

Confidence Interval Calculation for Binomial Proportions

Keith Dunnigan
Statking Consulting, Inc.

Introduction:

One of the most fundamental and common calculations in statistics is the estimation of a population proportion and its confidence interval (CI). Estimating the proportion of successes in a population is simple and involves only calculating the ratio of successes to the sample size.

The most common method for calculating the confidence interval is sometimes called the Wald method, and is presented in nearly all statistics textbooks. It is so widely accepted and applied, that for many it is the only method they have used. For most others it is the technique of first choice. Careful study however reveals that it is flawed and inaccurate for a large range of n and p , to such a degree that it is ill-advised as a general method^{1,2}. Because of this many statisticians have reverted to the exact Clopper-Pearson method, which is based on the exact binomial distribution, and not a large sample normal approximation (as is the Wald method). Studies have shown however that this confidence interval is very conservative, having coverage levels as high as 99% for a 95% CI, and requiring significantly larger sample sizes for the same level of precision^{1,2,3}. An alternate method, called the Wilson Score method is often suggested as a compromise. It has been shown to be accurate for most parameter values and does not suffer from being over-conservative, having coverage levels closer to the nominal level of 95% for a 95% CI.

In this discussion a brief review of the Wald, Wilson-Score, and exact Clopper Pearson methods of calculating confidence intervals for binomial proportions will be presented, focusing on differences between the Wald and Wilson Score methods. Sample size calculations for the Wald and Wilson Score methods will also be discussed. SAS programs for these formulas will be also presented and applied to a worked out example, which can be readily modified for other data. Finally the differences between the methods will be discussed in general.

Background – Confidence Interval Calculation

Binomial Distributed Random Variables

In standard statistical methodology, a Bernoulli random variable X_i is defined to have two possible values: Success ($X_i = 1$, with probability p), and Failure ($X_i = 0$, with probability $q = 1 - p$). From this, the mean and variance of a Bernoulli random variable may be calculated:

$$\mu = E(X_i) = (1)(p) + (0)(q) = p$$

$$E(X_i^2) = (1^2)(p) + (0^2)q = p$$

$$\text{Var}(X_i) = E(X_i^2) - [E(X_i)]^2 = pq$$

A binomial random variable X is defined as the sum of n independent Bernoulli random variables (X_1, \dots, X_n) . From this the mean and variance are easily obtained as np and npq . The probability of each value x of a Binomial distributed random variable X is defined through its probability mass function:

$$X \sim \text{Bin}(n, p) \Leftrightarrow P\{X = x | n, p\} = \binom{n}{x} p^x q^{n-x}$$

Wald and Wilson Score Confidence Interval Formulas

The Wald, Wilson Score, and Clopper-Pearson methods of calculating CI's all assume that the variable of interest (the number of successes) can be modeled as a Binomial random variable. The difference between the first two methods can be seen most easily by examining the difference in the derivations^{4,5}. The derivation of the Wald and Wilson Score confidence intervals begin similarly:

Since the binomial is the sum of n independent Bernoulli random variables, for large values of n , the central limit theorem is valid and X has approximately a normal distribution. The estimator for the population proportion \hat{p} is equal to X/n , and since it differs from X only by a constant, also has approximately a normal distribution. The mean and variance of \hat{p} are easily obtained:

$$E[X/n] = (1/n) E[X] = np/n = p$$

$$V[X/n] = (1/n)^2 V[X] = npq/n^2 = pq/n$$

Subtracting off the mean and standard deviation from \hat{p} then gives a standard normal random variable and the following equation can be used to derive the endpoints of a 95% confidence interval:

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{pq/n}} < z_{\alpha/2}\right) = 1 - \alpha \quad (1)$$

The endpoints can be derived by taking the left side of equation 1 and solving it for \hat{p} after replacing the $<$ signs with equals signs:

$$\frac{\hat{p} - p}{\sqrt{pq/n}} = \pm z_{\alpha/2} \quad (2)$$

At this point the Wald and Wilson Score methods diverge. The traditional Wald method completes the algebra to the following step before making an approximation:

$$p = \hat{p} \pm z_{\alpha/2} \sqrt{p(1-p)/n} \quad (3)$$

At this point the Wald method replaces the population values p and q in the right side of the equation with their approximations \hat{p} and \hat{q} to obtain the traditional Wald confidence interval formula for a proportion:

$$p = \hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}\hat{q}/n} \quad (4)$$

The Wilson Score method does not make the approximation in equation 3. The result is more involved algebra (which involves solving a quadratic equation), and a more complicated solution. The result is the Wilson Score confidence interval for a proportion:

$$p = \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}} \quad (5)$$

Clopper Pearson Exact Confidence Interval Formula

The formula for the Clopper Pearson confidence interval is shown below⁶. It is also commonly shown in several other algebraically identical forms^{1,3,4}.

$$\frac{1}{1 + \frac{n-x+1}{x} F_{2(n-x+1), 2x, \alpha/2}} \leq p \leq \frac{\frac{x+1}{n-x} F_{2(x+1), 2(n-x), \alpha/2}}{1 + \frac{x+1}{n-x} F_{2(x+1), 2(n-x), \alpha/2}}$$

Sample Size Formulas for the Wald and Wilson Score Methods

The sample size formula for the Wald method may be obtained straightforwardly from equation 4. If we define the precision as one half the length of the confidence interval, then the sample size required to obtain a precision d is:

$$n = \frac{(z_{\alpha/2}^2) \hat{p} \hat{q}}{d^2}$$

Where you would replace the random values \hat{p} and \hat{q} by the assumed constant values p_0 and q_0 .

The sample size required to produce a Wilson Score confidence interval with a lower confidence limit of L , may be derived by using equation 5, setting $p = L$, and solving for n . (Again, \hat{p} and \hat{q} are also replaced by the constant values p_0 and q_0). After some simplification this gives:

$$\sqrt{4np_0q_0 + z_{\alpha/2}^2} = \frac{2(p_0 - L)}{z_{\alpha/2}}n + (1 - 2L)z_{\alpha/2}$$

$$\frac{4(p_0 - L)^2}{z_{\alpha/2}^2}n^2 + (4(p_0 - L)(1 - 2L) - 4p_0q_0)n + ((1 - 2L)^2 - 1)z_{\alpha/2}^2 = 0 \quad (6)$$

Equation 6 is solved using the quadratic formula, where it turns out that only one of the roots is positive. The final equation then becomes:

$$n = z_{\alpha/2}^2 \left[\frac{-((p_0 - L)(1 - 2L) - p_0q_0) + \sqrt{[(p_0 - L)(1 - 2L) - p_0q_0]^2 - (p_0 - L)^2((1 - 2L)^2 - 1)}}{2(p_0 - L)^2} \right] \quad (7)$$

Sample Size for the Exact Clopper Pearson Method

Some sample size tables have been calculated for the Clopper Pearson Exact Confidence interval and are available in the literature⁴.

SAS Example 1 – Confidence Interval Calculation

The SAS code for calculating the confidence interval for one proportion will now be illustrated for the Wald, Wilson Score, and Exact methods by presenting a worked out example.

In this example a new xray imaging method is to be evaluated in a clinical study for its effectiveness in detecting the presence or absence of a specific disease state. It is evaluated by a separate procedure also, which is deemed the ‘gold standard’ and is assumed perfect. Through previous clinical experience as well as pre-clinical

experimentation, it is believed and assumed that the success rate of the new procedure in detecting the presence or absence of this disease is > 90% and that the point estimate obtained in the study will be at least 90%. The current standard of care in imaging technology, and the best that is commercially available is an 80% success rate. We wish to select a sufficient sample size for the study in order that the low end of a one-sided confidence interval for p will be at least 80%.

In this example we work backwards by trial and error to find the appropriate sample size by constructing confidence intervals for $\hat{p} = 0.90$ and various values of n. The SAS code used is given in the appendix, where a data step is used outside the macro to input various values of n and \hat{p} . (This is because it is desired to view a number of related confidence intervals on the same output). The entries after the cards statement can easily be modified to calculate 1 or 2 sided confidence intervals for any set parameters (alpha, n, and \hat{p}). The SAS output produced is shown in Table 1.

The un-symmetric nature of the Score and Exact confidence intervals is illustrated in this example. The symmetric nature of the Wald confidence interval may lead to upper limits over 100% or lower limits under 0, which is seen here for n=24.

The conservative hierarchy of the confidence intervals (in this range of p) can be seen in this example. From Table 1 we see that in order to insure a lower confidence limit over 80%, we need to select only 25 subjects using the Wald interval. The Wilson Score interval requires an additional 19 subjects (44 total) for a sample size increase of 76%. Finally the Clopper-Pearson exact confidence interval requires an additional 6 subjects over the Wilson Score (50 total), which is an increase of 13.6%.

Table 1: SAS Output for Example 1

Confidence Intervals for a Single Proportion									
n	p	alpha	Sides on CI	LCL (Wald)	UCL (Wald)	LCL (Wilson Score)	UCL (Wilson Score)	LCL (Exact)	UCL (Exact)
24	0.9	0.05	1	0.79927	1.00073	0.75574	0.96321	0.73879	0.97764
25	0.9	0.05	1	0.80131	0.99869	0.75938	0.96250	0.74321	0.97665
26	0.9	0.05	1	0.80323	0.99677	0.76278	0.96182	0.74733	0.97570
29	0.9	0.05	1	0.80837	0.99163	0.77182	0.95992	0.75820	0.97302
32	0.9	0.05	1	0.81277	0.98723	0.77944	0.95820	0.76728	0.97059
35	0.9	0.05	1	0.81659	0.98341	0.78597	0.95663	0.77500	0.96839
38	0.9	0.05	1	0.81995	0.98005	0.79163	0.95520	0.78165	0.96637
43	0.9	0.05	1	0.82475	0.97525	0.79959	0.95306	0.79093	0.96336
44	0.9	0.05	1	0.82561	0.97439	0.80100	0.95266	0.79256	0.96281
45	0.9	0.05	1	0.82644	0.97356	0.80236	0.95227	0.79413	0.96227
45	0.9	0.05	1	0.82644	0.97356	0.80236	0.95227	0.79413	0.96227
48	0.9	0.05	1	0.82878	0.97122	0.80615	0.95117	0.79851	0.96073
49	0.9	0.05	1	0.82951	0.97049	0.80732	0.95082	0.79986	0.96024
50	0.9	0.05	1	0.83021	0.96979	0.80846	0.95047	0.80117	0.95976

Coverage Probabilities

Example 1 illustrated the conservative nature of the three confidence interval methods under discussion. The question is, is the change in confidence interval method from Wald, to Wilson Score, to Exact worth the often substantial increase in sample size? The answer to this question is provided by examining the coverage probabilities for each of these methods. Coverage probabilities for these intervals (among others), were examined by Stein Vollset² as a function of p , for sample sizes of 10, 100, and 1000. These graphs are reproduced below for sample sizes $n=10$ and $n=100$, where 'W' denotes Wald, 'S' denotes Wilson Score, and 'MAX' denotes the exact Clopper Pearson.

Figure 1: Coverage Probabilities for $n=10$
(Source: Vollset 1993)

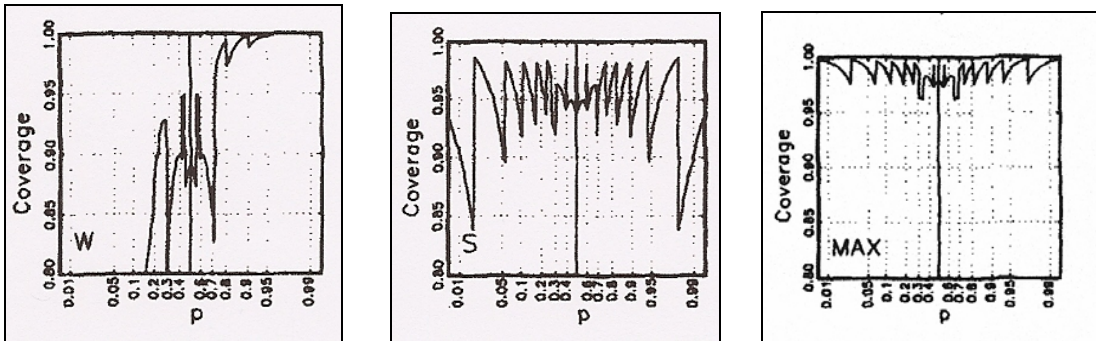
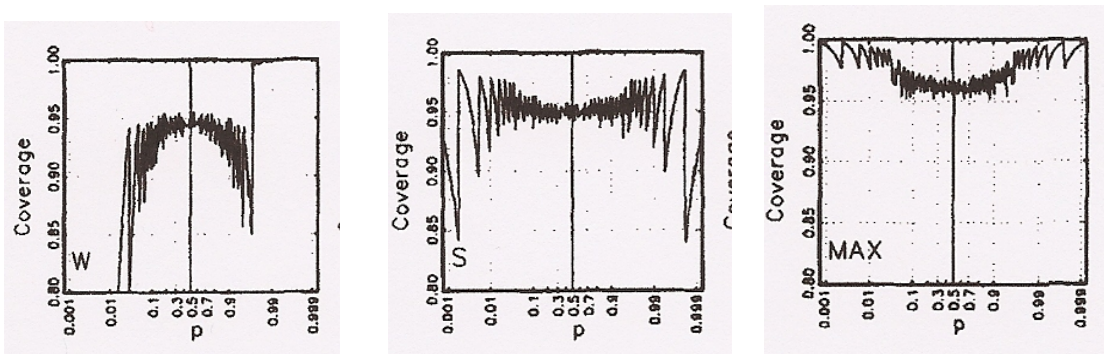


Figure 2: Coverage Probabilities for $n=100$
(Source: Vollset 1993)



Comparing the Wald and Wilson score coverage probabilities, the increase in sample size is justified. It can be seen from Figures 1 and 2, that the coverage probability drops substantially for the Wald interval as the proportion varies up or down from 50%, but reverses course and becomes hyper-conservative at extreme values. The smaller sample sizes for such values are deceptive: One is not really estimating a 95% confidence

interval, as the coverage probabilities either drop substantially below 90% or increase to over 99%.

Comparing the Wilson Score to the Exact method, the sample size increase is not justified. In this range of parameter values ($p=90\%$, $n=44-50$) we can see that the coverage probabilities of the exact interval are very conservative and exceed those of the Wilson Score, often by as much as 3 to 4 percent. The exact confidence interval here is closer to a 98% or 99% confidence interval than a 95% confidence interval. On the other hand the Wilson Score confidence interval may have a coverage level as low as 94 or 93% for some values of p (this involves some interpolation as a plot for $n=44$ is not available). If some tolerance in the coverage level can be tolerated, the Wilson Score interval is the method of choice.

SAS Example 2 – Sample Size Calculation

In this example SAS is used to calculate the sample size for the Wald and Wilson Score confidence intervals. The SAS code computes the sample size based on equations 5 and 7, and is included in the appendix. The SAS output is shown below in Table 2, where we can see that the sample size estimates, after being rounded up, match those obtained in example 1 (25 and 44).

Table 2: SAS Output for Example 2

Sample Size Calculation for a CI of One Proportion					
alpha	p	Precision (Half CI Width) (Wald)	Lower Confidence Limit (Wilson Score)	Sample Size (Wald)	Sample Size (Wilson Score)
0.05	0.9	0.1	0.8	24.3499	43.2887

Summary

From the Vollset plots it can be seen that for small values of n (<100), the Wald confidence interval has a coverage probability that may be alternately very liberal or very conservative. For higher values of n and intermediate values of p ($10\% < p < 90\%$), the Wald interval may be satisfactory. It should be avoided for general use.

The Wilson Score confidence interval is recommended for general use. The question is, when should one consider using the exact method instead? In examining the Vollset plots, some good rules of thumb suggest themselves:

- 1) For small ranges of n (15-25), and if some tolerance can be accepted on the coverage (1% or 2%), consider using the exact method only if $p > 90\%$ or $<10\%$.

- 2) For intermediate ranges of n (≈ 100), and if some tolerance can be accepted on the coverage (1% or 2%), consider using the exact method only if $p > 99\%$ or $< 1\%$.
- 3) If no tolerance can be accepted on the coverage and sample size conservation is paramount, consider using the Wilson Score with a smaller alpha level (e.g., $\alpha = 0.04$ for a 95% CI), and compare the resulting sample sizes for both methods.

The exact method is recommended for the case of very rare or common events. As event probabilities drop increasingly below 1% or above 99%, the exact method should be strongly considered. This is especially the case with extremely rare or common events. A good example might be in cell biology or chemistry, where one may deal with concentration levels in the parts per million. In such situations, the exact method should be the method of choice.

Finally the exact method should be considered in situations where there is no tolerance in the coverage level. For instance, if a 95%CI must not have *any* possibility of being below 95% in coverage. The exact method is guaranteed to never fall below the nominal level. The cost of this is its highly conservative nature and inflated sample size. However, another possibility in this situation is to use the Wilson Score method with a lower level of alpha.

References

- 1) Agresti, Alan and Brent Coull. "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions" *The American Statistician*, May 1998 Vol. 52, No. 2, pp 119-126.
- 2) Stein Vollset. "Confidence Intervals For a Binomial Proportion" *Statistics in Medicine*, Vol 12, pp 809-824, 1993.
- 3) Sauro, Jeff and James Lewis. "Estimating Completion Rates from Small Samples using Binomial Confidence Intervals" *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting – 2005*. pp 2100-2104.
- 4) K. Krishnamoorthy and Jie Peng. "Some Properties of the Exact and Score Methods for Binomial Proportion and Sample Size Calculation" *Communications in Statistics – Simulation and Computation*, 36: pp 1171-1186, 2007.
- 5) Devore, Jay. *Probability and Statistics for Engineering and the Sciences*, 3rd Edition, 1991 Brooks/Cole Publishing Company, A Division of Wadsworth, Belmont CA, pp 268, 269.
- 6) Casella, George and Roger Berger. *Statistical Inference* 1990 Brooks Cole Publishing Company, A Division of Wadsworth, Belmont CA, pp 449.

Appendix – SAS Code

Example 1:

```
/******  
** Program: Propci.sas  
** Author: Keith Dunnigan  
**  
** Purpose: Calculate the confidence interval for one proportion.  
**  
** Description: Calculate Wald, Wilson Score, and exact Clopper-Pearson  
**              confidence intervals for a variety of n, p, and alpha  
**              combinations for either a one or two-sided interval.  
**  
*****/  
  
%macro runprog;  
  
data propci;  
  set parms;  
  if sided = 1 then do;  
    zalpha=probit(1-(alpha));  
  end;  
  else if sided = 2 then do;  
    zalpha=probit(1-(alpha/2));  
  end;  
  
** Wald;  
  
  q = 1 - p;  
  stder = sqrt(p*q/n);  
  Wald_lcl = p - zalpha * stder;  
  Wald_ucl = p + zalpha * stder;  
  
** Wilson Score;  
  
  part1 = p + ((zalpha**2)/(2*n));  
  part2 = sqrt( (p*q/n) + ((zalpha**2)/(4*n**2)) );  
  part3 = 1 + (zalpha**2)/n;  
  Wilson_lcl = (part1 - (zalpha * part2))/ part3;  
  Wilson_ucl = (part1 + (zalpha * part2))/ part3;  
  
** Exact Clopper Pearson;  
  
  x = round (n*p,0.1);  
  
  * Calculate the lower limit.;  
  
  v1 = 2*(n-x+1);  
  v2 = 2*x;  
  
  if sided = 1 then do;  
    a = 1-(alpha);  
  end;  
  else if sided = 2 then do;  
    a = 1-(alpha/2);
```

```

end;

coef = (n-x+1)/x;
fscore = finv(a,v1,v2);
exact_lcl = 1/(1+coef*fscore);

* Calculate the upper limit.;

v11 = 2*(x+1);
v22 = 2*(n-x);
fscore2 = finv(a,v11,v22);
coef2 = (x+1)/(n-x);
numer = coef2*fscore2;
exact_ucl = numer/(1+numer);
run;

options nodate;
title 'Confidence Intervals for a Single Proportion';
proc print data = propci split = '_' noobs;
  var n p alpha sided Wald_lcl Wald_ucl Wilson_lcl Wilson_ucl exact_lcl exact_ucl;
  label wald_lcl = 'LCL_(Wald)'
        wald_ucl = 'UCL_(Wald)'
        wilson_lcl = 'LCL_(Wilson_Score)'
        wilson_ucl = 'UCL_(Wilson_Score)'
        sided = 'Sides_on_CI'
        exact_lcl = 'LCL_(Exact)'
        exact_ucl = 'UCL_(Exact)';
run;

%mend runprog;

data parms;
infile cards;
input n p alpha sided;
cards;
24 0.9 0.05 1
25 0.9 0.05 1
26 0.9 0.05 1
29 0.9 0.05 1
32 0.9 0.05 1
35 0.9 0.05 1
38 0.9 0.05 1
43 0.9 0.05 1
44 0.9 0.05 1
45 0.9 0.05 1
45 0.9 0.05 1
48 0.9 0.05 1
49 0.9 0.05 1
50 0.9 0.05 1
run;

%runprog;

```

Example 2:

```

/*****
** Program: Propci_SamSize.sas
** Author: Keith Dunnigan
**
** Purpose: Calculate the sample size to produce a Wald confidence
**          interval for one proportion of specified precision d
**          for a variety of d, p, and alpha combinations for either
**          a one or two-sided interval.
**
**          Calculate the sample size to produce a Wilson Score Confidence
**          interval for one proportion of specified lower confidence
**          limit L for a variety of L, p, and alpha combinations for
**          either a one or two-sided interval.
**
*****/

data propsamsiz;
  infile cards;
  input d p alpha sided L;
  width = 2 * d;

  if sided = 1 then do;
    zalpha=probit(1-(alpha));
  end;
  else if sided = 2 then do;
    zalpha=probit(1-(alpha/2));
  end;

  ** Wald;

  q = 1 - p;
  n1 = (zalpha**2 * p * q) / d**2;

  ** Wilson Score;

  part1b = (p-L)*(1-2*L) - (p*q) ;
  part2b = ((p-L)**2) * ( (1-(2*L))**2 - 1);
  part3b = 2*((p-L)**2);
  n2 = (zalpha**2) * ( (-1*part1b + sqrt ( (part1b**2) - part2b) ) / part3b );

cards;
0.1 0.9 0.05 1 0.80
run;

options nodate;
title "Sample Size Calculation for a CI of One Proportion";
proc print data = propsamsiz split = '_' noobs;
  var alpha p d L n1 n2;
  label d = 'Precision_(Half CI Width)_(Wald)'
        L = 'Lower_Confidence Limit_(Wilson Score)'
        n1 = 'Sample_Size_(Wald)'
        n2 = 'Sample_Size_(Wilson Score)';
run;
```