Journal of Modern Applied Statistical Methods

Volume 6 | Issue 1 Article 15

5-1-2007

Better Binomial Confidence Intervals

James F. Reed III
Lehigh Valley Hospital and Health Network

Journal of Modern Applied Statistical Methods
May, 2007, Vol. 6, No. 1, 153-161

1538 – 9472/07/$95.00

Better Binomial Confidence Intervals

James F. Reed III
Lehigh Valley Hospital & Health Network

The construction of a confidence interval for a binomial parameter is a basic analysis in statistical inference. Most introductory statistics textbook authors present the binomial confidence interval based on the asymptotic normality of the sample proportion and estimating the standard error - the Wald method. For the one sample binomial confidence interval the Clopper-Pearson exact method has been regarded as definitive as it eliminates both overshoot and zero width intervals. The Clopper-Pearson exact method is the most conservative and is unquestionably a better alternative to the Wald method. Other viable alternatives include Wilson's Score, the Agresti-Coull method, and the Borkowf SAIFS-z.

Key words: Binomial distribution, confidence intervals, coverage probability, Wald method, Clopper-Pearson Method, Score Method, Agresti-Coull method.

Introduction

The International Committee of Medical Journal editors indicated that confidence intervals are preferred over simple point estimates and p-values. This applies to over 300 international medical/scientific journals. Most introductory statistics textbook authors present the binomial confidence interval based on the asymptotic normality of the sample proportion and estimating the standard error. This approximate method is referred to as the Wald interval. In order to avoid approximation, some advanced statistics textbooks recommend the Clopper-Pearson exact binomial confidence interval. Other methods, asymptotic as well as exact, have been proposed and appear sporadically in introductory textbooks. There is a rather large set of articles, primarily in the statistics literature, about these and other less common methods of constructing binomial confidence intervals.

The purpose of this article is to provide a review of alternatives to the Wald method for computing a binomial confidence interval and provide a set of tractable and better methods of constructing binomial confidence intervals for a single proportion.

Methodology

When a binomial confidence interval is reported, the computational method is rarely given. This may imply that there is only one standard method for computing a binomial confidence interval - the Wald method (W). The W binomial confidence interval, either with or without a continuity correction, is found in every introductory statistics text. Typically, a warning or rule of thumb for determining when not to use W is included, but usually ignored. Occasionally, the Wald with a continuity correction (WCC) is included. For a single proportion the W and WCC lower bound (LB) and upper bound (UB) are defined as:

James Reed III is the Interim Chief of Health Studies and Director of Research at Lehigh Valley Hospital and Health Network. He has published over 100 journal articles and book chapters. His interests include applied statistical analyses, medical education, and statistical methods in simulation studies. Email: James_F.Reed@lvh.com

$$\text{W LB} = p - z_{\alpha/2}\sqrt{[pq/n]}$$
$$\text{W UB} = p + z_{\alpha/2}\sqrt{[pq/n]},$$
$$\text{WCC LB} = p - (z_{\alpha/2}\sqrt{[pq/n]}+1/(2n))$$
$$\text{WCC UB} = p + (z_{\alpha/2}\sqrt{[pq/n]}+1/(2n))$$

where $p = r/n$, $q = 1-p$, r=number of successes, and n is the total sample size.

Even though these two confidence interval methods are similar to large-sample formulas for means, both the W and WCC confidence intervals behave poorly in terms of zero width intervals and overshoot (Beal, 1987; Vollset, 1993; Newcombe, 1998; Pires, 2002; Rieczigel, 2003; Agresti, 2003). For instance, when r=0 or n, W and WCC have zero width or degenerate confidence intervals. Despite the known poor performance of the W and WCC confidence intervals, they continue to dominate in statistics textbooks, typically accompanied by warnings that when np is small, usually less than 5 or 10, exact or score methods should be used. A slightly different version of the rule of thumb requires that npq should be greater than or equal to 5. A better rule is to not compute confidence bounds for a proportion using the W method but rather to use one of the better methods. For small proportions the calculated lower bound can be below zero. Conversely, when a proportion approaches one, such as in the sensitivity and specificity of diagnostic or screening tests, and the upper bound may exceed one. This overshoot is avoided by truncating the interval to lie within [0, 1]. Overshoot and zero width confidence intervals may be avoided by a variety of better methods.

One of the standard measures of binomial confidence interval performance is the coverage probability, $C(\pi|n,\alpha)$. Given X=k,n, and α, let $\delta(\pi|k,n,\alpha)=1$ if $\pi \in [LB(k,n,\alpha), UB(k,n,\alpha)]$, and $\delta(\pi|k,n,\alpha)=0$ otherwise. Then, $C(\pi|n,\alpha)$ for a given π is:

$$C(\pi|n,\alpha)=\Sigma\, P(X=k|n,\pi)\ \delta(\pi|k,n,\alpha)$$

Figure 1 shows the 95% confidence interval coverage probability of the standard Wald methods {W, WCC} as a function of π, $\pi \in [0,1]$, for n=20. The coverage probability curves demonstrate the subnomial coverage for values of π near 0 and 1.

The Clopper-Pearson (CP) binomial confidence interval is the best-known exact method for interval estimation and is considered by most to be the gold standard (Clopper & Pearson, 1934). The CP confidence interval eliminates overshoot and zero width intervals and is strictly conservative. The CP lower and upper limits are defined by inverting the exact binomial tests with equal-tailed acceptance regions.

CP LB=0 if x=0, $(\alpha/2)^{1/n}$ if x=n.

$$LB=[1+(n-r+1)/(r\times F_{2r,\,2(n-r+1),\,1-\alpha/2})]^{-1}$$

CP UB=1-$(\alpha/2)^{1/n}$ if x=0, 1 if x=n.

$$UB=[1+(n-r)/(r\times F_{2(r+1),\,2(n-r),\alpha/2})]^{-1}$$

Fleiss (1981) preferred a more computationally intense binomial confidence interval with a continuity correction (SCC) attributed to Wilson (Wilson, 1927). For a single proportion, Wilson's Score (S) and Wilson's Score with continuity correction (SCC) LB and UB are defined as:

$$\text{S LB}=(2np+z^2-z\sqrt{\{z^2+4npq\}})/2(n+z^2)$$

$$\text{S UB}=(2np+z^2+z\sqrt{\{z^2+4npq\}})/2(n+z^2)$$

SCC LB =
$$[2np+z^2-1-z\sqrt{\{z^2-2-1/n+4p(nq+1)\}}]/(2n+2z^2)$$

SCC UB =
$$[2np+z^2+1+z\sqrt{\{z^2+2-1/n+4p(nq-1)\}}]/(2n+2z^2)$$

Blyth and Still (1983) investigated the performance of W, WCC, CP, Sterne's binomial confidence interval method (Sterne, 1954), and Pratt's (P) approximate confidence interval method (Pratt, 1968). Their results demonstrate the need for a continuity correction even when n is large. Blythe and Still then suggested a modification to W (WBS). While the WBS was an improvement over W and WCC, they concluded that it still was not

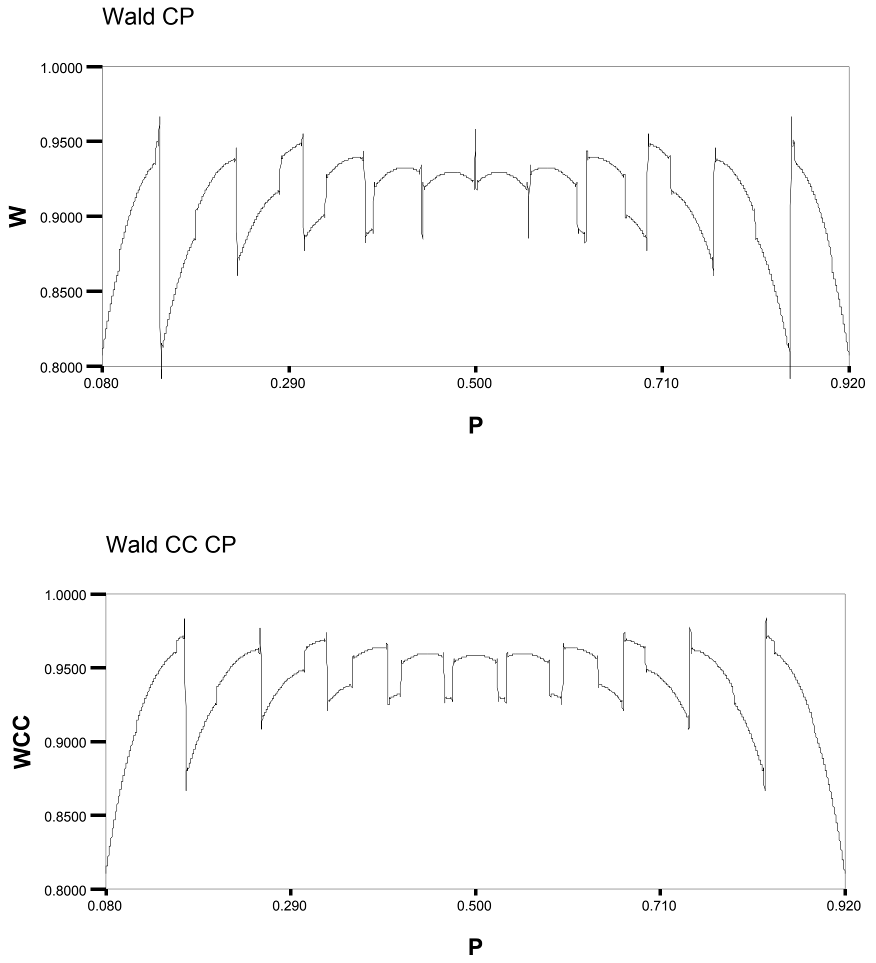


Figure 1. Coverage Probabilities (n=20) for the Wald and Wald CC Binomial Confidence Interval Methods.

satisfactory. The LB and UB for WBS are defined as:

$$LB = p - [z/\surd(n-z^2-2z/\surd n-1/n)][\surd(pq)+1/2n],$$

except LB=0 when r=0.

$$UB = p + [z/\surd(n-z^2-2z/\surd n-1/n)][\surd(pq)+1/2n],$$

except UB=1 for r=n.

Vollset (Vollset, 1993) compared thirteen methods for computing binomial confidence intervals using evaluative criteria of C(P), interval width, and errors relative to limits. Vollset proposed a mean Pratt (MP), a modification of P that is a closed form approximation to the mid-P exact interval. Define the UB of P as:

$$P\ UB=[1+(r+1)/(n-r))^2((A-b)/c)^3]^{-1},$$

with

$$A=81(r+1)(n-r)-9n-8,$$

$$B=3z\surd[9(r+1)(n-r)(9n+5z^2)+n+1],$$

and

$$C=81(r+1)2-9(r+1)(2+z^2)+1.$$

For P LB, replace r with r-1 and z with -z.

The Vollset MP lower and upper bound are then defined as:

$$\text{MP LB}=\{P_l(r)+P_l(r+1)\}/2,$$
$$\text{MP UB}=\{P_u(r)+P_u(r-1)\}/2$$

Vollset argued that W and WCC were unsatisfactory and the Clopper-Pearson, Pratt's approximation, SCC, MP, S and SCC are methods that may be safely used in all applications.

Newcombe (1998) compared seven methods for constructing two-sided binomial confidence intervals (W, WCC, S, SCC, Clopper-Pearson, mid-P and a likelihood-based method). The W and WCC were quickly judged as being inadequate, highly anti-conservative, asymmetrical in coverage, and incurred a higher risk of unacceptable boundary limits. Newcombe argued that neither W nor WCC should be acceptable methods for the scientific literature since other methods are tractable and all perform much better. Newcombe further argued that the use of the simple asymptotic standard error of a proportion should be restricted to sample size planning and introductory teaching purposes. Newcombe preferred three methods: the Clopper-Pearson method, the Score method and mid-P binomial based method.

Agresti and Coull, in noting the poor performance of the Wald interval and conservativeness of the Clopper-Pearson interval, proposed a straightforward adjustment - the add 4 to Wald. They suggested that by simply adding two successes and two failures and then use the Wald formula. Alternatively, one could add $z^2/2$ successes and $z^2/2$ failures before computing the Wald confidence interval. The latter is preferred. The Agresti-Coull adjusted Wald (AC) lower and upper bounds are:

$$LB=p'-z\sqrt{[p'q'/n']},$$
$$UB=p'+z\sqrt{[p'q'/n']}, \text{ where}$$
$$p'=(2r+z^2)/(2n+z^2), \text{ and } n'=n+z^2$$

Pires (2002) compared twelve methods for constructing confidence intervals for a binomial proportion and concluded that a clear classification of conservative methods included the Clopper-Pearson, the Score, and two arcsine transformation methods. A second tier of recommended confidence interval construction methods included a Bayesian method and the SCC.

Agresti (2003) argued for reducing the effects of discreteness in binomial confidence intervals by inverting two-sided tests rather than two one-sided tests. In most statistical practice, for interval estimation of a proportion or a difference or ratio of proportions, the inversion of the asymptotic score test is the best choice. If one wants to be a bit more conservative, mid-P adaptations or the Clopper-Pearson are recommended. For teaching purposes, the Wald-type interval plus and minus a normal-score multiple of a standard error is simplest.

Reiczigel compared four methods for constructing binomial confidence intervals: Wilson's Score, Agresti and Coull Adjusted Wald, the Clopper-Pearson, the mid-P, and Sterne's interval (Rieczigel, 2003). Unique to this study is the recommendation of using the Sterne interval and the Agresti-Coull adjusted Wald interval for binomial confidence intervals.

Tobi et al. (2005) compared the performance of seven approximate methods and the exact Copper-Pearson exact confidence intervals for small proportions. Three criteria were used to evaluate the performance of confidence intervals; coverage, confidence interval width, and aberrant confidence intervals. They concluded that: (1) one should compute confidence intervals for small proportions even when the number of events equals zero, (2) report what method has been used for confidence interval calculation, (3) the W method should be discarded, and (4) the Clopper-Pearson and the SCC are the best choices to calculate confidence intervals for small proportions.

Borkowf (2005) argued that even though the Agresti-Coull method binomial confidence intervals are substantially better than the Wald method, it can yield sub nominal coverage for some values of π for moderate sample sizes. A binomial confidence interval, which results in near nominal coverage and is easy to calculate by first augmenting the original data with a single imaginary failure to compute the lower confidence bound and a single imaginary success to compute the upper confidence bound is proposed - a single augmentation with an imaginary failure or success (SAIFS) method. The lower and upper SAIFS confidence bounds are then:

$$\text{SAIFS LB} = p_1 - \xi_{1-\alpha/2} \sqrt{[p_1 q_1/n]}$$

and

$$\text{UB} = p_2 + \xi_{1-\alpha/2} \sqrt{[p_2 q_2/n]},$$

with

$$p_1=(r + 0)/(n+1) \text{ and } p_2=(r+1)/(n+1)$$

Borkowf (2005) evaluated two forms of the SAIFS. The first uses the z-quantiles ($\xi_{1-\alpha/2}$) and the second used the t-quantiles ($\tau_{n-1,\ 1-\alpha/2}$). Compared to the Clopper-Pearson method, the SAIFS method using either the z or t quantiles results in confidence intervals with mean widths that are narrower for proportion parameters near 0 or 1 and whose coverage probabilities are marginally better over all values of π. The SAIFS-Z is preferred.

Figure 2 shows the 95% confidence interval coverage probability as a function of π, $\pi \in [0,1]$, for n=20 for CP, WBS, S, SCC, AC, and SAIFS-Z. Note that the sawtooth appearance of the coverage functions is due to the discontinuities for values of p corresponding to any lower or upper limits in the set of n+1 confidence intervals. The Clopper-Pearson and Borkowf SAIFS-z methods give at least nominal coverage for all values of $\pi \in [0,1]$, with severe over coverage near 0 and 1. The Score CC method gives at least nominal coverage for all values of $\pi \in [0,1]$ and avoids the over coverage of either the Clopper-Pearson or Score methods. The Score and Agresti-Coull methods yield nearly nominal coverage for all values of $\pi \in [0,1]$.

Conclusion

For the one sample binomial confidence interval, a new generation of introductory and medical statistics textbooks should emphasize the poor performance properties of W, WCC and include better binomial confidence methods. At least one from the set of Clopper-Pearson, S, SCC, Agresti-Coull, or the SAIFS-z methods should be mentioned. With the widespread use of laptop computers and access to computing resources on the internet, the complexity of computing binomial confidence intervals should not be an issue. The question remains as to which method to use. The Clopper-Pearson exact method has been regarded as definitive as it eliminates both overshoot and zero width intervals. The Clopper-Pearson exact method is the most conservative and is unquestionably a better alternative to the W when constructing and reporting binomial confidence intervals. In terms of programming ease, the Clopper-Pearson is easily programmed as are the Blythe & Still, Wilson's Score, Score with a continuity correction, the Agresti-Coull method, and the Borkowf SAIFS-z.

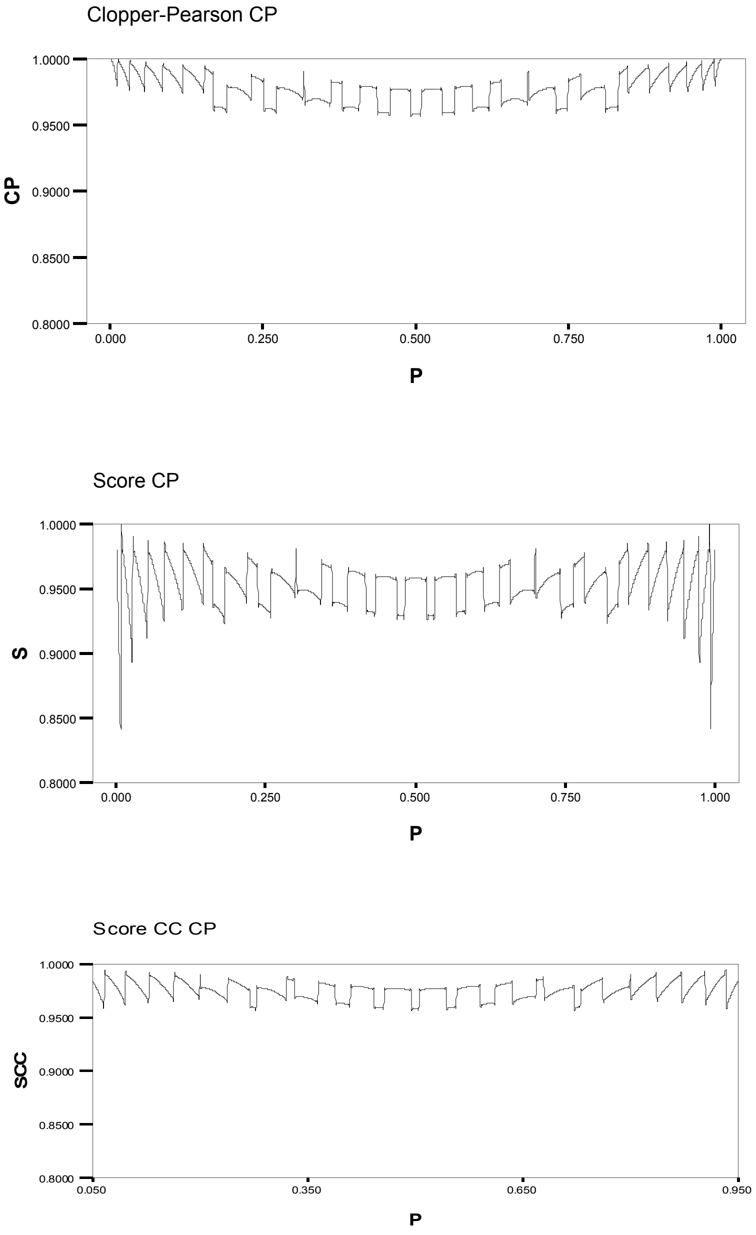
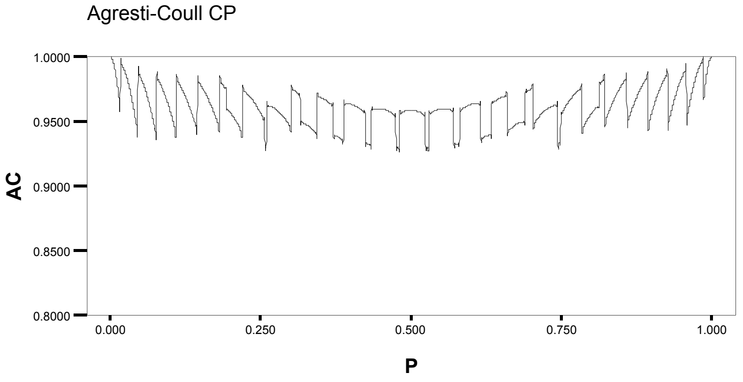


Figure 2. Coverage Probabilities (n=20) for the Clopper-Pearson, Score, Score CC, Agresti-Coull, and Borkowf SAIFS-z Binomial Confidence Interval Methods.



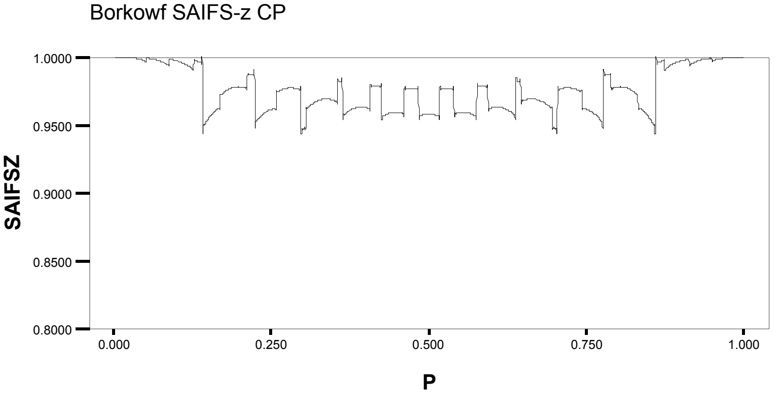


Figure 2 (Continued). Coverage Probabilities (n=20) for the Clopper-Pearson, Score, Score CC, Agresti-Coull, and Borkowf SAIFS-z Binomial Confidence Interval Methods.

Table 1. Methods for Calculation of Confidence Intervals for a Single Proportion

Method		Formula
Clopper-Pearson	CP	$LB=0$ if $x=0$, $(\alpha/2)^{1/n}$ if $x=n$.
		$LB=[1+(n-r+1)/(r \times F_{2r,\ 2(n-r+1),\ 1-\alpha/2})]^{-1}$
		$UB=1-(\alpha/2)^{1/n}$ if $x=0$, 1 if $x=n$.
		$UB=[1+(n-r)/(r \times F_{2(r+1),\ 2(n-r),\alpha/2})]^{-1}$
Score (Wilson)	S	$LB=(2np+z^2-z\surd\{z^2+4npq\})/2(n+z^2)$
		$UB=(2np+z^2+z\surd\{z^2+4npq\})/2(n+z^2)$
Score (w/CC)	SCC	$LB=[2np+z^2-1-z\surd\{z^2-2-1/n+4p(nq+1)\}]/(2n+2z^2)$
		$UB=[2np+z^2+1+z\surd\{z^2+2-1/n+4p(nq-1)\}]/(2n+2z^2)$
Agresti-Coull	AC	$LB=p'-z\surd[p'q'/n']$
		$UB=p'+z\surd[p'q'/n']$, where
		$p'=(2r+z^2)/(2n+z^2)$, and $n'=n+z^2$.
Borkowf	SAIFS	$LB = p_1 - \xi_{1-\alpha/2} \surd[p_1q_1/n]$
		$UB = p_2 + \xi_{1-\alpha/2} \surd[p_2q_2/n]$, with
		$p_1=(r + 0)/(n+1)$ and $p_2=(r+1)/(n+1)$, where
		$\xi_{1-\alpha/2}$ are z-quantiles or $\tau_{n-1,\ 1-\alpha/2}$ the t-quantiles

References

Agresti A. & Coull B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. The *American Statistician, 52*, 119-126.

Agresti, A. & Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions. *Biometrics*, *57*, 963-71.

Agresti, A. (2003). Dealing with discreteness: Making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Statistical Methods Medical Research, 12*, 3-21.

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

Bonett, D. G. & Price, R. M. (2005). Confidence intervals for a ratio of binomial proportions based on paired data. *Statistical Methods Medical Research,15.*

Borkowf, C. B. (2005). Constructing binomial confidence intervals with near nominal coverage by adding a single imaginary failure or success. *Statistical Methods Medical Research*, *25*.

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

Fleiss, J. H. (1981). *Statistical methods for rates and proportions* (2nd Ed.). New York: John Wiley & Sons.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistical Methods Medical Research, 17*, 857-72.

Pires, A. M. (2002). *Confidence intervals for a binomial proportion: Comparison of methods and software evaluation*. Proceedings of the Conference ComStat 2002. http://www.math.ist.utl.pt/~apires.

Pratt, J. W. (1968). A normal approximation for binomial, F, Beta, and other common, related tail probabilities. *Journal of the American Statistical Association, 63*, 1457-1483.

Radhakrishna, S., Murthy, B. N., Nair, N. G. K., Jayabal, P., & Jayasri, R. (1992). Confidence intervals in medical research. *Indian Journal of Medical Research [B], 96*, 199-205.

Reiczigel, J. (2003). Confidence intervals for the binomial parameter: Some new considerations. *Statistical Methods Medical Research, 22*, 611-21.

Sterne, T. E. (1954). Some remarks on confidence or fiducial limits'. *Biometrika, 41*, 275-278.

Tobi, H., van den Berg, P. B., & deJong-van den Berg, L. T. W. (2005). Small proportions: What to report for confidence intervals. *Pharmacoepidemiology and Drug Safety, 14*, 239-247.

Vollset, S. E. (1993). Confidence intervals for a binomial proportion. *Statistical Methods Medical Research, 12*, 809-24.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22,* 209-212.