Wald interval performs poorly unless n is quite large (e.g., Ghosh 1979, Blyth and Still 1983). The Clopper–Pearson exact interval is typically treated as the "gold standard" (e.g., Böhning 1994; Leemis and Trivedi 1996; Jovanovic and Levy 1997; and most mathematical statistics texts). However, this procedure is necessarily conservative, because of the discreteness of the binomial distribution (Neyman 1935), just as the corresponding exact test (without supplementary randomization on the boundary of the critical region) is conservative. For any fixed parameter value, the actual coverage probability can be much larger than the nominal confidence level unless n is quite large, and we believe it is inappropriate to treat this approach as optimal for statistical practice.

A compromise solution is the confidence interval based on inverting the approximately normal test that uses the null, rather than estimated, standard error; that is, its

Alan Agresti is Professor, Department of Statistics, University of Florida, Gainesville, FL 32611-8545 (E-mail: aa@stat.ufl.edu). Brent A. Coull is a post-doc, Department of Biostatistics, Harvard School of Public Health, Boston MA 02115. This work was partially supported by a grant from the National Institutes of Health. The authors thank the referees and Thomas Santner for helpful suggestions.

endpoints are the p_0 solutions to the equations $(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n} = \pm z_{\alpha/2}$. This confidence interval, apparently first discussed by Edwin B. Wilson (1927), has the form

$$\left(\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{[\hat{p}(1-\hat{p}) + z^2_{\alpha/2}/4n]/n}\right)/(1 + z^2_{\alpha/2}/n). \tag{2}$$

This inversion of what is the score test for p is called the *score confidence interval.* (Score tests, and in particular their standard errors, are based on the log likelihood at the null hypothesis value of the parameter, whereas Wald tests are based on the log likelihood at the maximum likelihood estimate; see, e.g., Agresti 1996, pp. 88-95.) This article shows that the score confidence interval tends to perform much better than the exact or Wald intervals in terms of having coverage probabilities close to the nominal confidence level. It can be recommended for use with nearly all sample sizes and parameter values. In addition, we show that a simple adaptation of the Wald interval also performs well even for small samples.

At first glance, the score confidence interval formula seems awkward to interpret, compared to (1). Letting $z = z_{\alpha/2}$, however, the midpoint of this interval is the weighted average

$$\hat{p}\left(\frac{n}{n+z^2}\right) + \frac{1}{2}\left(\frac{z^2}{n+z^2}\right),$$

which falls between $\hat{p}$ and $1/2$, with the weight given to $\hat{p}$ approaching 1 asymptotically. This midpoint shrinks the sample proportion towards .5, the shrinking being less severe as n increases. The coefficient of z in the term that is added to and subtracted from this midpoint to form the score confidence interval has square equal to

$$\frac{1}{n+z^2}\left[\hat{p}(1-\hat{p})\left(\frac{n}{n+z^2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z^2}{n+z^2}\right)\right].$$

This has the form of a weighted average of the variance of a sample proportion when $p = \hat{p}$ and the variance of a sample proportion when $p = 1/2$, using $n + z^2$ in place of the usual sample size n.

2. COMPARING ACTUAL COVERAGE PROBABILITIES TO NOMINAL CONFIDENCE LEVELS

For a fixed value of a parameter, the actual coverage probability of an interval estimator is the (a priori) probability that the interval contains that value. In many cases, such as with discrete distributions, this varies according to the parameter value. In statistical theory, the confidence coefficient is defined to be the infimum of such coverage probabilities for all possible values of that parameter. Most practitioners, however, probably interpret confidence coefficients in terms of "average performance" rather than "worst possible performance." Thus, a possibly more relevant description of performance is the long-run percentage of times that the procedure is correct when it is used repeatedly for a variety of data sets in various problems with possibly different parameter values.

For any confidence interval procedure for estimating p, the actual coverage probability at a fixed value of p is

$$C_n(p) = \sum_{k=0}^{n} I(k,p)\binom{n}{k} p^k(1-p)^{n-k},$$

where $I(k,p)$ equals 1 if the interval contains p when $X = k$ and equals 0 if it does not contain p. We summarize this, using the alternative description of performance, by averaging over the possible values that p can take. We obtained results $\bar{C}_n = \int_0^1 C_n(p)g(p)dp$ for three beta densities $g(p)$ for this averaging: (1) the uniform distribution (mean = .50, std. dev. = $1/\sqrt{12}$ = .29); (2) bell-shaped with values relatively near the middle (mean = .50, std. dev. = .10); (3) skewed with values relatively near 0 (mean = .10, std. dev. = .05) or, by symmetry, near 1. Due to space considerations, we report results here mainly for the first case, but similar results occurred in the other two cases. Though this evaluation may suggest a Bayesian approach to inference, we restrict attention in this article to comparing the three standard methods decribed previously, in which the user makes no assumption about such a distribution for p.

Table 1 shows the mean of the actual coverage probabilities for the uniform averaging of the parameter values (i.e., $\bar{C}_n$ with $g(p) = 1$, $0 \le p \le 1$) at various sample sizes, for nominal 95% Wald, score, and exact confidence intervals (the three other methods listed in that table are discussed

Table 1. Mean Coverage Probabilities of Nominal 95% Confidence Intervals for the Binomial Parameter p, with Root Mean Square Errors in Parentheses, for Sampling p from a Uniform Distribution

Method	n = 5	n = 15	n = 30	n = 50	n = 100
Exact	.990	.980	.973	.969	.965
	(.041)	(.031)	(.026)	(.022)	(.017)
Score	.955	.953	.952	.952	.951
	(.029)	(.019)	(.014)	(.012)	(.008)
Wald	.641	.819	.875	.901	.922
	(.400)	(.238)	(.170)	(.133)	(.094)
Wald with *t*	.664	.837	.886	.905	.926
	(.391)	(.233)	(.167)	(.131)	(.093)
Mid-*P*	.978	.964	.958	.955	.953
	(.033)	(.021)	(.017)	(.013)	(.010)
Continuity-corrected Score	.987	.979	.973	.969	.965
	(.039)	(.030)	(.025)	(.021)	(.016)

in Section 4). The mean actual coverage probabilities for the Wald interval tend to be much too small. On the other hand, the exact interval is very conservative. For instance, for this method, $\bar{C}_n = .990$ when $n = 5$, .980 when $n = 15$, and .973 when $n = 30$. By contrast, $\bar{C}_n$ for the score method is close to the nominal confidence level, even for $n = 5$ where it is .955. Figure 1, which plots $\bar{C}_n$ as a function of n for the three interval estimators with the uniform and skewed beta weightings, illustrates their performance. Similar results were obtained with the bell-shaped weighting and using .90 nominal confidence coefficient, but are not reported here.

To describe how far actual coverage probabilities typically fall from the nominal confidence level, Table 1 also reports $\sqrt{\int_0^1 (C_n(p) - .95)^2 dp}$, the uniform-weighted root mean squared error of those probabilities about that confidence level. These values indicate that the variability about the nominal level is much smaller for the score confidence interval than for the Wald or exact confidence intervals. The improved performance of the score method relative to the Wald method is no surprise and simply adds to other evidence of this type accumulated over the years (e.g., Ghosh 1979; Vollset 1993). Some readers, though, may be surprised at just how much better the score method does than the exact method. The exact interval remains quite conservative even for moderately large sample sizes when p tends to be near 0 or 1. The Wald interval is also especially inadequate when p is near 0 or 1, partly a consequence of using $\hat{p}$ as its midpoint when the binomial distribution is highly skewed.

Even though the score intervals tend to have considerably higher actual coverage probabilities than the Wald intervals, they are not necessarily wider. In fact, unless the sample proportions fall near 0 or 1, they are shorter. Direct comparison of the formulas for the two interval widths yields that the score interval is narrower than the Wald interval whenever $\hat{p}$ falls within $\sqrt{(n + z^2)/(8n + 4z^2)}$ of 1/2. In particular, since this term decreases in the limit toward $1/\sqrt{8} = .35$ as n increases or $|z|$ decreases, the score interval is narrower than the Wald interval whenever $\hat{p}$ falls in (.15, .85) for any n and any nominal confidence level. See Ghosh (1979) for additional results about the relative lengths of the two types of interval. This comparison has limited relevance, since the actual coverage probabilities of the two methods differ. We mention this, however, to stress that the inadequacy of the Wald approach is not that the intervals are too short.

For fixed n and p, the expected width of an interval estimator is a useful measure of its performance. Figure 2 illustrates the relative sizes of the expected widths for the nominal 95% Wald, score, and exact intervals by plotting them as a function of p, for $n = 15$. For small n, the score intervals tend to be much shorter than exact intervals. The narrowness of the Wald intervals as p approaches 0 or 1 reflects the fact that when $x = 0$ or n, that interval is degenerate at 0 or at 1. By contrast, when $x = 0$, the score interval is $[0, z^2/(n + z^2)] = [0, 3.84/(n + 3.84)]$ and the exact interval is $[0, 1 - (.025)^{1/n}]$, which is approximately $[0, -\log(.025)/n] = [0, 3.69/n]$; the latter shows an extension of the "rule of $3/n$" (Jovanovic and Levy 1997) from the .95 upper confidence bound to .95 confidence limits.

Is anything sacrificed by using the score intervals? Well, since they are not "exact," they are not guaranteed to have coverage probabilities uniformly bounded below by the nominal confidence level, and their actual confidence coefficient (the infimum of such probabilities) is, in fact, well below it. Vollset's (1993) plots of the coverage probabilities as a function of p, for various methods, are illuminating for

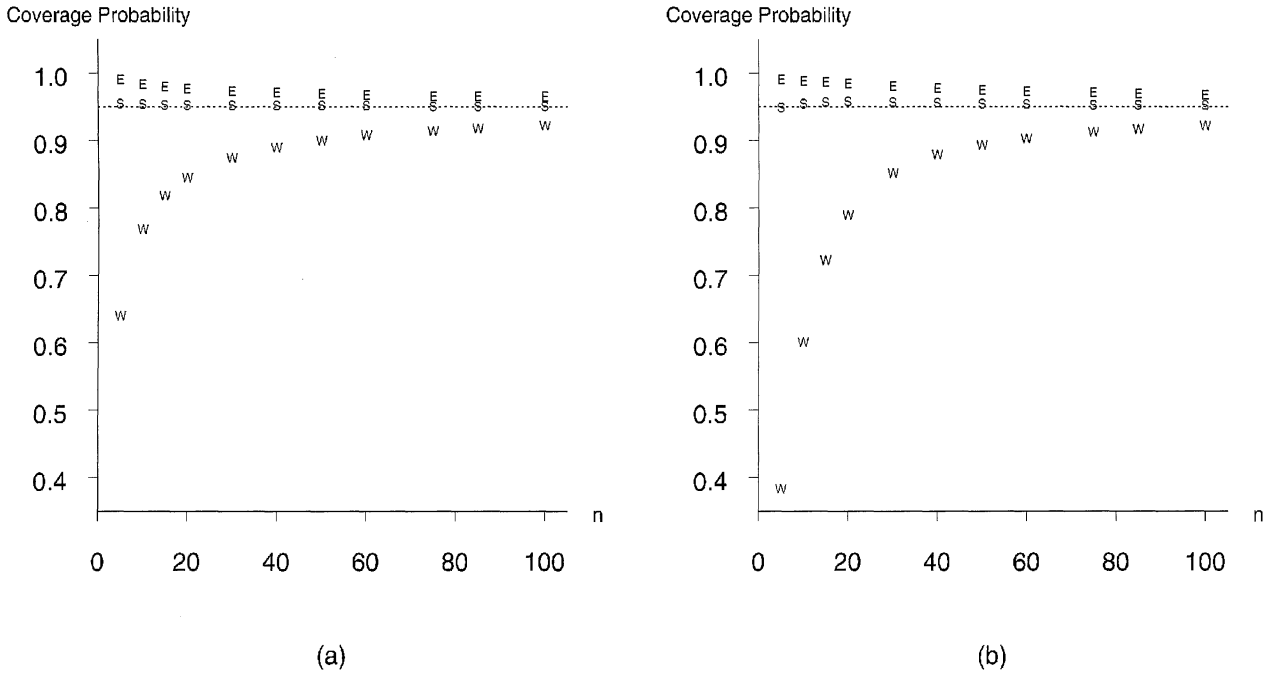

Figure 1. Mean Coverage Probability as a Function of Sample Size for the Nominal 95% Exact (E), Score (S), and Wald (W) Intervals, When p has (a) a Uniform (0,1) Distribution and (b) a Beta Distribution with $\mu = .10$ and $\sigma = .05$.

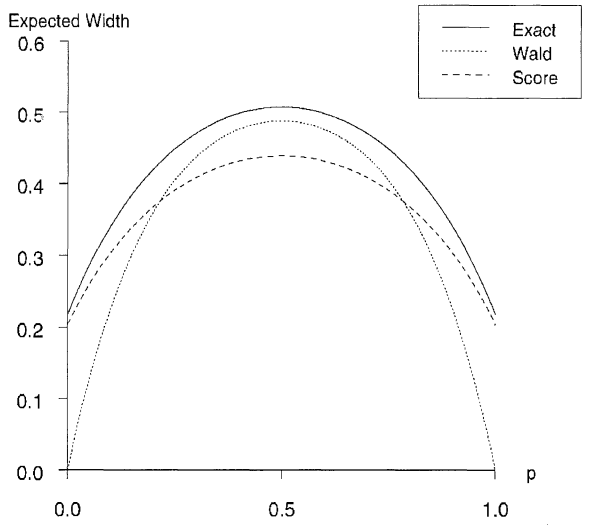

Figure 2. A Comparison of Expected Widths for the Nominal 95% Exact, Wald, and Score Intervals When n = 15.

describing the behavior of the methods. The score method has two very narrow regions of values for p, one near 0 and one near 1, at which the actual coverage probability falls seriously below the nominal confidence level, and this badly affects the actual confidence coefficient. These regions get closer to 0 and to 1 as n increases. For $n = 10$ with nominal 95% confidence intervals, for instance, there is a minimum coverage of .835 at $p = .018$ and $p = .982$, whereas at $n = 100$, there is a minimum coverage of .838 at $p = .002$ and $p = .998$.

We now explain why this happens. There is a region of values $[0, r)$ for p that falls in the score confidence interval only when $X = 0$. The upper bound r of this region is the lower endpoint of the confidence interval when $X = 1$, which for large n is approximately $(1 + z^2/2 - z\sqrt{4 + z^2}/2)/n$. The coverage probability just below r is approximately $P(X = 0) = [1 - (1 + z^2/2 - z\sqrt{4 + z^2}/2)/n]^n \approx \exp\{-(1 + z^2/2 - z\sqrt{4 + z^2}/2)\}$. The analogous remark applies for values of p near 1. This limiting coverage probability is .800 for nominal 90% intervals, .838 for 95% intervals, and .889 for 99% intervals. See Huwang (1995) for related remarks. In particular, the actual confidence level does not converge to the nominal level as n increases.

Though this may seem problematic, the portion of the [0, 1] parameter space over which the actual coverage probability drops seriously below the nominal confidence level is small. Table 2 illustrates. The proportion of the parameter space for which the coverage probability of the nominal 95% score interval falls below .90 is no more than .01 when $n \geq 20$. That table also shows that the proportion of parameter values for which the coverage probability is within .02 of .95 is much higher for the score than the exact interval. In fact, the score coverage probability is closer than the exact coverage probability to .95 over more than 90% of the parameter space, for the sample sizes reported.

3. THE "ADD TWO SUCCESSES AND TWO FAILURES" ADJUSTED WALD INTERVAL

The poor performance of the Wald interval is unfortunate, since it is the simplest approach to present in elementary statistics courses. We strongly recommend that instructors present the score interval instead. Santner (1998) makes the same recommendation. Of course, many instructors will hesitate to present a formula such as (2) in elementary courses. The shrinkage representation of the score approach suggests, however, that for constructing 95% confidence intervals (for which $z^2 = 1.96^2 \approx 4$ and the midpoint of the score interval is $(X + z^2/2)/(n + z^2) \approx (X + 2)/(n + 4)$) an instructor will not go far wrong in giving the following advice: "Add two successes and two failures and then use the Wald formula (1)." That is, this "adjusted Wald" interval uses the usual simple formula presented in such courses, but with $(n + 4)$ trials and point estimate $\tilde{p} = (X + 2)/(n + 4)$.

The midpoint of this interval, $\tilde{p} = (X + 2)/(n + 4)$, is nearly identical to the midpoint of the 95% score interval. It is identical to the Bayes estimate (mean of the posterior distribution) for the beta prior distribution with parameters 2 and 2, which has mean .50 and standard deviation .224 and which shrinks the sample proportion toward .50 somewhat more than does the uniform prior. This simple adjustment to the ordinary Wald interval changes it from highly liberal to slightly conservative, on the average, and a bit more conservative than the score method. Figure 3 illustrates, showing the mean actual coverage probability $\bar{C}_n$ for the nominal 95% Wald and adjusted Wald intervals as a function of n, for the uniform and skewed weightings of p. The adjusted Wald confidence interval behaves surprisingly well, even for very small sample sizes.

Figure 4 shows the actual coverage probabilities as a function of p for the Wald, adjusted Wald, and Clopper–Pearson exact intervals when $n = 5$ and $n = 10$. The im-

Table 2. Proportion of Parameter Space for which (a) Nominal 95% Score Interval has Actual Coverage Probability Below .90; (b) Nominal 95% Score and Exact Intervals Have Actual Coverage Probabilities Between .93 and .97; (c) Actual Coverage Probability is Closer to .95 for Score Interval than Exact Interval

n	Score coverage Prob. below .90	Coverage .93–.97		Coverage closer to .95 for Score than Exact
		Score	Exact	
5	.042	.463	.000	.944
10	.019	.608	.077	.963
20	.010	.792	.297	.925
30	.006	.882	.395	.977
50	.003	.939	.615	.961
100	.002	.968	.830	.961

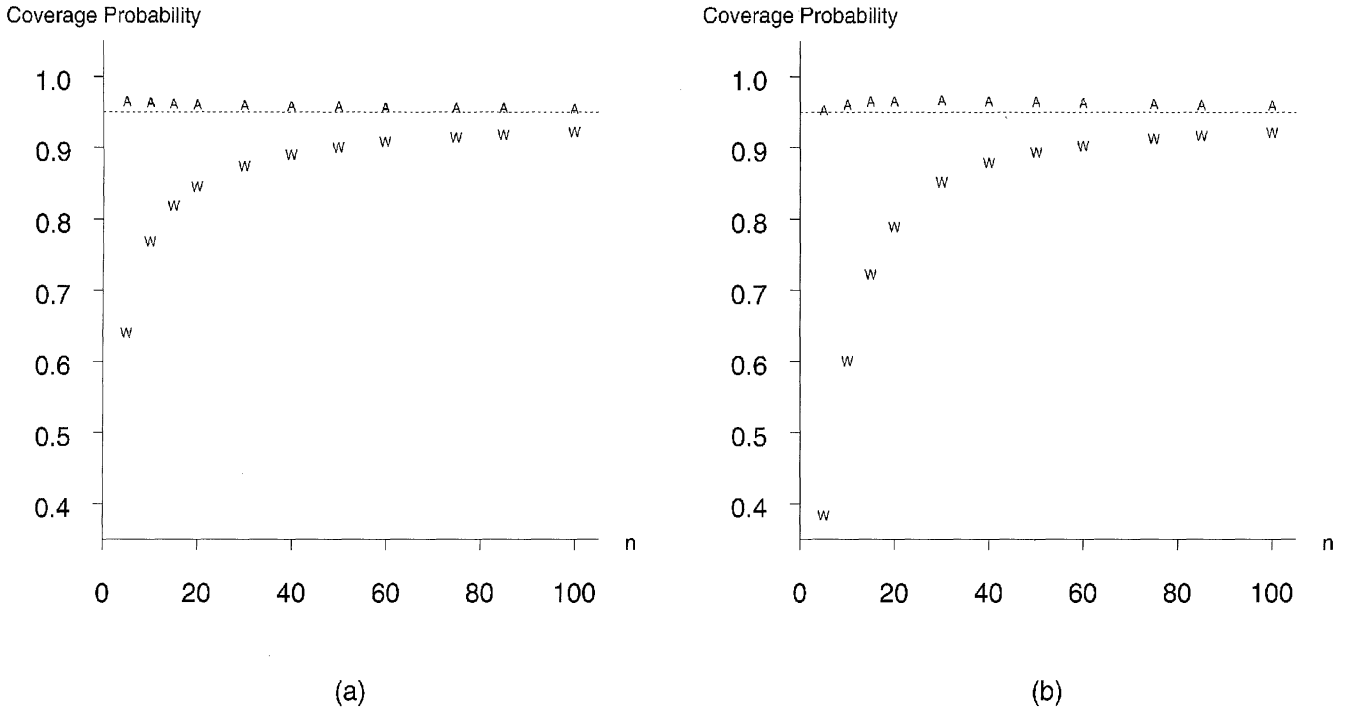

Figure 3. Mean Coverage Probability as a Function of Sample Size for the Nominal 95% Wald (W) and Adjusted Wald (A) Intervals, When p has (a) a Uniform (0,1) Distribution and (b) a Beta Distribution with μ = .10 and σ = .05.

provement of the adjusted Wald interval over the ordinary Wald interval is dramatic. The adjusted Wald interval also has the advantage, relative to the score interval, of not having spikes with seriously low coverage near $p = 0$ and 1. This is because this interval's rather crude bounds contain 0 when $X = 0$ or 1 and contain 1 when $X = n-1$ or n. For

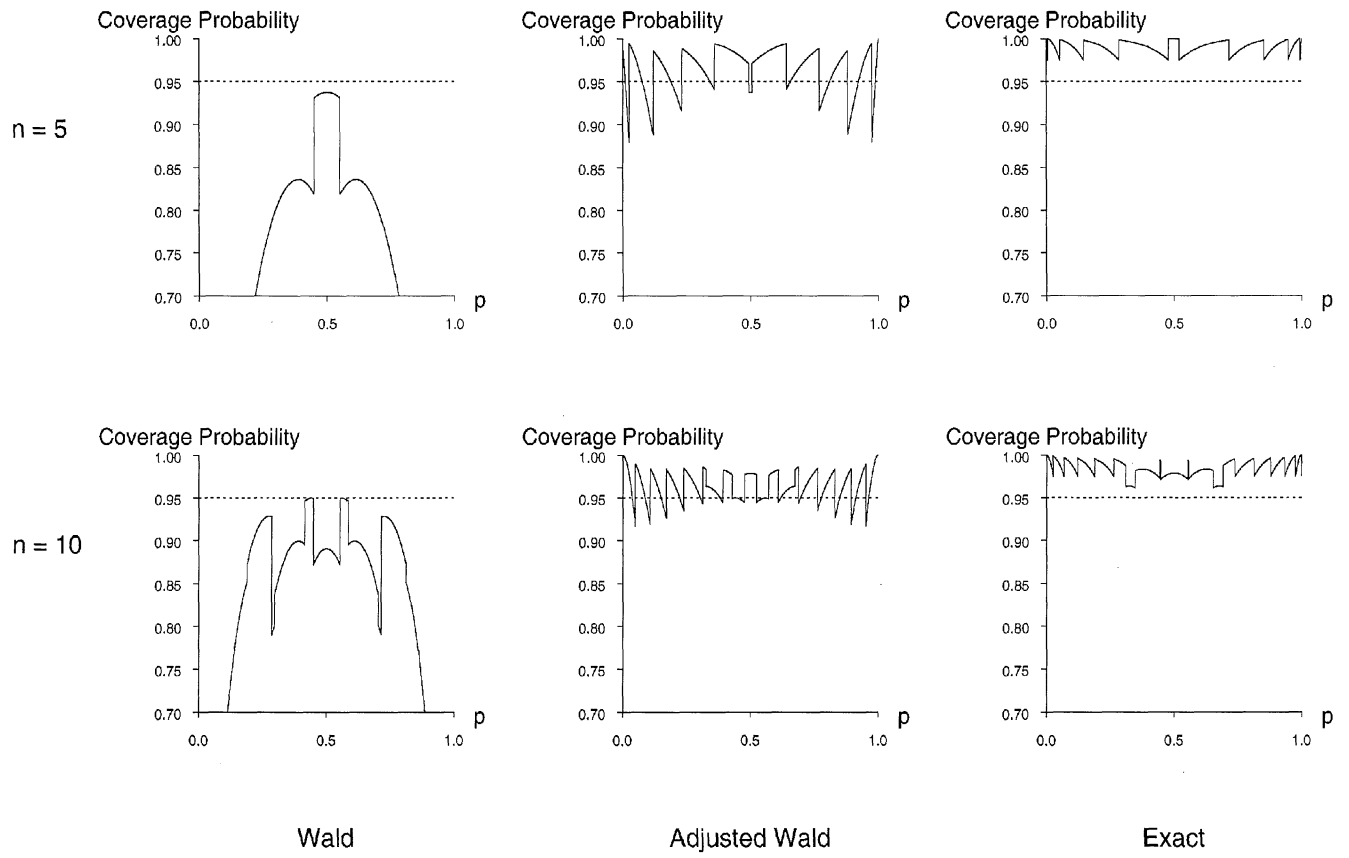

Figure 4. A Comparison of Coverage Probabilities for the Nominal 95% Wald, Adjusted Wald, and Exact Intervals.

instance, the minimum coverage probability for the nominal 95% adjusted Wald interval is .917 for $n = 10$ and never falls below .92 for $n > 10$. The proportion of the parameter space for which the actual coverage probability falls within .02 of .95 is slightly less than reported in Table 2 for the score interval, but the proportion of times its actual coverage probability is closer to .95 than the exact interval is still at least .94 for the sample sizes reported in that table. See Chen (1990) for results about coverage properties of related intervals using Bayes estimates as midpoints.

Introductory statistics textbooks have an awkward time with sample size recommendations for the Wald interval. Most simple recommendations tend to be inadequate (Leemis and Trivedi 1996). Our results suggest that if one tells students to add two successes and two failures before they form the Wald 95% interval, it is not necessary to present such sample size rules, since the "add two successes and two failures" confidence interval behaves adequately for practical application for essentially any n regardless of the value of p.

One can use the adjusted Wald interval without regarding its midpoint $\tilde{p} = (X + 2)/(n + 4)$ as the preferred point estimate of p. However, this rather strong shrinkage toward .5 might often provide a more appealing estimate than $\hat{p}$. The mean square error of $\tilde{p}$ equals $[np(1 - p) + 16(p - .5)^2]/(n + 4)^2$, which is smaller than that of $\hat{p}$ when p is within $\sqrt{3n^2 + 8n + 4}/(6n + 4)$ of .5; this interval of values of p decreases from (.113, .887) to (.211, .789) as n increases. Interestingly, Wilson (1927) mentioned this shrinkage estimator as a reasonable alternative to the sample proportion or the Laplace estimator $(X + 1)/(n + 2)$. Letting S denote X, the number of successes, Wilson stated, "As the distribution of chances of an observation is asymmetric, it is perhaps unfair to take the central value as the best estimate of the true probability; but this is what is actually done in practice. . . . Those who make the usual allowance of 2σ for drawing an inference would use $(S + 2)/(n + 4)$."

In recognition of his pioneering work, predating the famous articles by Neyman and Pearson on confidence intervals, we suggest that statisticians refer to $\tilde{p} = (X + 2)/(n + 4)$ as the Wilson point estimator of p and refer to the score confidence interval for p as the Wilson method. See Stigler (1997) for an interesting summary of Edwin B. Wilson's career. Other highlights included service as the first professor and head of the Department of Vital Statistics at Harvard School of Public Health in 1922, the Wilson–Hilferty normal approximation for the chi-squared distribution in 1931, and the Wilson–Worcester introduction of the median lethal dose (LD 50) in bioassay.

4. OTHER INTERVAL ESTIMATION METHODS FOR p

Although the focus of this article is comparison of the Wald, score, and exact intervals, which are the methods commonly presented in statistics textbooks, we next briefly discuss some alternative methods. Some elementary textbooks (e.g., Siegel 1988), perhaps recognizing the poor performance of the Wald intervals, suggest using ordinary t confidence intervals for a mean for interval estimation of a proportion. These intervals are wider than the Wald intervals, of course, but we found that mean coverage probabilities are still seriously deficient. Table 1 illustrates for the uniform weighting.

Other, more complex, methods exist for constructing exact confidence intervals, such as presented by Blyth and Still (1983) and Duffy and Santner (1987). Our evaluations of these intervals indicated that they perform better than the Clopper–Pearson intervals but not as well as the score intervals, still showing considerable conservatism. To reduce the conservativeness inherent in exact methods for discrete distributions, many authors recommend using tests and confidence intervals based on the mid-P value, namely half the probability of the observed result plus the probability of more extreme results (Lancaster 1961). The mid-P confidence interval is the inversion of the adaptation of the exact test that uses the mid-P value. Results in Vollset (1993) suggest that the mid-P interval tends to perform well but is somewhat more conservative than the score interval, typically having actual coverage probability greater than (and

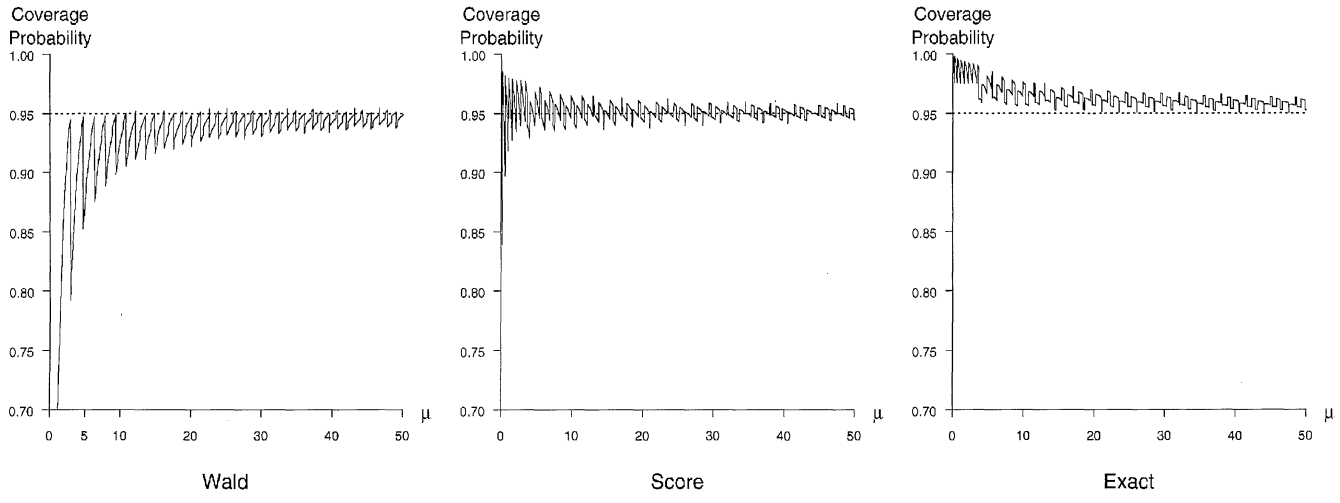

Figure 5. A Comparison of Coverage Probabilities for the Nominal 95% Wald, Score, and Exact Intervals for a Poisson Mean.

never much less than) the nominal confidence level. Our evaluations agreed with this, and are also illustrated in Table 1. We feel this is a reasonable method to use, especially if one is concerned that p may be very close to 0 or 1. It is more complex computationally than the score and adjusted Wald intervals, but like those intervals it has the advantage of being shorter than the exact interval.

Yet another alternative method is a continuity-corrected version of the score interval, based on the normal continuity correction for the binomial. This interval approximates the Clopper–Pearson interval, however, and our evaluations and results in Vollset (1993, Fig. 2) suggest that it is often as conservative as the exact interval itself. Again, Table 1 illustrates, and we do not recommend this approach.

Finally, we mention two other methods that perform well. The confidence interval based on inverting the likelihood-ratio test is similar to the score interval in terms of how it compares with the exact interval, but it is more complex to construct. Not surprisingly, Bayesian confidence intervals with beta priors that are only weakly informative also perform well in a frequentist sense (see, e.g., Carlin and Louis 1996, pp. 117–123).

In deciding whether to use the score interval, some may be bothered by its poor coverage for values of p just below the lower boundary of the interval when $X = 1$ and just above the upper boundary of the interval when $X = n - 1$. One could then use an adapted version that replaces the lower endpoint by $-\log(1-\alpha)/n$ when $X = 1$ and the upper endpoint by $1 + \log(1-\alpha)/n$ when $X = n - 1$. (e.g., at $p = -\log(1-\alpha)/n$, $P(X = 0) = [1 + \log(1-\alpha)/n]^n \approx 1 - \alpha$.) This adaptation improves the minimum coverage considerably. For instance, the nominal 95% interval has minimum coverage probability converging to .895 for large n, which is the large-sample coverage probability at p just below the lower endpoint of the interval when $X = 2$.

5. CONCLUSION AND EXTENSIONS

The Clopper–Pearson interval has coverage probabilities bounded below by the nominal confidence level, but the typical coverage probability is much higher than that level. The score and adjusted Wald intervals can have coverage probabilities lower than the nominal confidence level, yet the typical coverage probability is close to that level. In forming a 95% confidence interval, is it better to use an approach that guarantees that the actual coverage probabilities are *at least* .95 yet typically achieves coverage probabilities of about .98 or .99, or an approach giving narrower intervals for which the actual coverage probability could be less than .95 but is usually quite *close* to .95? For most applications, we would prefer the latter. The score and adjusted Wald confidence intervals for p provide shorter intervals with actual coverage probability usually nearer the nominal confidence level. In particular, even though the score and adjusted Wald intervals leave something to be desired in terms of satisfying the usual technical definition of "95% confidence," the operational performance of those methods is better than the exact interval in terms of how most practitioners interpret that term.

Results similar to those in this article also hold in other discrete problems. For instance, similar comparisons apply for score, Wald, and exact confidence intervals for a Poisson parameter μ, based on an observation X from that distribution. Figure 5 illustrates, plotting the actual coverage probabilities when the nominal confidence level is .95. Here, the score interval for μ results from inverting the approximately normal test statistic $z = (X - \mu_0)/\sqrt{\mu_0}$, the Wald interval results from inverting $z = (X - \mu_0)/\sqrt{X}$, and the endpoints of the exact interval, $(1/2)(\chi^2_{2X,.025}, \chi^2_{2(X+1),.975})$, result from equating tail sums of null Poisson probabilities to .025 (Garwood 1936; for n independent Poisson observations, $X_1, \ldots, X_n$, the same formulas apply if one lets $X = \sum X_i$ and $\mu = E(X) = nE(X_i)$). For another discrete example, see Mehta and Walsh (1992) for a comparison of exact with mid-P confidence intervals for odds ratios or for a common odds ratio in several 2×2 contingency tables.

Exact inference has an important place in statistical inference of discrete data, in particular for sparse contingency table problems for which large-sample chi-squared statistics are often unreliable. However, approximate results are sometimes more useful than exact results, because of the inherent conservativeness of exact methods.

[Received February 1997. Revised November 1997.]

REFERENCES

Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: Wiley.

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

Böhning, D. (1994), "Better Approximate Confidence Intervals for a Binomial Parameter," *Canadian Journal of Statistics*, 22, 207–218.

Carlin, B. P., and Louis, T. A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, London: Chapman and Hall.

Chen, H. (1990), "The Accuracy of Approximate Intervals for a Binomial Parameter," *Journal of the American Statistical Association*, 85, 514–518.

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.

Duffy, D. E., and Santner, T. J. (1987), "Confidence Intervals for a Binomial Parameter Based on Multistage Tests," *Biometrics*, 43, 81–93.

Garwood, F. (1936), "Fiducial Limits for the Poisson Distribution," *Biometrika*, 28, 437–442.

Ghosh, B. K. (1979), "A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter," *Journal of the American Statistical Association*, 74, 894–900.

Huwang, L. (1995), "A Note on the Accuracy of an Approximate Interval for the Binomial Parameter," *Statistics & Probability Letters*, 24, 177–180.

Jovanovic, B. D., and Levy, P. S. (1997), "A Look at the Rule of Three," *The American Statistician*, 51, 137–139.

Lancaster, H. O. (1961), "Significance Tests in Discrete Distributions," *Journal of the American Statistical Association*, 56, 223–234.

Laplace, P. S. (1812), *Théorie Analytique des Probabilités*, Paris: Courcier.

Leemis, L. M., and Trivedi, K. S. (1996), "A Comparison of Approximate Interval Estimators for the Bernoulli Parameter," *The American Statistician*, 50, 63–68.

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.

Neyman, J. (1935), "On the Problem of Confidence Limits," *Annals of Mathematical Statistics*, 6, 111–116.

Santner, T. J. (1998), "A Note on Teaching Binomial Confidence Intervals," *Teaching Statistics*, 20, 20–23.

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

Siegel, A. F. (1988), *Statistics and Data Analysis*. New York: Wiley.

Stigler, S. M. (1997), "Edwin Bidwell Wilson," in *Leading Personalities in Statistical Sciences*, eds. N. L. Johnson and S. Kotz, New York: Wiley, pp. 344–346.

Vollset, S. E. (1993), "Confidence Intervals for a Binomial Proportion," *Statistics in Medicine*, 12, 809–824.

Wilson, E. B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, 22, 209–212.