Teacher's Corner

Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures

Alan AGRESTI and Brian CAFFO

The standard confidence intervals for proportions and their differences used in introductory statistics courses have poor performance, the actual coverage probability often being much lower than intended. However, simple adjustments of these intervals based on adding four pseudo observations, half of each type, perform surprisingly well even for small samples. To illustrate, for a broad variety of parameter settings with 10 observations in each sample, a nominal 95% interval for the difference of proportions has actual coverage probability below .93 in 88% of the cases with the standard interval but in only 1% with the adjusted interval; the mean distance between the nominal and actual coverage probabilities is .06 for the standard interval, but .01 for the adjusted one. In teaching with these adjusted intervals, one can bypass awkward sample size guidelines and use the same formulas with small and large samples.

KEY WORDS: Binomial distribution; Score test; Small sample; Wald test.

1. INTRODUCTION

Let X denote a binomial variate for n trials with parameter p, denoted bin(n, p), and let $\hat{p} = X/n$ denote the sample proportion. For two independent samples, let X_1 be bin(n_1, p_1), and let X_2 be bin(n_2, p_2). Let z_a denote the $1-a$ quantile of the standard normal distribution. Nearly all elementary statistics textbooks present the following confidence intervals for p and $p_1 - p_2$:

- An approximate $100(1-\alpha)\%$ confidence interval for p is

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

- An approximate $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}. \quad (2)$$

These confidence intervals result from inverting large-sample *Wald tests*, which evaluate standard errors at the maximum likelihood estimates. For instance, the interval for p is the set of p_0 values for which $|\hat{p} - p_0|/\sqrt{\hat{p}(1-\hat{p})/n} < z_{\alpha/2}$; that is, the set of p_0 having P value exceeding α in testing $H_0 : p = p_0$ against $H_a : p \neq p_0$ using the approximately normal test statistic. The intervals are sometimes called *Wald intervals*. Although these intervals are simple and natural for students who have previously seen analogous large-sample formulas for means, a considerable literature shows that they behave poorly (e.g., Ghosh 1979; Vollset 1993; Newcombe 1998a, 1998b). This can be true even when the sample size is very large (Brown, Cai, and DasGupta 1999). In this article, we describe simple adjustments of these intervals that perform much better but can be easily taught in the typical non-calculus-based statistics course.

These references showed that a much better confidence interval for a single proportion is based on inverting the test with standard error evaluated at the null hypothesis, which is the *score test* approach. This confidence interval, due to Wilson (1927), is the set of p_0 values for which $|\hat{p} - p_0|/\sqrt{p_0(1-p_0)/n} < z_{\alpha/2}$, which is

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

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

The midpoint is a weighted average of $\hat{p}$ and $1/2$, and it equals the sample proportion after adding $z^2_{\alpha/2}$ pseudo observations, half of each type. The square of the coefficient of $z_{\alpha/2}$ in this formula is 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 $n + z^2_{\alpha/2}$ in place of the usual sample size n. For the 95% case, Agresti and Coull (1998) used this representation to motivate approximating the score interval by the ordinary Wald interval (1)

Alan Agresti is Professor, and Brian Caffo is a Graduate Student, Department of Statistics, University of Florida, Gainesville, FL 32611-8545 (E-mail: AA@STAT.UFL.EDU). This work was partially supported by grants from the National Institutes of Health and the National Science Foundation. The authors appreciate helpful comments from Brent Coull and Yongyi Min.

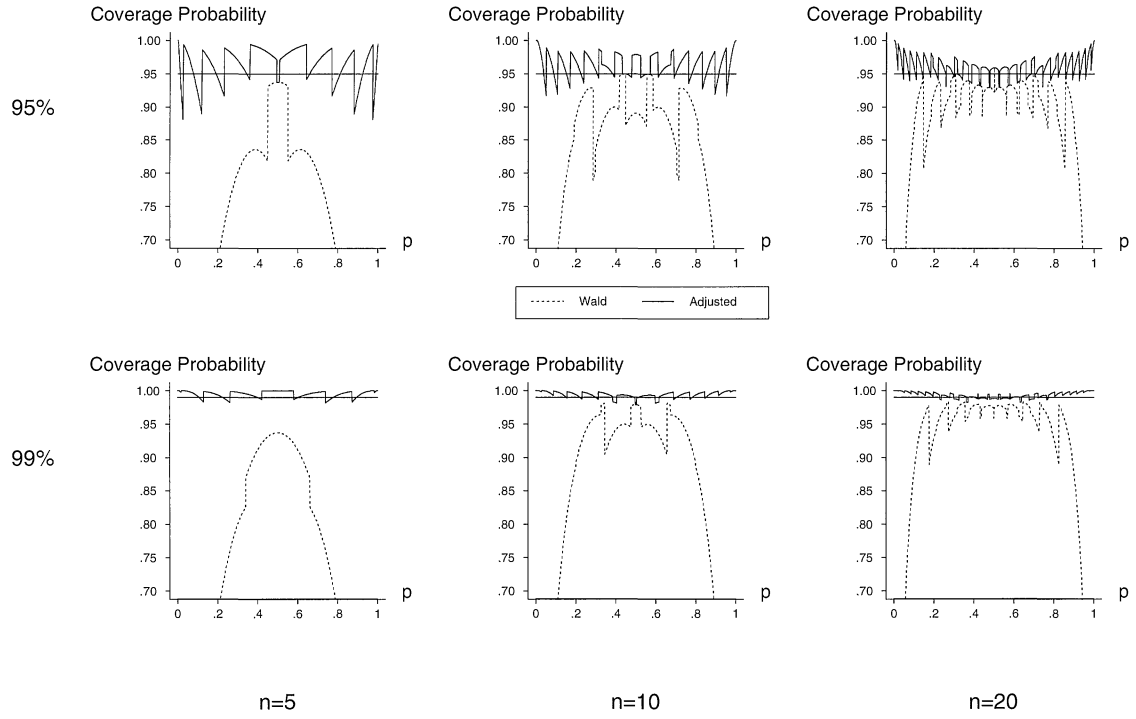

Figure 1. Coverage probabilities for the binomial parameter p with the nominal 95% and 99% Wald confidence interval and the adjusted interval based on adding four pseudo observations, for n = 5, 10, 20.

after adding $z^2_{.025} = 1.96^2 \approx 4$ pseudo observations, two of each type. That is, their adjusted "add two successes and two failures" interval has the simple form

$$\tilde{p} \pm z_{.025}\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}\,, \tag{3}$$

but with $\tilde{n} = (n+4)$ trials and $\tilde{p} = (X+2)/(n+4)$. The midpoint equals that of the 95% score confidence interval (rounding $z_{.025}$ to 2.0 for that interval), but the coefficient of $z_{.025}$ uses the variance $\tilde{p}(1-\tilde{p})/\tilde{n}$ at the weighted average

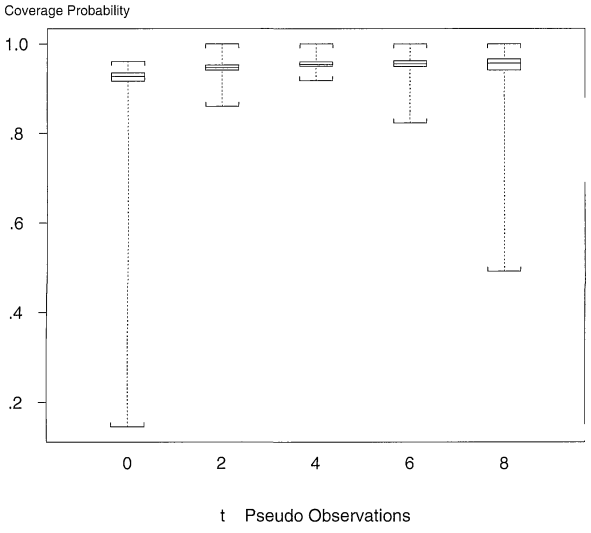

Figure 2. Boxplots of coverage probabilities for nominal 95% adjusted confidence intervals based on adding t pseudo observations; distributions refer to 10,000 cases, with n1 and n2 each chosen uniformly between 10 and 30 and p1 and p2 chosen uniformly between 0 and 1.

$\tilde{p}$ of $\hat{p}$ and $1/2$ rather than the weighted average of the variances; by Jensen's inequality, the adjusted interval is wider than the score interval.

For small samples, the improvement in performance of the adjusted interval compared to the ordinary Wald interval is dramatic. To illustrate, Figure 1 shows the actual coverage probabilities for the nominal 95% Wald and adjusted intervals plotted as a function of p, for $n = 5$, 10, and 20. For all n great improvement occurs for p near 0 or 1. For instance, Brown et al. (1999) stated that when $p = .01$, the size of n required such that the actual coverage probability of a nominal 95% Wald interval is uniformly at least .94 for all n above that value is $n = 7963$, whereas for the adjusted interval this is true for every n; when $p = .10$ the values are $n = 646$ for the Wald interval and $n = 11$ for the adjusted interval. The Wald interval behaves especially poorly with small n for p near the boundary, partly because of the nonnegligible probability of having $\hat{p} = 0$ or 1 and thus the degenerate interval [0, 0] or [1, 1]. Agresti and Coull (1998) recommended the adjusted interval for use in elementary statistics courses, since the Wald interval behaves poorly yet the score interval is too complex for most students. Many students in non-calculus-based courses are mystified by quadratic equations (which are needed to solve for the score interval) and would have difficulty using the weighted average formula above. In such courses, it is often easier to show how to adapt a simple method so that it works well rather than to present a more complex method.

Let $I_t(n,x)$ denote the adjustment of the Wald interval that adds $t/2$ successes and $t/2$ failures. With confidence levels $(1-\alpha)$ other than .95, the Agresti and Coull approximation of the score interval uses $I_t(n,x)$ with $t = z^2_{\alpha/2}$

Table 1. Summary of Performance of Nominal 95% Confidence Intervals for $p_1 - p_2$ Based on Adding t Pseudo Observations, Averaging with Respect to a Uniform Distribution for (p_1,p_2).

Characteristic	n	Number of Pseudo Observations t					Hybrid Score	Approximate Bayes
		0	2	4	6	8		
Coverage	10	.891	.949	.960	.958	.945	.954	.952
	20	.924	.949	.956	.955	.948	.953	.951
	30	.933	.949	.954	.954	.949	.950	.951
	30, 10	.895	.948	.959	.959	.950	.950	.952
Distance	10	.059	.014	.013	.020	.035	.014	.012
	20	.026	.008	.008	.012	.022	.009	.007
	30	.017	.006	.006	.008	.016	.008	.006
	30, 10	.055	.018	.012	.013	.023	.010	.011
Length	10	.647	.670	.673	.668	.659	.654	.647
	20	.480	.487	.488	.487	.485	.481	.477
	30	.398	.401	.401	.401	.401	.398	.396
	30, 10	.537	.551	.553	.551	.545	.537	.536
Cov. Prob. $< .93$	10	.880	.090	.010	.100	.235	.072	.046
	20	.404	.016	.002	.046	.175	.020	.008
	30	.180	.005	.000	.023	.131	.009	.002
	30, 10	.934	.112	.004	.028	.173	.029	.018

NOTE: Table reports mean of coverage probabilities $C_t(n,p_1; n,p_2)$, mean of distances $|C_t(n,p_1; n,p_2) - .95|$ from nominal level, mean of expected interval lengths, and proportion of cases with $C_t(n,p_1; n,p_2) < .93$.

instead of $t = 4$, for instance adding 2.7 pseudo observations for a 90% interval and 5.4 for a 99% interval. Many instructors in elementary courses will find it simpler to tell students to use the same constant for all cases. One will do reasonably well, especially at high nominal confidence levels, by the recipe of always using $t = 4$. The performance of the adjusted interval $I_4(n, x)$ is much better than the Wald interval (1) for the usual confidence levels. To illustrate, Figure 1 also shows coverage probabilities for nominal 99% intervals, when $n = 5$, 10, 20. Since the .95 confidence level is the most common in practice and since this "add two successes and two failures" adjustment provides strong improvement over the Wald for other levels as well, it is simplest for elementary courses to recommend that adjustment uniformly. Of the elementary texts that recommend adjustment of the Wald interval by adding pseudo observations, some (e.g., McClave and Sincich 2000) direct students to use $I_4(n, x)$ regardless of the confidence coefficient whereas others (e.g., Samuels and Witmer 1999) recommend $t = z^2_{\alpha/2}$.

The purpose of this article is to show that a simple adjustment, adding two successes and two failures (total), also works quite well for two-sample comparisons of proportions. The simple Wald formula (2) improves substantially after adding a pseudo observation of each type to each sample, regarding sample i as $(n_i + 2)$ trials with $\tilde{p}_i = (X_i + 1)/(n_i + 2)$. There is no reason to expect an optimal interval to result from this method, or in particular from adding the same number of pseudo observations to each sample or even the same number of cases of each type, but we restricted attention to this form because of the simplicity of explaining it in a classroom setting.

2. COMPARING PERFORMANCE OF WALD INTERVALS AND ADJUSTED INTERVALS

For the two-sample comparison of proportions, we now study the performance of the Wald confidence formula (2) after adding t pseudo observations, $t/4$ of each type to each sample, truncating when the interval for $p_1 - p_2$ contains values < -1 or > 1. Denote this interval by $I_t(n_1, x_1; n_2, x_2)$, or I_t for short, so I_0 denotes the ordinary Wald interval. Our discussion refers mainly to the .95 confidence coefficient, but our evaluations also studied .90 and .99 coefficients. Let $C_t(n_1, p_1; n_2, p_2)$, or C_t for short, denote the true coverage probability of a nominal 95% confidence interval I_t. We investigated whether there is a t value for which $|C_t(n_1, p_1; n_2, p_2) - .95|$ tends to be small for most

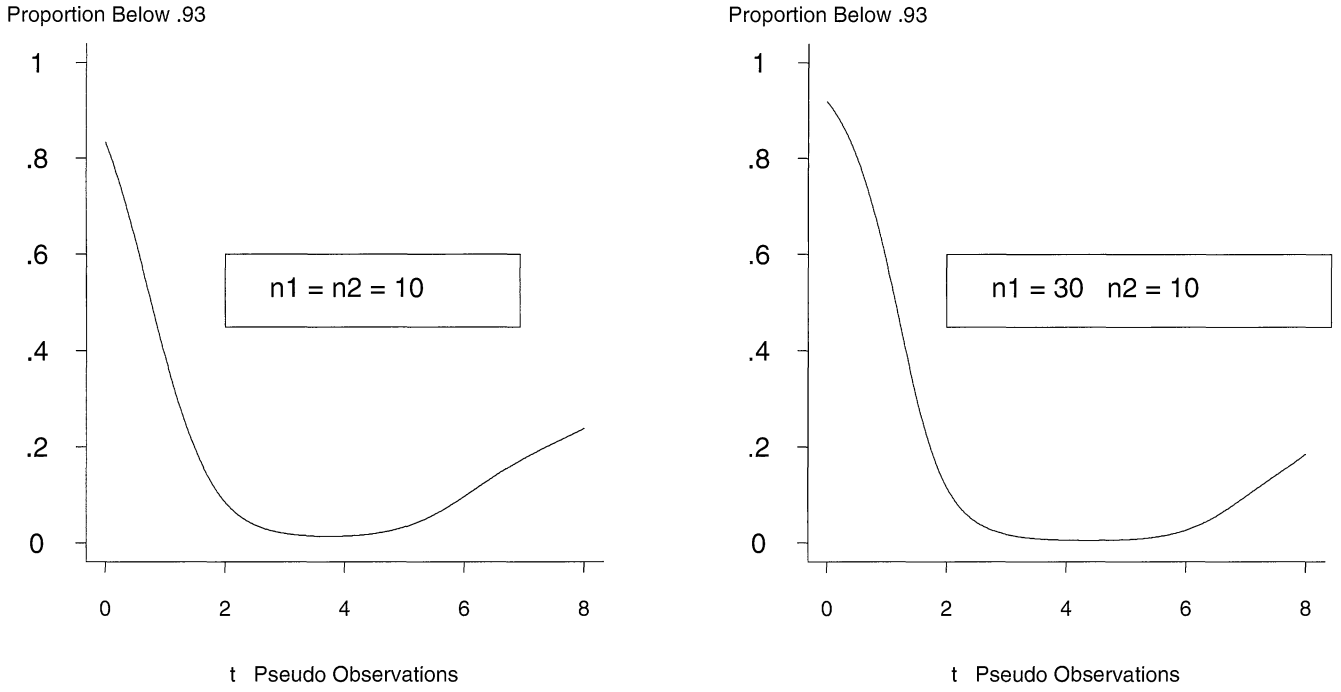

Figure 3. Proportion of (p1, p2) cases with p1 and p2 chosen uniformly between 0 and 1 for which nominal 95% adjusted confidence intervals based on adding t pseudo observations have actual coverage probabilities below ,93, for n1 = n2 = 10 and n1 = 30, n2 = 10.

(p_1, p_2), even with small n_1 and n_2, with C_t rarely very far (say .02) below .95. To explore the performance for a variety of t with small n_i, we randomly sampled 10,000 values of $(n_1, p_1; n_2, p_2)$, taking p_1 and p_2 independently from a uniform distribution over [0,1] and taking n_1 and n_2 independently from a uniform distribution over $\{10, 11, \ldots, 30\}$. For each realization we evaluated $C_t(n_1, p_1; n_2, p_2)$ for t between 0 and 8. Figure 2 illustrates results, showing skeletal box plots of C_t for $t = 0, 2, 4, 6, 8$ (i.e., adding 0, .5, 1, 1.5, 2 observations of each type to each sample).

The ordinary 95% Wald interval behaves poorly. Its coverage probabilities tend to be too small, and they converge to 0 as each p_i moves toward 1 or 0. The coverages for I_t improve greatly for the positive values of t. The case I_4 with four pseudo observations behaves especially well, having relatively few poor coverage probabilities. For instance, the proportion of cases for $t = (0, 2, 4, 6, 8)$ that had $C_t < .93$ were (.572, .026, .002, .046, .171). Similarly, the proportion of nominal 99% intervals that had actual coverage probability below .97 were (.310, .012, .000, .000, .000), and the proportion of nominal 90% intervals that had ac-

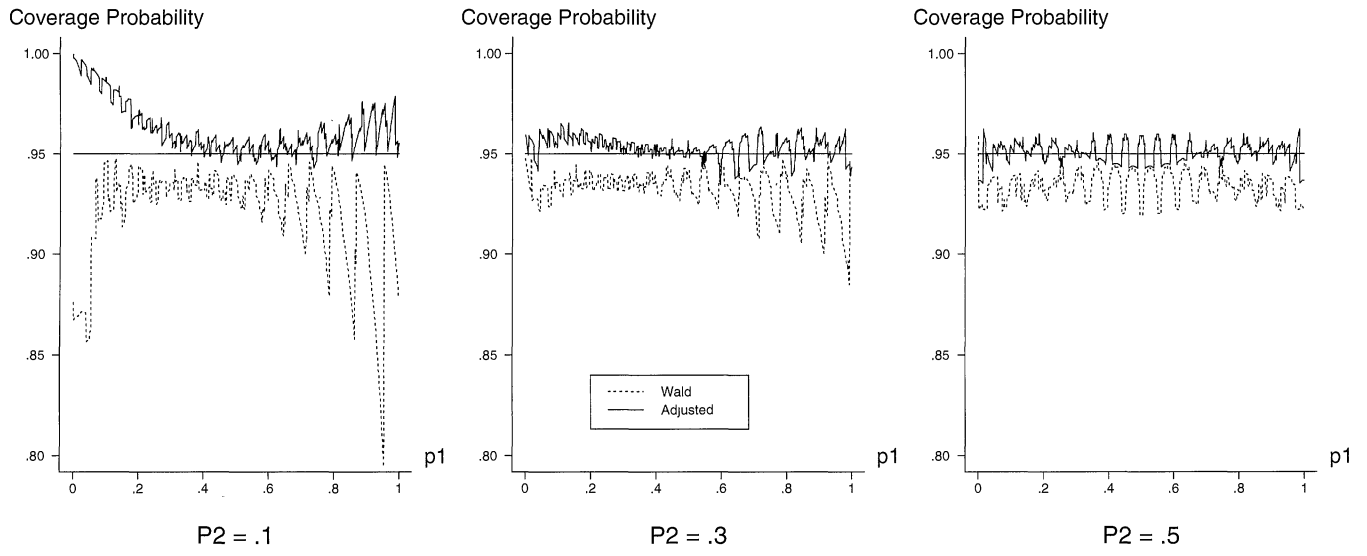

Figure 4. Coverage probabilities for nominal 95% Wald and adjusted confidence intervals (adding t = 4 pseudo observations) as a function of p1 when p2 = .1, .3, .5, with n1 = n2 = 20.

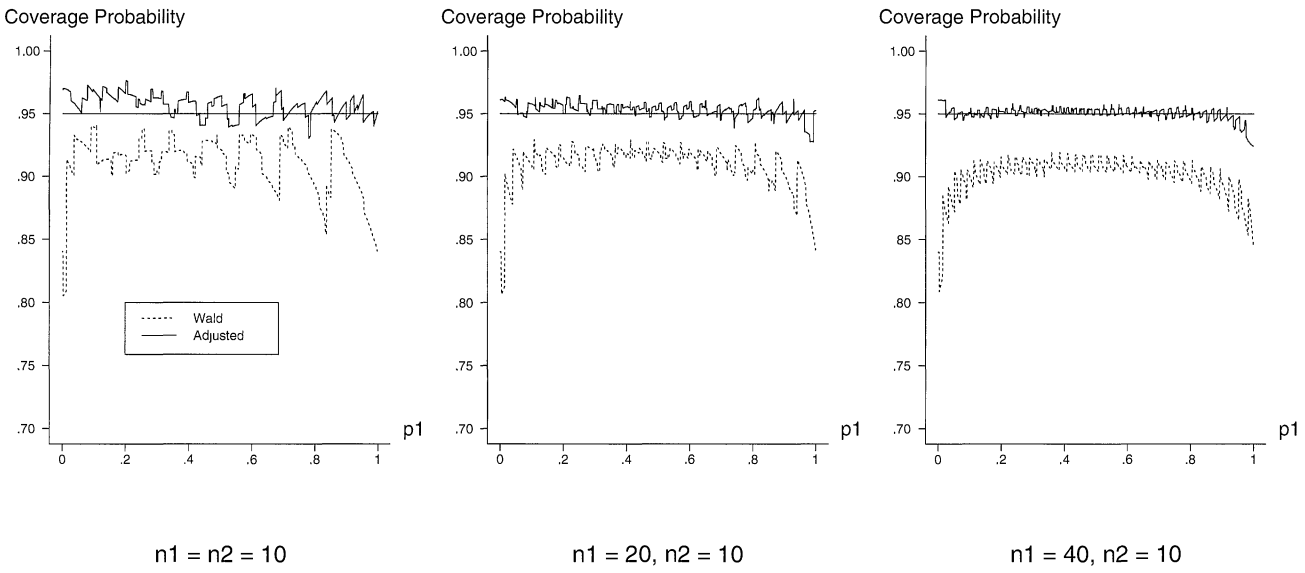

Figure 5. Coverage probabilities for nominal 95% Wald and adjusted confidence intervals (adding t = 4 pseudo observations) as a function of p1 when p2 = .3 when n1 = n2 = 10, n1 = 20, n2 = 10, and n1 = 40, n2 = 10.

tual coverage probability below .88 were (.623, .045, .016, .131, .255). The pattern exhibited here is illustrative of a variety of results from analyzing C_t more closely, as we now discuss.

We analyzed the performance of the I_t interval for various fixed (n_1, n_2) combinations. Table 1 summarizes some characteristics, in an average sense based on taking (p_1, p_2) uniform from the unit square, for $(n_1, n_2) = (10, 10), (20, 20), (30, 30), (30, 10)$. Although the adjusted interval I_4 tends to be conservative, it compares well to other cases in the mean of the distances $|C_t - .95|$ and especially the proportion of cases for which $C_t < .93$. For $n_i = 10$, for instance, the actual coverage probability is below .93 for 88% of such cases with the Wald interval, but for only 1% of them with I_4. Figure 3 shows the proportions of coverage probabilities that are below .93 as a function of t, for $(n_1, n_2) = (10, 10)$ and (30, 10). The improvement over the ordinary Wald interval from adding $t = 4$ pseudo observations is substantial. Remaining figures concentrate on this particular adjustment, which fared well in a variety of evaluations we conducted.

Averaging performance over the unit square for (p_1, p_2) can mask poor behavior in certain regions, and in practice certain pairings (e.g., $|p_1 - p_2|$ small) are often more common or more important than others. Thus, besides studying these summary expectations, we plotted C_t as a function of p_1 for various fixed values of p_2, $p_1 - p_2$, and p_1/p_2. To illustrate, Figure 4 plots the Wald coverage C_0 and the coverage C_4 for the adjusted interval, fixing p_2 at .1, .3, and .5, for $n_1 = n_2 = 20$. The poor coverage spikes for the Wald interval disappear with I_4, but this adjustment is quite conservative when p_1 and p_2 are both close to 0 or both close to 1. The adjustment I_4 performs reasonably well, and much better than the Wald interval, even with very small or unbalanced sample sizes. Figure 5 illustrates, plotting C_0 and C_4 as a function of p_1 with p_2 fixed at .3, for $(n_1, n_2) = (10, 10), (20, 10)$, and $(40, 10)$. Figure 6 shows C_0 and C_4 as a function of p_1 when $p_1 - p_2 = 0$ or .2 and when the relative risk $p_1/p_2 = 2.0$ or 4.0, when $n_1 = n_2 = 10$. In Figures 4–6, only rarely does the adjusted interval have coverage significantly below the nominal level. On the other hand, Figures 4 and 6 show that it can be very conservative when p_1 and p_2 are both close to 0 or 1, say with $(p_1 + p_2)/2$ below about .2 or above about .8 for the small sample sizes studied here. This is preferred, however, to the very low coverages of the Wald interval in these cases. Figures 7 and 8 illustrate their behavior, showing surface plots of C_0 and C_4 over the unit square when $n_1 = n_2 = 10$. The spikes at values of p_i in Figures 4 and 5 become ridges at values of $p_1 - p_2$ in these figures.

The poor performance of the Wald interval does not occur because it is too short. In fact, for moderate-sized p_i it tends to be too long. For instance, when $n_1 = n_2 = 10$, I_0 has greater expected length than I_4 for p_2 between .11 and .89 when $p_1 = .5$ and for p_2 between .18 and .82 when $p_1 = .3$. When $n_1 = n_2 = n$ and when $\hat{p}_1 = \hat{p}_2 = \hat{p}$, I_0 has greater length than I_t when $\hat{p}$ falls within $\sqrt{.25 - n(4n + t)/[24n^2 + 12nt + 2t^2]}$ of .5. For all $t > 0$, this interval around .5 shrinks monotonically as n increases to $.50 \pm .50/\sqrt{3}$, or $(.21, .79)$, which applies also to the Agresti and Coull (1998) adjusted interval in the single-sample case. As in the single-proportion case, the Wald interval suffers from having the maximum likelihood estimate exactly in the middle of the interval.

There is nothing unique about $t = 4$ pseudo observations in providing good performance of adjusted intervals in the one- and two-sample problems. For instance, Figure 3 and Table 1 show that other adjustments often work well. A region of t values provide substantial improvement over the Wald interval, with values near $t = 2$ being less conservative than $t = 4$. We emphasized the case $t = 4$ earlier for the two-sample case because it rarely has poor coverage. We believe it is worth permitting some conservativeness to

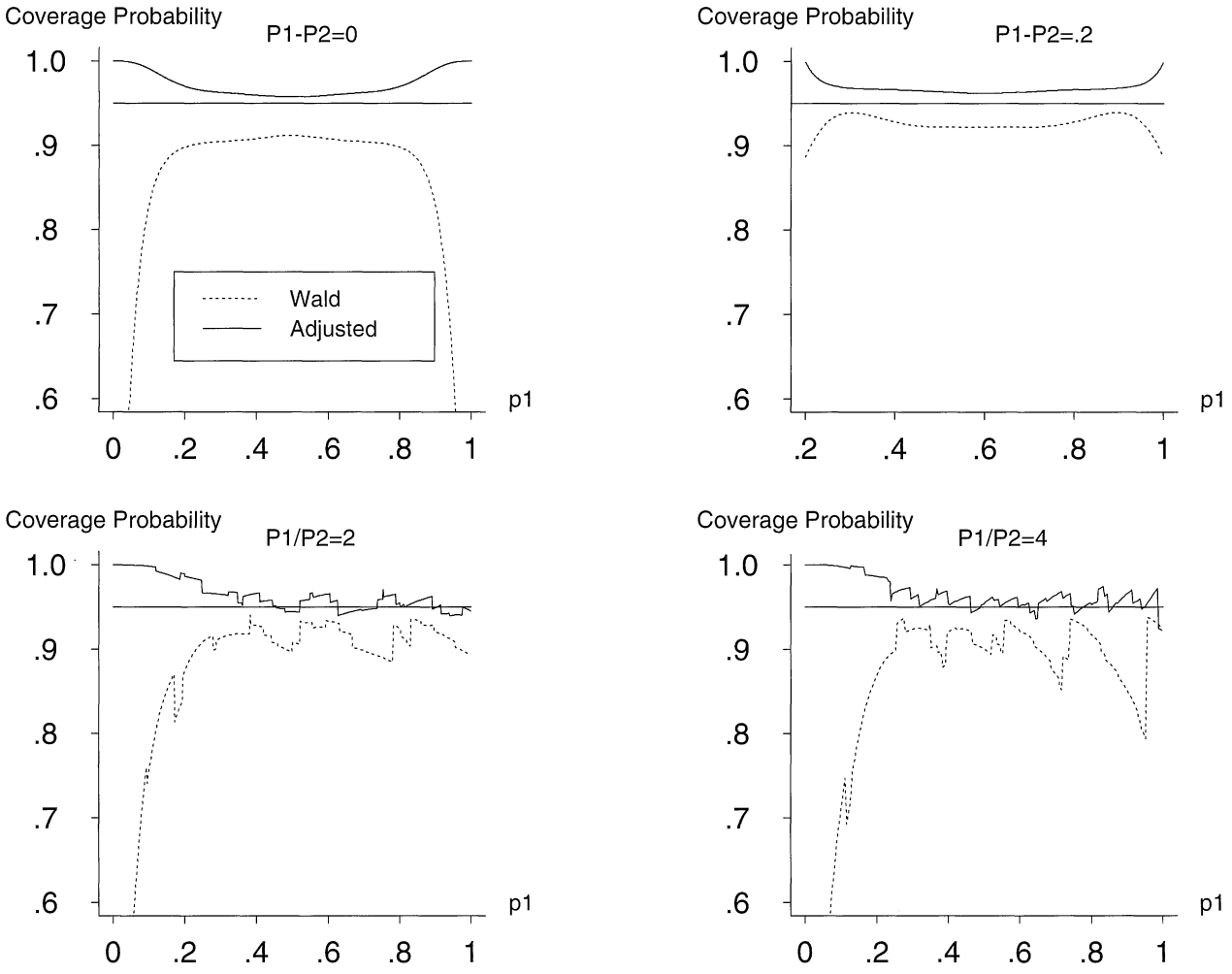

Figure 6. Coverage probabilities for nominal 95% Wald and adjusted confidence intervals (adding t = 4 pseudo observations) as a function of p1 when p1 − p2 = 0 or .2 and when p1/p2 = 2 or 4, for n1 = n2 = 10.

ensure that the coverage probability rarely falls much below the nominal level. In the one-sample case the adjusted interval $I_2(n, x)$ is better than $I_4(n, x)$ in approximating the score interval with small confidence levels, such as 90%. An advantage of the interval $I_2(n, x)$ for p is consistency between the single-sample case and our recommended adjustment $I_4(n_1, x_1; n_2, x_2)$ for two samples. For instance, as $n_2 \to \infty$ and the second sample yields a perfect estimate, the resulting "add two successes and two failures" two-sample interval uses the first sample in the same way as does the "add one success and one failure" single-sample interval. However, for the single-sample problem we prefer the $I_4(n, x)$ interval, since .95 is by far the most common confidence level in practice and this interval works somewhat better than $I_2(n, x)$ in that case.

3. COMPARING THE ADJUSTED INTERVAL WITH OTHER GOOD INTERVALS

Many methods have been proposed for improving on the ordinary Wald confidence interval for $p_1 - p_2$. Since this article dicusses methods appropriate in elementary statistics courses, it focuses on the simple I_t adjustment rather than methods that may be suggested by statistical principles. To find a good method more generally, one approach is to invert a test of $H_0 : p_1 - p_2 = \Delta$ that has good properties, such as using the large-sample score test (Mee 1984) or profile likelihood methods (Newcombe 1998b). The score test of $p_1 - p_2 = 0$ is the familiar Pearson chi-squared test, so this approach has the advantage that the confidence interval is consistent with the most commonly taught test of the same nominal level. The method of obtaining the confidence interval is too complex for elementary courses, however, partly because the test of $p_1 - p_2 = \Delta$ requires finding the maximum likelihood estimates of (p_1, p_2) for the standard error subject to the constraint $\hat{p}_1 - \hat{p}_2 = \Delta$.

Newcombe (1998b) evaluated various confidence interval methods for $p_1 - p_2$. He proposed a method that performs substantially better than the Wald interval and similar to the score interval, while being computationally simpler (although too complex for most elementary statistics courses). His method is a hybrid of results from the single-sample score intervals for p_1 and p_2. Specifically, let (ℓ_i, u_i) be the roots for p_i in $z_{\alpha/2} = |\hat{p}_i - p_i| / \sqrt{p_i(1 - p_i)/n_i}$. Newcombe's

hybrid score interval is

$$\left[(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\ell_1(1-\ell_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}},\right.$$
$$\left.(\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{\ell_2(1-\ell_2)}{n_2}}\right]$$

Compared to the adjusted interval I_4, the hybrid score interval also is conservative when p_1 and p_2 are both close to 0 or 1; overall, it is less conservative, however, with mean coverage probability closer to the nominal level (see Table 1). Likewise, it tends to be a bit shorter. It has a somewhat higher proportion of cases with coverage probability being too small, mainly for values of $|p_1 - p_2|$ near 1; for the 10,000 randomly selected cases with n_i also random between 10 and 30, the minimum coverage probability was .92 for the 95% adjusted interval and .86 for the 95% hybrid score interval.

The adjusted interval I_4 and the hybrid score interval both have a greater tendency for distal non-coverage then mesial non-coverage. For instance, for the 10,000 randomly selected cases, the mean probability for which the lower limit exceeds $p_1 - p_2$ when $p_1 - p_2 > 0$ or the upper limit is less than $p_1 - p_2$ when $p_1 - p_2 < 0$ was .030 for I_4 and .033 for the 95% hybrid score interval, whereas the mean probability for which the upper limit is less than $p_1 - p_2$ when $p_1 - p_2 > 0$ or the lower limit exceeds $p_1 - p_2$ when $p_1 - p_2 < 0$ was .013 for I_4 and .014 for the 95% hybrid score. As t increases for I_t, the ratio of incidence of distal non-coverage to mesial non-coverage increases; for these randomly selected cases, for t = (0, 2, 4, 6, 8) it equals (.7, 1.2, 2.2, 4.3, 8.1). Unlike the adjusted interval and the Wald interval, the hybrid score interval cannot produce *overshoot*,

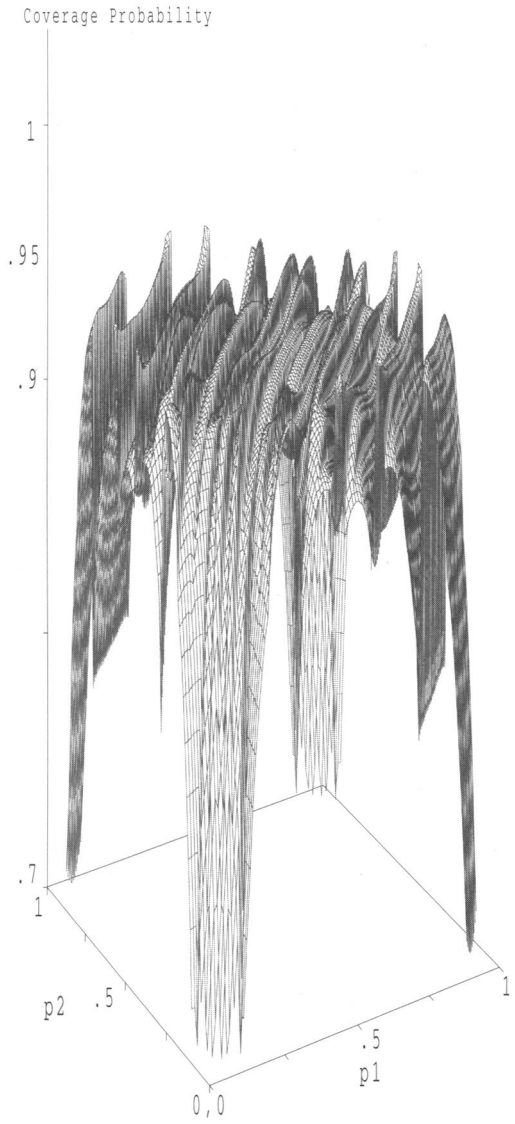

Figure 7. Coverage probabilities for 95% nominal Wald confidence interval as a function of p1 and p2, when n1 = n2 = 10.

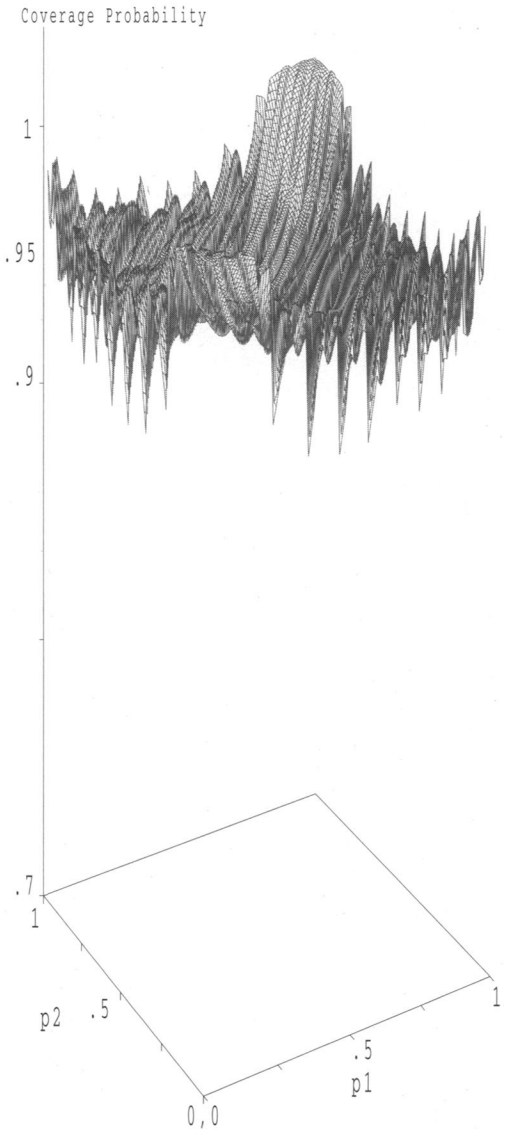

Figure 8. Coverage probabilities for 95% nominal adjusted confidence interval (adding t = 4 pseudo observations) as a function of p1 and p2, when n1 = n2 = 10.

with the interval for $p_1 - p_2$ extending below -1 or above $+1$ and thus requiring truncation. Overshoot for I_t is less common as t increases. For instance, for these randomly selected cases, the mean probability of overshoot for $t =$ (0, 2, 4, 6, 8) was (.048, .033, .016, .006, .000).

Since standard intervals for p and $p_1 - p_2$ improve greatly with adjustment corresponding to shrinkage of point estimates, one would expect intervals resulting from a Bayesian approach with comparable shrinkage also to perform well in a frequentist sense. Carlin and Louis (1996, pp. 117–123) provided evidence of this type for estimating p. For $p_1 - p_2$, consider independent uniform prior distributions for p_1 and p_2. The posterior distribution of p_i is beta with mean $\tilde{p}_i = (X_i+1)/(n_i+2)$ and variance $\tilde{p}_i(1-\tilde{p}_i)/(n_i+3)$. Using a crude normal approximation for the distribution of the difference of the posterior beta variates leads to the interval

$$(\tilde{p}_1 - \tilde{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+3} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+3}}. \qquad (4)$$

This has the same center as the adjusted interval I_4 but uses $n_i + 3$ instead of $n_i + 2$ in the denominators of the standard error. For elementary courses, this interval was suggested by Berry (1996, p. 291). Like Newcombe's hybrid score interval, it tends to perform quite well, being slightly shorter and less conservative than I_4 but suffering occasional poorer coverages (see Table 1). For sample size combinations we considered, its minimum coverage probability was only slightly below that for the adjusted interval. If conservativeness is a concern (e.g., if both p_i are likely to be close to 0), the approximate Bayes and hybrid score intervals are slightly preferable to I_4.

The adjusted interval I_4 (and the similar approximate Bayes interval (4)) is simpler than other methods that improve greatly over the Wald interval. Thus, we believe it is appropriate for elementary statistics courses. We do not claim optimality in any sense or that other methods may not be better for some purposes. Some applications, for instance, may require that the true confidence level be no lower than the nominal level, mandating a method that is necessarily conservative (e.g., Chan and Zhang 1999). Also, we recommend I_4 for interval estimation and not for an implicit test of $H_0 : p_1 - p_2 = 0$, although such a test would be more reliable than one based on the Wald interval. For a significance test, we would continue to teach the Pearson chi-squared test in elementary courses. The test based on I_4 is too conservative when the common value of p_i under the null is close to 0 or close to 1, for most sample sizes more conservative than the Pearson test for such p_i. Although the adjusted interval is not guaranteed to be consistent with the result of the Pearson test, it usually does agree. For instance, for common values (.1, .2, .3, .4, .5) of p_i, the 95% version of I_4 and the Pearson test with nominal significance level of .05 agree with probability (.972, .996, .9996, 1.000, 1.000) when $n_1 = n_2 = 30$ and (1.0, 1.0, 1.0, 1.0, 1.0) when $n_1 = n_2 = 10$.

Finally, an alternative way to improve the Wald method is with a continuity correction (Fleiss 1981, p. 29). As with other continuity corrections, this generally results in conservative performance, usually more so than the adjusted interval. However, the coverage probabilities, like those of the Wald interval, can dip substantially below the nominal level when both p_i are near 0 or 1.

4. TEACHING THE ADJUSTED INTERVALS

Agresti and Coull (1998) motivated their adjusted interval (3) for a single proportion as a simple approximation for the score 95% confidence interval. We know of no such simple motivation for the adjusted interval for the two-sample comparison, other than the similarity with the Bayesian interval (4). A problem for future research is to study whether theoretical support exists for this simple yet effective adjustment, such as Edgeworth or saddlepoint expansions that might provide improved approximations for the tail behavior of $\hat{p}_1 - \hat{p}_2$.

The motivation needed for teaching in the elementary statistics course is quite different. How can one motivate adding pseudo observations? In the single-sample case we remind students that the binomial distribution is highly skewed as p approaches 0 and 1, and because of this perhaps $\hat{p}$ should not be the midpoint of the interval. As support for this, we have students use the software ExplorStat (available at http://www.stat.ufl.edu/∼dwack/). Through simulation it shows how operating characteristics of statistical methods change as students vary sample sizes and population distributions. For instance, when p takes values such as .10 or .90, students observe a relatively high proportion of Wald intervals failing to contain p when n is 30, the sample size their text suggests is adequate for large-sample inference for a mean.

Most students, however, seem more convinced by specific examples where the Wald method seems nonsensical, such as when $\hat{p} = 0$ or 1. We often use data from a questionnaire administered to the students at the beginning of term. For instance, one of us (Agresti) taught a class to 24 honors students in fall 1999. In response to the question, "Are you a vegetarian?", 0 of the 24 students responded "yes," yet they realized that the Wald interval of [0, 0] was not plausible for a corresponding population proportion. We have also used homework exercises such as estimating the probability of success for a new medical treatment when all 10 subjects in a sample experience success, or estimating the probability of death due to suicide when a sample of 30 death records has no occurrences. (Again, the Wald interval is [0, 0], but the National Center for Health Statistics reports that in the United States the probability of death due to suicide is about .01.) Although one can amend the Wald method to improve its behavior when $\hat{p} = 0$ or 1, such as by replacing the endpoints by ones based on the exact binomial test, making such exceptions from a general recipe distracts students from the main idea of taking the estimate plus and minus a normal-score multiple of a standard error.

Why *four* pseudo observations? In the single-sample case we explain that this approximates the results of a more complex method that does not require estimating the unknown standard error; here, we explain the concept of inverting the test with null standard error, or finding solutions of $(\hat{p} - p) = 2\sqrt{p(1-p)/n}$ that do not require estimating $\sqrt{p(1-p)/n}$. In the two-sample case one could explain that this approximates a statistical analysis that represents prior beliefs about each p_i by a uniform distribution. (Some instructors, of course, will prefer a more fully Bayesian approach, as in Berry 1996.)

The poor performance of the ordinary Wald intervals for p and for $p_1 - p_2$ is unfortunate, since they are the simplest and most obvious ones to present in elementary courses. Also unfortunate for these intervals is the difficulty of providing adequate sample size guidelines. Introductory textbooks provide a variety of recommendations, but these have inadequacies (Leemis and Trivedi 1996; Brown et al. 1999). And, needless to say, most texts do not indicate what to do when the guidelines are violated, other than perhaps to consult a statistician. The results in this article suggest that for the "add two successes and two failures" adjusted confidence intervals, one might simply bypass sample size rules. The adjusted intervals have safe operating characteristics for practical application with almost all sample sizes. In fact, we note in closing (and with tongue in cheek) that the adjusted intervals $I_4(n, x)$ and $I_4(n_1, x_1; n_2, x_2)$ have the advantage that, as with Bayesian methods, one can do an analysis without having any data. In the single-sample case the adjusted sample then has $\tilde{p} = 2/4$, and the 95% confidence interval is $.5 \pm 2\sqrt{(.5)(.5)/4}$, or [0, 1]. In the two-sample case the adjusted samples have $\tilde{p}_1 = 1/2$ and $\tilde{p}_2 = 1/2$, and the 95% confidence interval is $(.5 - .5) \pm 2\sqrt{[(.5)(.5)/2] + [(.5)(.5)/2]}$, or $[-1, +1]$. Both analyses are uninformative, as one would hope from a frequentist approach with no data. No one will get into too much trouble using them!

[Received September 1999. Revised February 2000.]

REFERENCES

Agresti, A., and Coull, B. A. (1998), "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions," *The American Statistician*, 52, 119–126.

Berry, D. A. (1996), *Statistics: A Bayesian Perspective*, Belmont, CA: Wadsworth.

Brown, L. D., Cai, T. T., and DasGupta, A. (1999), "Confidence Intervals for a Binomial Proportion and Edgeworth Expansions," technical report 99-18, Purdue University, Statistics Department.

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

Chan, I. S. F., and Zhang, Z. (1999), "Test-Based Exact Confidence Intervals for the Difference of Two Binomial Proportions," *Biometrics*, 55, 1202–1209.

Fleiss, J. L. (1981), *Statistical Methods for Rates and Proportions* (2nd ed.), New York: Wiley.

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

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

McClave, J. T., and Sincich, T. (2000), *Statistics* (8th ed.), Englewood Cliffs, NJ: Prentice Hall.

Mee, R. W. (1984), "Confidence Bounds for the Difference Between Two Probabilities," *Biometrics*, 40, 1175–1176.

Newcombe, R. (1998a), "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods," *Statistics in Medicine*, 17, 857–872.

——— (1998b), "Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods," *Statistics in Medicine*, 17, 873–890.

Samuels, M. L., and Witmer, J. W. (1999), *Statistics for the Life Sciences* (2nd ed.), Englewood Cliffs, NJ: Prentice Hall.

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.