How Much Confidence Should You Have in **Binomial Confidence Intervals?**

Eric A. Suess

Daniel Sultana

Gary Gongwer

Among 18-year-old students, what percentage has clear career goals? Suppose you ask a random sample of $n = 25$ such students from your school and find that $x = 8$ have specific careers in mind. So their proportion in your sample is $\hat{p} = x/n = 8/25 = 0.32$. From this information, you might guess the population proportion, p, for your school is somewhere around 1/3. However, the number of Yes answers, X, is a random variable. Specifically, X is a binomial random variable with $n = 25$ trials and p = probability of a success = P(Yes). How close to $\hat{p}$ might the true value of p lie? The formula for the traditional 95% confidence interval (CI) shown in many elementary statistics books is:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

In our case, this gives 0.32 ± 0.183. Here, 0.32 is the *point estimate* of p and 0.183 is the *margin of error* for the estimate. Notice that both 0.32 and 0.183 are computed from X. According to this formula, we would be 95% confident that the interval 0.137 and 0.503 captures the true value of p. That's a pretty long confidence interval, but with so little data, we can't expect great precision. Of course, if we interviewed more people, we would get a shorter CI.

Now we consider whether our 95% confidence in such intervals is justified. The traditional formula displayed above is based on two assumptions. The argument goes like this:

> ***First:*** The binomial distribution of X is *approximately normal* for large n. So the distribution of $\hat{p} = X/n$ is approximately normal with mean p, and variance $\sigma^2 = p(1-p)/n$, and $(\hat{p} - \mathrm{p})/\sigma$ is approximately standard normal. Thus,
>
> $$P\{-1.96 \le (\hat{p} - p)/\sigma \le 1.96\} = 0.95.$$
>
> Manipulating the inequality in this expression, we find there is 95% probability that p lies in the interval $\hat{p} \pm 1.96\,\sigma$. But, this expression is useless just as it stands. We cannot use it to calculate a CI because p is unknown and so σ is also unknown.
>
> ***Second:*** In order to get a CI, we also assume that σ^2 is well approximated by $\hat{p}(1 - \hat{p})/n$. So, under the square root in the displayed formula for the traditional CI, we assume it is okay to use the estimate $\hat{p}$ instead of the true value of p.

Especially for small values of n, there are good theoretical reasons to be skeptical of both these assumptions. The normal distribution is continuous and symmetrical. The binomial distribution that it is supposed to approximate is discrete and may be skewed when p differs from 1/2. Perhaps more importantly, one has to wonder how much error in the length of the CI arises from using $\hat{p}$ as an estimate of p to get the margin of error. If the CI is longer or shorter than it should be, that would affect the chance it covers the true value of p.

Moreover, there is a serious practical problem with the traditional CI. If $\hat{p} = 0$ or 1, then the estimated margin of error becomes 0 and we have a CI of 0 length. For example, if we sampled 25 cattle at random from the United States and found none of them had mad cow disease, an alleged 99.99% CI would 'guarantee' that the entire United States is free of the disease. How wonderful it would be if life were so simple!

In this article, we will see two things: (1) For small n, the true coverage probability of the traditional CI is

Eric A. Suess, eric.suess@csueastbay.edu, is associate professor of statistics at California State University, East Bay. His particular interest is in computational statistics.

Daniel Sultana, dsultana@horizon.csueastbay.edu, is a master's student in statistics at California State University, East Bay, and a statistician at the California Environmental Protection Agency in the area of cancer prevention.

Gary Gongwer, ggongwer@horizon.csueastbay.edu is a master's student in statistics at California State University, East Bay, and a mathematics teacher at Moreau Catholic High School in Hayward, California.

often distressingly far below 95%, and (2) a very simple modification of the traditional CI works much better.

Exploring Coverage Probabilities of The Traditional CI

What do we mean by "coverage probability"? To answer this question, consider the values $n=25$ for the number of trials and $p=0.3$ for the population proportion. In this case, the random variable X takes 26 values: $x = 0, 1, 2, \ldots 25$. As it turns out, it is sufficient for us to look at the values 2 through 14, and we show them in the first column of Table 1. The second column shows the corresponding possible values of the estimate $\hat{p} = x/n$. The next two columns show the lower and upper confidence limits based on the traditional CI. Notice the interval (0.137, 0.503) mentioned earlier is one of these (look at the box in row 8).

x	Est.	LCL	UCL	Probability
...				
2	0.08	-0.0263	0.1863	
3	0.12	-0.0074	0.2474	
4	0.16	0.0163	0.3037	0.0572
5	0.20	0.0432	0.3568	0.1030
6	0.24	0.0726	0.4074	0.1472
7	0.28	0.1040	0.4560	0.1712
8	0.32	0.1371	0.5029	0.1651
9	0.36	0.1718	0.5482	0.1336
10	0.40	0.2080	0.5920	0.0916
11	0.44	0.2454	0.6346	0.0536
12	0.48	0.2842	0.6758	0.0268
13	0.52	0.3242	0.7158	
14	0.56	0.3654	0.7546	
...				
P{CI covers 0.30} = **0.9493**				

Table 1. Illustrating the Coverage Probability of the Traditional Confidence Interval when $p = 0.30$. This is the sum of the nine probabilities shown in the last column. None of the omitted values smaller than $x = 2$ or greater than $x = 14$ has a CI that covers 0.30.

So far, we have used only the parameter $n=25$. Now we begin to use $p=0.30$. We notice that the CIs resulting from values of x from 4 through 12 cover (include) this value of p, even though the upper end of the CI for $x=4$ is just barely larger than 0.30.

For the last column of Table 1, we compute the binomial probabilities for these outcomes $x=4, 5, \ldots 12$, based on the parameters $n=25$ and $p=0.30$. For example, the first of the relevant probabilities is computed as

$$P\{X = 4\} = \binom{25}{4}\; 03^4 0.7^{21} = 0.0572$$

The coverage probability for $p=0.30$ is the sum of all nine of these probabilities:

$$\begin{aligned} P(\text{Cover}) &= P\{X=4\} + P\{X=5\} + \cdots + P\{X=12\} \\ &= 0.0572 + 0.1030 + \cdots + 0.0268 = 0.9493. \end{aligned}$$

Thus, the coverage probability of the traditional 95% CI is 94.93% when $n=25$ and $p=0.30$. This result is very close to the promised 95% confidence level. So what's the problem?

x	Est.	LCL	UCL	Probability
...				
2	0.08	-0.0263	0.1863	
3	0.12	-0.0074	0.2474	
4	0.16	0.0163	0.3037	
5	0.20	0.0432	0.3568	0.0910
6	0.24	0.0726	0.4074	0.1363
7	0.28	0.1040	0.4560	0.1662
8	0.32	0.1371	0.5029	0.1680
9	0.36	0.1718	0.5482	0.1426
10	0.40	0.2080	0.5920	0.1025
11	0.44	0.2454	0.6346	0.0628
12	0.48	0.2842	0.6758	0.0329
13	0.52	0.3242	0.7158	
14	0.56	0.3654	0.7546	
...				
P{CI covers 0.31} = **0.9024**				

Table 2. Illustrating the Coverage Probability When $p = 0.31$. This is the sum of the eight probabilities shown in the last column. In contrast to Table 1, the confidence interval on row $x = 4$ does not cover $p = 0.31$, so its probability is not included.

The problem is that if we change to $p=0.31$, the interval corresponding to $x=4$ no longer covers p, and the coverage probability drops to 90.24%. Thus, what is supposed to be a 95% CI has nowhere near 95% coverage probability. The probability column of the table changes a bit with $p = 0.31$, but most of this difference results from the loss of the probability corresponding to $x = 4$ (see Table 2 and Figure 1).

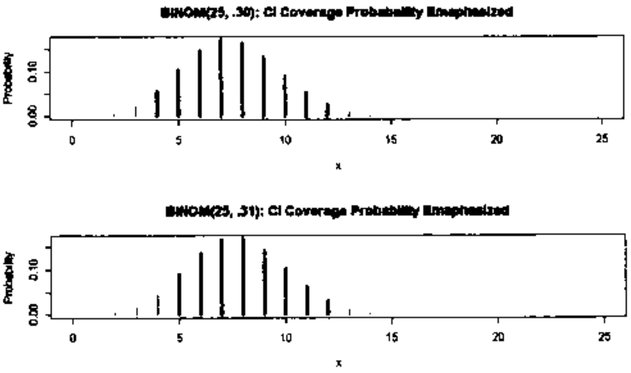

Figure 1. Comparing the coverage probabilities of traditional confidence intervals for $p=0.30$ (top) and $p = 0.31$. A small change in p can result in a large change in the coverage probability of the confidence interval.

We can see that there are "lucky" values of p, such as 0.30, with coverage probabilities close to 95% and "unlucky" ones, such as 0.31, with much smaller coverage probabilities. Unfortunately, it turns out the traditional CI has many more unlucky values of p than lucky ones.

To get a more comprehensive view of the generally bad performance of the traditional 95% CI, we can use the R software package to step through 2,000 values of p ranging from near 0 to near 1 and plot the coverage probability for each of these values of p. The results are shown in Figure 2. It is clear that, for most values of p, the coverage probability is below 95%—often very much below. The two heavy dots in this figure show the coverage probabilities for p=0.30 and 0.31 illustrated in Tables 1 and 2.

Because $n=25$ is a very small number of subjects, it makes sense to see what happens to coverage probabilities for larger values of n. If we look at graphs similar to Figure 2, but with $n=50$ and $n=100$, they unfortunately show very little improvement—and then mainly for values of p near 1/2 (see Figure 3, where $n=100$). The fundamental problem remains: The coverage probability falls far below 95% for many values of p. It seems many unlucky combinations of n and p persist, even for surprisingly large values of n. *The traditional 95% CI for binomial proportions simply cannot be relied upon to provide the promised level of confidence.*

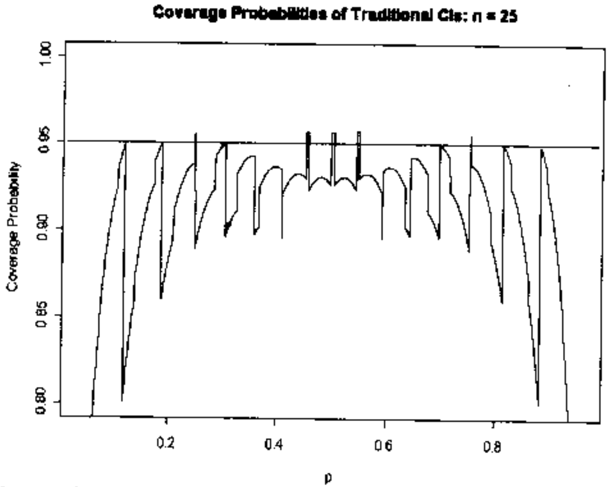

Figure 2. Coverage probabilities for traditional confidence intervals are mostly below 95%. As p changes continuously, the discreteness of the binomial distribution causes some abrupt changes in the coverage probability. Two heavy dots show the coverage probabilities at p = 0.30 and 0.31, which were computed in Tables 1 and 2.

Modified Confidence Intervals

Many proposals have been made to improve the coverage probabilities of CIs for the binomial proportion. Perhaps the simplest of these is the rule to "add two successes and two failures" to the data. This means that X is adjusted to $X_+ = X+2$ and n is adjusted to $n_+=n+4$. Then, the modified point estimate is $\hat{p}_+= X+/n_+=(X + 2)/(n + 4)$. The effect is to "shrink" the distance between the point estimate and 1/2. In order to compute the modified CI, simply use $\hat{p}_+$ and n_+ in place of $\hat{p}$ and n, respectively. This kind of modified CI is called the Agresti-Coull CI or the "plus-four" CI. For our poll with eight Yes answers out of 25, the adjusted results are shown in Table 3, along with our earlier results using the traditional CI.

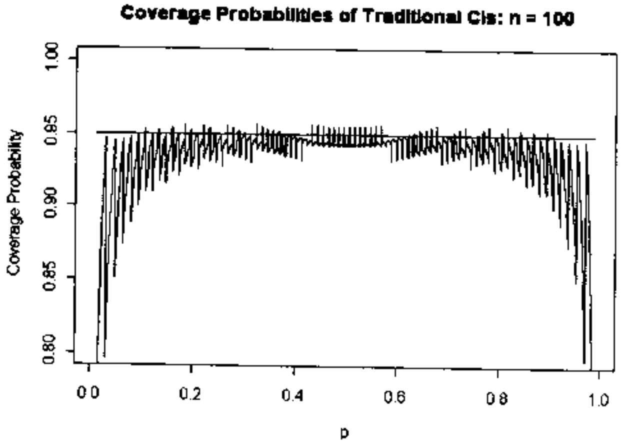

Figure 3. Even for a sample as large as 100, traditional "95% confidence" intervals have coverage probabilities far below 95% for many values of p.

Type of CI	Point Est.	Margin of Error	CI	Length
Traditional	.320	.183	(.137,.503)	.366
Plus-Four	.345	.173	(.172,.518)	.346

Table 3. Comparison of Traditional and Plus-four Confidence Intervals Based on 8 Yes Answers out of 25 Subjects.

Coverage probabilities of 95% plus-four CIs for n = 25 are shown in Figure 4. While coverage probabilities for these CIs are seldom exactly 95%, they are mainly much closer to 95% than for the traditional intervals. Also,

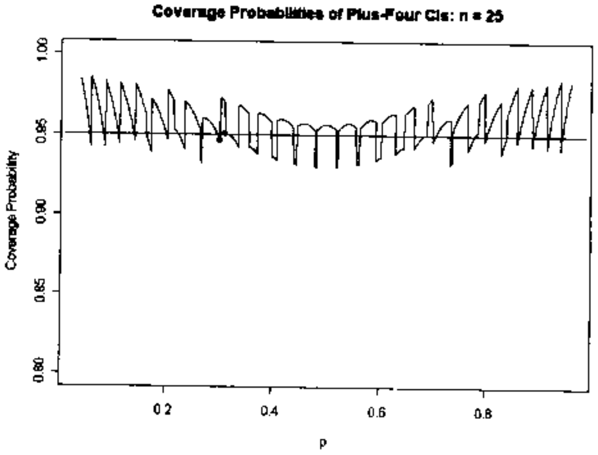

Figure 4. Confidence intervals based on the rule "add two Successes and two Failures." Two heavy dots show the coverage probabilities at p = 0.30 and 0.31 for this type of confidence interval. Coverage probabilities here are generally much closer to 95% than those in Figure 2.

coverage probabilities exceed 95% for many values of p and fall below 95% for relatively few values of p.

In particular, returning to our earlier examples, for samples of size 25, the coverage probability of the plus-four CI is 94.68% for $p = 0.30$ and 95.06% for $p = 0.31$. Both probabilities are remarkably close to 95% (see the heavy dots in Figure 4). Many values of p in the vicinity of 0.3 have larger coverage probabilities, and some have smaller coverage probabilities.

Lengths of Confidence Intervals

Of course, one can always improve coverage by making confidence intervals longer. At an absurd extreme, an all-purpose 100% CI for p—and a totally useless one—would be the interval (0, 1). So it is reasonable to ask how the average lengths of the plus-four CIs compare with the average lengths of the traditional ones. Has the increased coverage of the plus-four CIs come at the cost of an undue increase in their average length?

To show how the expected (or average) lengths are computed for a particular type of CI, we consider traditional CIs based on $n = 25$ subjects. Because the expected length depends on the value of p, we use $p = 0.30$ for an example, as we did in Table 1. Each value of $x = 0, 1, ..., 25$ yields its own CI, so we must view the length of a CI as a random variable L and compute $E(L)$.

Table 4, abbreviated to show only a few values of x, illustrates how to do this. The lower and upper confidence limits (LCL and UCL, respectively) are found, as in Table 1, for each value of x. If LCL falls below 0 or UCL falls above 1, then it is replaced by 0 or 1, respectively. Next, the length L is found by subtraction. Finally, the possible values of L are multiplied by their corresponding probabilities, and the 26 products are summed to give the expected length. For $p = 0.30$, the traditional CI has expected length 0.3498.

x	UCL	LCL	Length	Prob.	Product
0	0.0000	0.0000	0.0000	0.0001	0.0000
1	0.1168	0.0000	0.1168	0.0014	0.0002
2	0.1863	0.0000	0.1863	0.0074	0.0014
3	0.2474	0.0000	0.2474	0.0243	0.0060
4	0.3037	0.0163	0.2874	0.0572	0.0164
5	0.3568	0.0432	0.3136	0.1030	0.0323
6	0.4074	0.0726	0.3348	0.1472	0.0493
7	0.4560	0.1040	0.3520	0.1712	0.0603
8	0.5029	0.1371	0.3657	0.1651	0.0604
9	0.5482	0.1718	0.3763	0.1336	0.0503
10	0.5920	0.2080	0.3841	0.0916	0.0352
...					
25	1.0000	1.0000	0.0000	0.0000	0.0000
Sum of 26 products = E (Length) = 0.3498					

Table 4. Illustrating the Computation of the Average Length of a Traditional Confidence Interval for $n = 25$, $p = 0.30$.

For values of p in (0,1) and $n=25$, Figure 5 shows the average lengths of traditional and plus-four CIs. Our computation in Table 4 corresponds to one point on the curve for the traditional CIs.

What can we conclude from Figure 5? For extreme values of p, the plus-four CIs tend to be longer because the adjusted point estimates $\hat{p}_+$ are nearer 1/2 than are corresponding estimates $\hat{p}$. Recall that the maximum value of $p(1-p)$ occurs at $p=1/2$. For values of $\hat{p}$ near 1/2, the adjustment does not make much change in the point estimates, but it does have the effect of increasing n by 4, and so it decreases the margin of error and shortens the average CI a little. The plus-four adjustment appears to lengthen the CIs for values of p near 0 or 1 as necessary to achieve roughly 95% coverage and shorten them for values of p near 1/2 in a way that does no harm. Overall, it seems the adjustment used to make the 95% plus-four CIs has resulted in a reasonable tradeoff between coverage probability and length.

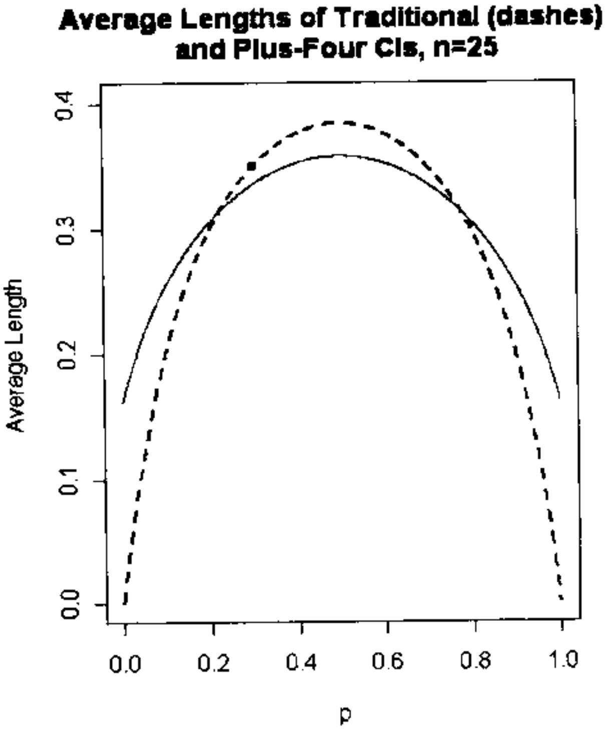

Figure 5. Comparing average lengths of traditional and plus-four confidence intervals. For p near 0 and 1, the plus-four intervals are longer and therefore have coverage probabilities nearer 95%. The heavy dot shows the average length of the traditional confidence interval for $p = 0.30$, as computed in Table 4.

Better, but Not Perfect

The papers by Agresti and Coull and by Brown, Cai, and Dasgupta have called widespread attention among professional statisticians to the bad behavior of the traditional CI for a small or moderate number of trials. These papers suggest a number of alternatives to the traditional CI, of which the plus-four CI is recommended as the simplest to explain and the easiest to compute. However, the plus-four adjustment is not a magical cure for every situation. One possible difficulty is at the 99% confidence level: When p starts to get near 0 or 1, plus-four CIs are surprisingly conservative, having very high coverage probabilities and unnecessarily long intervals. For example, see Figure 6, where $n = 50$. A general program in R for plotting coverage probabilities against p is available at *www.amstat.org/publications/stats/data.html* for those who wish to experiment with variations (types of CIs, values of n, or confidence levels) of the graphs shown in this article.

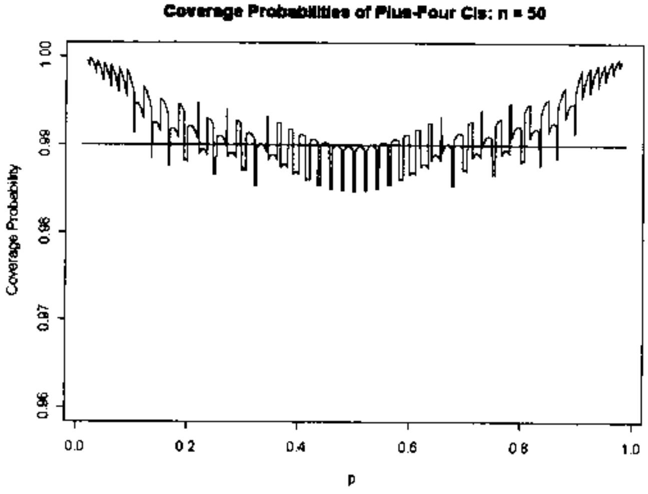

Figure 6. Illustrating the very conservative behavior of 99% plus-four confidence intervals for extreme values of p.

If the number of trials is several hundred or several thousand, as in many public opinion polls, the plus-four adjustment makes less difference. However, at the 95% level, it seems safe and easy just to use the plus-four interval, regardless of sample size. Recently, authors of some elementary texts (Moore and Devore, for example) have discussed and recommended plus-four CIs, especially when the number of trials is small.

There have been suspicions for some time that the traditional confidence interval for the binomial proportion might not perform well. So why has it taken until recently for statisticians to realize how bad it really is and seriously investigate alternatives? One can only speculate. However, graphs such as our Figures 2 and 3 carry a message that is instantly recognizable and almost impossible to ignore. These graphs require hundreds of thousands of computations. They would not have been made without modern statistical software or the imagination of those who figured out how to use such software to such striking effect. ■

Technical note:

The purpose of this note is to indicate a theoretical rationale for plus-four CIs. In the expression

$P\{-1.96 \le (\hat{p}-p)/\sigma \le 1.96\} = 0.95$

one can express σ in terms of p, square all three members of the inequality, and solve a quadratic equation to isolate p, obtaining a CI for p that depends on the normal approximation but does not approximate p by $\hat{p}$ to get the margin of error. The resulting point estimate of p is

$(X + \kappa^2/2)/(n+\kappa^2)$

and the margin of error is

$[\kappa/(n+\kappa^2)][n\hat{p}(1-\hat{p}) + \kappa^2/4]^{1/2}$

where $\kappa = 1.96$ for a 95% CI and is the value that cuts $1-\alpha/2$ from the upper tail of a standard normal distribution when an interval with confidence $1-\alpha$ is sought. This often is called the Wilson CI. The Wilson CI, with 2 instead of 1.96, is approximately the 95% plus-four CI. Accordingly, the plus-four adjustment works best for 95% CIs.

This article originated as a student project in a seminar class at California State University, East Bay, and is largely based on class notes and the first three references.

References

Agresti, A. and Coull, B.A. (1998). "Approximate Is Better than 'Exact' for Interval Estimation of Binomial Proportions." *The American Statistician,* 52:2, 119-126.

Brown, L.D., Cai, T.T., and Dasgupta, A.(2001). "Interval Estimation for a Binomial Proportion." *Statistical Science,* 16:2, 101-133.

Suess, E.A. and Trumbo, B.E. (in press). Chapter One. *Simulation and Estimation.* New York: Springer.

Moore, D.S. (2004). *The Basic Practice of Statistics* (3rd Ed.). New York: W.H. Freeman.

Devore, J. (2004). *Probability and Statistics for Engineering and the Sciences* (6th Ed.). New York: Duxbury.

Technical note:

The purpose of this note is to indicate a theoretical rationale for plus-four CIs. In the expression

$$P\{-1.96 \le (\hat{p}-p)/\sigma \le 1.96\} = 0.95,$$

one can express σ in terms of p, square all three members of the inequality, and solve a quadratic equation to isolate p, obtaining a CI for p that depends on the normal approximation but does not approximate p by $\hat{p}$ to get the margin of error. The resulting point estimate of p is

$$(X + \kappa^2/2)/(n+\kappa^2)$$

and the margin of error is

$$[\kappa/(n+\kappa^2)]\,[n\,\hat{p}\,(1-\hat{p}) + \kappa^2/4]^{1/2},$$

where $\kappa = 1.96$ for a 95% CI and is the value that cuts $1-\alpha/2$ from the upper tail of a standard normal distribution when an interval with confidence $1-\alpha^2$ is sought. This often is called the Wilson CI. The Wilson CI, with 2 instead of 1.96, is approximately the 95% plus-four CI. Accordingly, the plus-four adjustment works best for 95% CIs.