

# Constructing Confidence Intervals for the Differences of Binomial Proportions in SAS®

Will Garner, Gilead Sciences, Inc., Foster City, CA

## ABSTRACT

Given two independent binomial proportions, we wish to construct a confidence interval for the difference. The most widely known method is the Wald method (ie, normal approximation), but it can produce undesirable results in extreme cases (eg, when the proportions are near 0 or 1). Numerous other methods exist, broadly within two groups: approximate methods and exact methods. This paper presents 20 different methods for constructing such confidence intervals, 10 of which are available in SAS® 9.4 procedures, while 6 others can be computed in SAS® following formulas (or using PROC IML). The methods are compared by theoretical examples as well as a case study. Based on that, thoughts are given on which method to use.

## BACKGROUND (11 METHODS)

Suppose we have  $n_1$  subjects with  $x_1$  'successes' and  $n_2$  subjects with  $x_2$  'successes'. We can define the observed proportion of successes for the two groups as  $\hat{p}_1 = x_1/n_1$  and  $\hat{p}_2 = x_2/n_2$ , respectively. Let  $z$  correspond to the  $\alpha/2$ -percentile of the normal distribution. Thus, for a 95%CI (ie,  $\alpha=0.05$ ), the corresponding  $z$  value would be 1.96.

Newcombe (1998) provides 11 methods for constructing confidence intervals for the difference of two independent binomial proportions. They are as follows:

1. Wald method without continuity correction

$$(\hat{p}_1 - \hat{p}_2) \pm z \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

2. Wald method with continuity correction

$$(\hat{p}_1 - \hat{p}_2) \pm \left( \frac{1}{2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right) + z \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \right)$$

3. Haldane

The limits are  $\theta^* \pm w$  where

$$\theta^* = \frac{(\hat{p}_1 - \hat{p}_2) + z^2 v(1 - 2\hat{\psi})}{1 + z^2 u}$$

where

$$w = \frac{z}{1 + z^2 u} \sqrt{u\{4\hat{\psi}(1 - \hat{\psi}) - (\hat{p}_1 - \hat{p}_2)^2\} + 2v(1 - 2\hat{\psi})(\hat{p}_1 - \hat{p}_2) + 4z^2 u^2(1 - \hat{\psi})\hat{\psi} + z^2 v^2(1 - 2\hat{\psi})^2}$$

$$\hat{\psi} = (\hat{p}_1 + \hat{p}_2)/2, u = (1/n_1 + 1/n_2)/4, \text{ and } v = (1/n_1 - 1/n_2)/4$$

4. Jeffreys-Perks

This is the same as Haldane, but with

$$\hat{\psi} = \frac{1}{2} \left( \frac{x_1 + 1/2}{n_1 + 1} + \frac{x_2 + 1/2}{n_2 + 1} \right)$$

## 5. Mee

$$(\hat{p}_1 - \hat{p}_2) \pm z \sqrt{\lambda \left( \frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2} \right)}$$

where  $\tilde{p}_1$  and  $\tilde{p}_2$  are the maximum likelihood estimators for  $p_1$  and  $p_2$  under the null hypothesis that  $p_1 - p_2 = \delta$  and  $\lambda = 1$ . Miettinen and Nurminen provide a closed-form solution for the maximum likelihood given by the following:

$$\tilde{p}_1 = 2u\cos(w) - b/3a \quad \text{and} \quad \tilde{p}_2 = \tilde{p}_1 + \delta$$

where

$$\begin{aligned} w &= \frac{\pi + \cos^{-1}(v/u^3)}{3} \\ v &= \frac{b^3}{27a^3} - \frac{bc}{6a^2} + \frac{d}{2a} \\ u &= \frac{v}{|v|} \sqrt{\frac{b^2}{9a^2} + \frac{c}{3a}} \\ a &= 1 + \theta \\ b &= -(1 + \theta + \hat{p}_1 + \theta\hat{p}_2 + \delta(\theta + 2)) \\ c &= \delta^2 + \delta(2\hat{p}_1 + \theta + 1) + \hat{p}_1 + \theta\hat{p}_2 \\ d &= -\hat{p}_1\delta(1 + \delta) \\ \theta &= n_2/n_1 \end{aligned}$$

## 6. Miettinen-Nurminen

The interval is the same as Mee, but with  $\lambda = (n_1 + n_2)/(n_1 + n_2 - 1)$ .

## 7. True profile

The interval consists of all  $\delta$  satisfying

$$x_1 \ln \frac{\tilde{p}_1}{\hat{p}_1} + x_2 \ln \frac{\tilde{p}_2}{\hat{p}_2} + (n_1 - x_1) \ln \frac{(1 - \tilde{p}_1)}{(1 - \hat{p}_1)} + (n_2 - x_2) \ln \frac{(1 - \tilde{p}_2)}{(1 - \hat{p}_2)} \geq -\frac{z^2}{2}$$

8. Profile likelihood method based on 'exact' tail areas
9. Profile likelihood method based on 'mid-p' tail areas

Same as the 'exact' method (Method 8), but with  $k = 0.5$ .

## 10. Newcombe (Score) without continuity correction

The Wilson (Score) confidence limits without continuity correction for each individual binomial proportion,  $p_i = x_i/n_i$ , ( $i = 1, 2$ ) is given by

$$\frac{1}{2(n_i + z^2)} \left( (2n_i\hat{p}_i + z^2) \pm z\sqrt{4n_i\hat{p}_i(1 - \hat{p}_i) + z^2} \right)$$

Denote the lower and upper Wilson (Score) confidence limits for  $p_i$  as  $L_i$  and  $U_i$ . The Newcombe (Score) confidence limits are given by

$$\begin{aligned} \text{Lower limit: } & (\hat{p}_1 - \hat{p}_2) - \sqrt{(\hat{p}_1 - L_1)^2 + (U_2 - \hat{p}_2)^2} \\ \text{Upper limit: } & (\hat{p}_1 - \hat{p}_2) + \sqrt{(U_1 - \hat{p}_1)^2 + (\hat{p}_2 - L_2)^2} \end{aligned}$$

## 11. Newcombe (Score) with continuity correction

The method is the same as the Score method (Method 10) above, but the confidence intervals for each individual binomial proportion are obtained using the Wilson (Score) confidence limits with continuity correction, given by the following

$$\frac{2n_i\hat{p}_i + z^2}{2(n_i + z^2)} \pm \frac{\left(1 + z\sqrt{z^2 - 2 - \frac{1}{n_i} + 4\hat{p}_i(n_i(1 - \hat{p}_i) + 1)}\right)}{2(n_i + z^2)}$$

Note: The profile likelihood methods (Methods 8 and 9) are not discussed further here. Please see Newcombe (1998) for more information.

Methods 1, 2, 10, and 11 are available in SAS® 9.3, and Methods 5 and 6 were newly added in SAS® 9.4. Methods 3 and 4 can be computed using the formulas provided. Appendix A contains SAS® code that can be run in earlier versions (prior to 9.4) that can compute Methods 5 and 6. This code also computes Method 7. (Note: Method 7 is not available via SAS® 9.4 procedures.)

## SIX ADDITIONAL METHODS

The next 4 methods are available in SAS® 9.4 via PROC FREQ. The remaining 2 can be computed using the formulas. Method 13 was newly added for SAS® 9.4 and in prior releases, would also need to be computed using the listed formula.

## 12. Hauck-Anderson

$$(\hat{p}_1 - \hat{p}_2) \pm \left( \frac{1}{2\min(n_1, n_2)} + z\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1 - 1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2 - 1}} \right)$$

## 13. Agresti-Caffo

$$(\check{p}_1 - \check{p}_2) \pm z\sqrt{\frac{\check{p}_1(1 - \check{p}_1)}{n_1 + 2} + \frac{\check{p}_2(1 - \check{p}_2)}{n_2 + 2}}$$

where

$$\check{p}_1 = \frac{x_1 + 1}{n_1 + 2} \quad \text{and} \quad \check{p}_2 = \frac{x_2 + 1}{n_2 + 2}$$

## 14. Santner-Snell (Exact)

Denote the (true) difference of proportions as  $d = p_1 - p_2$ . The joint distribution of the responses is given by

$$f(x_1, x_2, n_1, n_2, d, p_2) = \binom{n_1}{x_1} (d + p_2)^{x_1} (1 - d - p_2)^{n_1 - x_1} \binom{n_2}{x_2} p_2^{x_2} (1 - p_2)^{n_2 - x_2}$$

$d$  is the parameter of interest and  $p_2$  is a nuisance parameter. Let  $\hat{d} = \hat{p}_1 - \hat{p}_2$  be the observed difference. The lower and upper confidence 95% confidence intervals are given by

$$d_L = \sup(\delta: P_U(\delta) > \alpha/2) \quad \text{and} \quad d_U = \inf(\delta: P_L(\delta) > \alpha/2)$$

where

$$P_L(\delta) = \sup_{p_2} \left( \sum_{A, T(a) \leq \hat{d}} f(x_1, x_2, n_1, n_2, \delta, p_2) \right)$$

$$P_U(\delta) = \sup_{p_2} \left( \sum_{A, T(a) \geq \hat{d}} f(x_1, x_2, n_1, n_2, \delta, p_2) \right)$$

The set  $A$  includes all 2x2 tables with row sums equal to  $n_1$  and  $n_2$  and  $T(a)$  denotes the value of the test statistic for table  $a$  in  $A$ . Here,  $T(a) = d_1 - d_2$ , which is the unstandardized risk difference.

#### 15. Chan-Zhang (Exact)

Same as Santner-Snell (Method 14), but using the standardized risk difference as the test statistic, which is normalized by the variance given in Miettinen-Nurminen (Method 6).

#### 16. Brown and Li's Jeffreys method

Using Jeffreys Beta(0.5, 0.5) prior, the confidence interval is given by

$$(\bar{p}_1 - \bar{p}_2) \pm z \sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}}$$

where

$$\bar{p}_1 = \frac{x_1 + 1/2}{n_1 + 1} \quad \text{and} \quad \bar{p}_2 = \frac{x_2 + 1/2}{n_2 + 1}$$

#### 17. Miettinen-Nurminen-Brown-Li

Laud and Dane (2014) remarked that the MN and BL methods appear to be somewhat complimentary. As such, they propose a weighted average of the two:

Lower limit:  $k \cdot \text{Lower}_{\text{MN}} + (1-k) \cdot \text{Lower}_{\text{BL}}$       Upper limit:  $k \cdot \text{Upper}_{\text{MN}} + (1-k) \cdot \text{Upper}_{\text{BL}}$

where  $\text{Lower}_{\text{MN}}$  ( $\text{Upper}_{\text{MN}}$ ) is the lower (upper) limit from the Miettinen-Nurminen Method (Method 6) and  $\text{Lower}_{\text{BL}}$  ( $\text{Upper}_{\text{BL}}$ ) is the lower (upper) limit from Brown and Li's Jeffreys method (Method 16).

They suggest using  $k = 2/3$  as a reasonable weight.

## IMPLEMENTATION IN SAS®

The following table records which version(s) of SAS® can implement the above methods and how.

Method	SAS® 9.2	SAS® 9.3	SAS® 9.4
1. Wald, no CC	PROC FREQ	PROC FREQ	PROC FREQ
2. Wald, CC	PROC FREQ	PROC FREQ	PROC FREQ
3. Haldane	Open Code	Open Code	Open Code
4. Jeffreys-Perks	Open Code	Open Code	Open Code
5. Mee	Open Code	Open Code	PROC FREQ
6. Miettinen-Nurminen	Open Code	Open Code	PROC FREQ
7. True profile	Open Code	Open Code	Open Code
8. 'Exact'	N/A	N/A	N/A
9. 'Mid-p'	N/A	N/A	N/A
10. Score, no CC	PROC FREQ	PROC FREQ	PROC FREQ
11. Score, CC	PROC FREQ	PROC FREQ	PROC FREQ
12. Hauck-Andersen	PROC FREQ	PROC FREQ	PROC FREQ
13. Agresti-Caffo	Open Code	Open Code	PROC FREQ
14. Santner-Snell (Exact)	PROC FREQ	PROC FREQ	PROC FREQ
15. Chan-Zhang (Exact)	N/A	PROC FREQ	PROC FREQ
16. Brown-Li	Open Code	Open Code	Open Code
17. Miettinen-Nurminen-Brown-Li	Open Code	Open Code	Open Code

Table 1. SAS® Implementation of the 17 Methods Listed Above

As noted above, all but Methods 8 and 9 are available in SAS® 9.4. The code to generate these CIs is listed below:

```
data testdata;
  input trial treat $ x n alpha;
  datalines;
  1 A 56 70 0.05
  1 B 48 80 0.05
  2 A 9 10 0.05
  2 B 3 10 0.05
  3 A 10 10 0.05
  3 B 0 20 0.05
;

data testdat1;
  set testdata;
  by trial;
  if first.trial then treatn = 1;
  else treatn = 2;
  y = n - x; p = x/n; z = probit(1-alpha/2);
run;

data testdat2a(keep=trial x y z rename=(x=x1 y=y1));
  set testdat1;
  where treatn = 1;
run;

data testdat2b(keep=trial x y rename=(x=x2 y=y2));
  set testdat1;
  where treatn = 2;
run;

data testdat2;
  merge testdat2a testdat2b;
  by trial;
run;

proc transpose data = testdat1 out = x_data(rename=(_NAME_=outcome COL1=count));
  by trial treat;
  var x y;
run;

/* Methods 1, 6 (9.4 only), 10, 12, and 13 (9.4 only) */
ods output PdiffCLs=asympl;
proc freq data=x_data;
  by trial;
  %if &sysver >= 9.4 %then %do;
    tables treat*outcome /riskdiff (CL=(WALD MN WILSON AC HA));
  %end;
  %else %do; tables treat*outcome /riskdiff (CL=(WALD WILSON HA)); %end;
  weight count;
run;

data asympl;
  set asympl;
  length method $25.;
  if Type = "Agresti-Caffo" then method = "13. Agresti-Caffo";
  else if Type = "Hauck-Anderson" then method = "12. Hauck-Anderson";
  else if Type = "Miettinen-Nurminen" then method = "6. Miettinen-Nurminen";
  else if index(Type,"Newcombe") > 0 then method = "10. Score, no CC";
  else if Type = "Wald" then method = "1. Wald, no CC";
  keep trial method LowerCL UpperCL;
run;
```

```

/* Method 5: MEE (9.4 only) */
%if &sysver >= 9.4 %then %do;
ods output PdiffCLs=asympt2;
proc freq data=x_data;
  by trial;
  tables treat*outcome /riskdiff(CL=(MN(CORRECT=NO)));
  weight count;
run;

data asympt2;
  set asympt2;
  length method $25.;
  method = " 5. Mee";
  keep trial method LowerCL UpperCL;
run;
%end;

/* Method 3: Haldane */
data asympt3;
  set testdat2;
  by trial;
  length method $25.;
  method = " 3. Haldane";
  p1 = x1/(x1+y1);
  p2 = x2/(x2+y2);
  psi = (x1/(x1+y1) + x2/(x2+y2))/2;
  u = (1/(x1+y1) + 1/(x2+y2))/4;
  v = (1/(x1+y1) - 1/(x2+y2))/4;
  w = z/(1+z*z*u)*sqrt(u*(4*psi*(1-psi)-(p1-p2)*(p1-p2)) + 2*v*(1-2*psi)*(p1-p2) +
4*z*z*u*u*(1-psi)*psi+z*z*v*v*(1-2*psi)*(1-2*psi));
  theta = ((p1-p2)+z*z*v*(1-2*psi))/(1+z*z*u);
  LowerCL = max(-1,theta - w);
  UpperCL = min(1,theta + w);
  keep trial method LowerCL UpperCL;
run;

/* Method 4: Jeffreys-Perks */
data asympt4;
  set testdat2;
  by trial;
  length method $25.;
  method = " 4. Jeffreys-Perks";
  p1 = x1/(x1+y1);
  p2 = x2/(x2+y2);
  psi = ((x1+0.5)/(x1+y1+1) + (x2+0.5)/(x2+y2+1))/2; /* Same as Haldane, but +1/2
success and failure */
  u = (1/(x1+y1) + 1/(x2+y2))/4;
  v = (1/(x1+y1) - 1/(x2+y2))/4;
  w = z/(1+z*z*u)*sqrt(u*(4*psi*(1-psi)-(p1-p2)*(p1-p2)) + 2*v*(1-2*psi)*(p1-p2) +
4*z*z*u*u*(1-psi)*psi+z*z*v*v*(1-2*psi)*(1-2*psi));
  theta = ((p1-p2)+z*z*v*(1-2*psi))/(1+z*z*u);
  LowerCL = max(-1,theta - w);
  UpperCL = min(1,theta + w);
  keep trial method LowerCL UpperCL;
run;

/* Method 16: Brown and Li's Jeffreys Method */
data asympt5;
  set testdat2;
  by trial;
  length method $25.;
  method = "16. Brown-Li";
  p1 = (x1+0.5)/(x1+y1+1);

```

```

    p2 = (x2+0.5)/(x2+y2+1);
    var = p1*(1-p1)/(x1+y1) + p2*(1-p2)/(x2+y2);
    LowerCL = max(-1, (p1-p2) - z*sqrt(var));
    UpperCL = min(1, (p1-p2) + z*sqrt(var));
    keep trial method LowerCL UpperCL;
run;

data asymp;
  set asymp1
  %if &sysver >= 9.4 %then %do; asymp2 %end; /* Add Method 5 if SAS 9.4 */
  asymp3
  asymp4
  asymp5
  ;
run;

/* Methods 2 and 11 */
ods output PdiffCLs=asymp_cc;
proc freq data=x_data;
  by trial;
  tables treat*outcome /riskdiff(correct CL=(wald wilson));
  weight count;
run;

data asymp_cc;
  set asymp_cc;
  length method $25.;
  if index(Type,"Newcombe") > 0 then method = "11. Score, CC";
  else if index(Type,"Wald") > 0 then method = " 2. Wald, CC";
  keep trial method LowerCL UpperCL;
run;

/* Exact methods: Methods 14 and 15 (Exact) */
ods output PdiffCLs=exact_ss;
proc freq data=x_data;
  by trial;
  tables treat*outcome /riskdiff(cl=(exact));
  weight count;
  exact riskdiff;
run;

data exact_ss;
  set exact_ss;
  length method $25.;
  method = "14. Santner-Snell";
  keep trial method LowerCL UpperCL;
run;

%if &sysver >= 9.3 %then %do; /* Add Method 15 if SAS 9.3+ */
ods output PdiffCLs=exact_cz;
proc freq data=x_data;
  by trial;
  tables treat*outcome /riskdiff(cl=(exact));
  weight count;
  exact riskdiff(method=fmscore);
run;

data exact_cz;
  set exact_cz;
  length method $25.;
  method = "15. Chan-Zhang";
  keep trial method LowerCL UpperCL;
run;

```

```

data exact;
  set exact_ss exact_cz;
run;

%end;
%else %do;

data exact;
  set exact_ss;
run;

%end;

/* Combine all of the outputs together */
data final;
  set asymp asymp_cc exact;
run;

/* Sort all of the outputs by trial and method */
proc sort data = final out = final;
  by trial method;
run;

```

Note: If running a version of SAS® prior to 9.4, Methods 5, 6, and 13 will need to be computed by hand. Code to do so appears in Appendix A. (This is also how Method 7 can be obtained as it is not included in SAS® 9.4.)

It is also worth noting that SAS® also has an asymptotic Farrington-Manning method available (via the `/riskdiff(CL=(FM))` option in the `table` statement). While reading the documentation, it appears similar to the Mee Method, but is designed for non-inferiority testing and should not be used in general for constructing confidence intervals for differences of binomials.

### THREE MORE METHODS

Beyond the two exact methods (Methods 14 and 15) above, there is another exact method by Agresti and Min (Method 18) that inverts one two-sided score test instead of two one-sided exact score tests. SAS® does not currently have an implementation for the Agresti-Min method, but StatXact does. The code using Proc StatXact appears in Appendix B. Wang (2010) also explored constructing smallest one-sided confidence intervals. Shan and Wang (2013) created an R package, “ExactCldiff” to implement this method (Method 19).

Pradhan and Banerjee developed a weighted profile likelihood method (Method 20) where the weights come from a noninformative Jeffreys’ prior. It is worth noting that without weights, this method reduces to the True Profile method (Method 7) proposed by Newcombe. The authors provided SAS® IML code in their paper.

### COMPARING THE METHODS

An important consideration when evaluating a method is to consider the coverage probability. The first question to ask is whether or not *average* coverage probability is desired or *minimum* coverage probability is desired. Approximate methods seek to achieve the former while exact methods ensure the latter.

Newcombe (1998) provided a detailed analysis of the initial 11 methods discussed above. Subsequently, Laud and Dane (2014) as well as Brown and Li (2005) examined coverage probabilities for a variety of approximate methods while Fagerland et al (2011) also considered exact methods.

The focus of this paper is on the application and interpretation of confidence intervals and not on the theoretical properties (beyond exact methods ensuring the desired coverage probability). Interested readers should consult the references listed to learn more.

Table II in Newcombe (1998) provides examples for the initial 11 methods. That table is recast below, with additional entries for the 9 other methods discussed above.



Method	56/70-48/80	9/10-3/10	10/10 – 0/20
1. Wald, no CC	0.0575, 0.3425	0.2605, 0.9395	1.0000*, 1.0000
2. Wald, CC	0.0441, 0.3559	0.1605, >1.0000*	0.9250, >1.0000*
3. Haldane	0.0535, 0.3351	0.1777, 0.8289	0.7482, 1.0000
4. Jeffreys-Perks	0.0531, 0.3355	0.1760, 0.8306	0.7431, >1.0000*
5. Mee	0.0533, 0.3377	0.1821, 0.8370	0.7225, 1.0000
6. Miettinen-Nurminen	0.0528, 0.3382	0.1700, 0.8406	0.7156, 1.0000
7. True profile	0.0547, 0.3394	0.2055, 0.8634	0.8252, 1.0000
8. 'Exact'	0.0529, 0.3403	0.1393, 0.8836	0.6915, 1.0000
9. 'Mid-p'	0.0539, 0.3393	0.1834, 0.8640	0.7411, 1.0000
10. Score, no CC	0.0524, 0.3339	0.1705, 0.8090	0.6791, 1.0000
11. Score, CC	0.0428, 0.3422	0.1013, 0.8387	0.6014, 1.0000
12. Hauck-Andersen	0.0494, 0.3506	0.1922, >1.0000*	0.9500, >1.0000*
13. Agresti-Caffo	0.0525, 0.3358	0.1600, 0.8400	0.6922, >1.0000*
14. Santner-Snell (Exact)	0.0390, 0.3533	0.1268, 0.8853	0.6915, 1.0000
15. Chan-Zhang (Exact)	0.0382, 0.3453	0.1608, 0.8795	0.6915, 1.0000
16. Brown-Li	0.0540, 0.3400	0.1869, 0.9040	0.7854, >1.0000*
17. Miettinen-Nurminen-Brown-Li	0.0532, 0.3388	0.1756, 0.8618	0.7389, 1.0000
18. Agresti-Min (Exact)	0.0506, 0.3406	0.1557, 0.8522	0.7000, 1.0000
19. Wang (Exact)	0.0308, 0.3571	0.1268, 0.8827	0.6915, 1.0000
20. Pradhan-Banerjee	0.0522, 0.3360	0.1581, 0.8210	0.6954, 0.9962*

**Table 2. Examples of Confidence Intervals from the 20 Methods Under 3 Scenarios**

CC: continuity correction \*aberrations (directly calculated limits outside [-1, 1] or incorrectly equal to difference)

Note: The Pradhan-Banerjee method (Method 20) computed an upper limit equal to 0.9962, which is less than the difference. Thus, the confidence interval does not even contain the point estimate. Methods 2, 4, 12, 13, and 16 resulted in upper limits above 1, which is outside the possible range for the difference. Method 1 resulted in a zero-width confidence interval.

## A CASE STUDY: EXAMPLE FROM AN HIV CLINICAL TRIAL

Consider Week 48 results of the PROGRESS trial as presented by Reyes (2010). There, the proportion of patients who were responders on a new treatment was 84/101 (83.2%) while the proportion of patients who were responders on the standard of care was 89/105 (84.8%). Non-inferiority (NI) testing was to be performed by constructing a confidence interval for the difference in proportions. If the lower bound of the confidence interval is above a pre-defined margin (eg, -12%), then NI can be concluded. The table below shows what the confidence intervals from the various methods above (excluding Methods 8 and 9) would be.

Method	84/101 – 89/105
1. Wald, no CC	-0.1162, 0.0843
2. Wald, CC	-0.1259, 0.0940
3. Haldane	-0.1152, 0.0834
4. Jeffreys-Perks	-0.1160, 0.0843
5. Mee	-0.1188, 0.0857
6. Miettinen-Nurminen	-0.1191, 0.0860
7. True profile	-0.1176, 0.0849
8. 'Exact'	N/A
9. 'Mid-p'	N/A
10. Score, no CC	-0.1177, 0.0851
11. Score, CC	-0.1245, 0.0918
12. Hauck-Andersen	-0.1216, 0.0898
13. Agresti-Caffo	-0.1168, 0.0850
14. Santner-Snell (Exact)	-0.1514, 0.1219
15. Chan-Zhang (Exact)	-0.1227, 0.0869
16. Brown-Li	-0.1170, 0.0852
17. Miettinen-Nurminen-Brown-Li	-0.1184, 0.0857
18. Agresti-Min (Exact)	-0.1202, 0.0877
19. Wang (Exact)	-0.1196, 0.1029
20. Pradhan-Banerjee	-0.1179, 0.0852

**Table 3. Confidence Intervals from an HIV Clinical Trial**

The method specified in the analysis plan was Chan-Zhang (Method 15). From the table above, using a 12% NI margin, we would fail to conclude non-inferiority. (Note: The study *did* meet its prespecified 20% NI margin.)

If, instead of the Chan-Zhang method, the Miettinen-Nurminen method (Method 6) was used, the lower bound would have been -11.9% and non-inferiority would be concluded using a 12% NI margin.

This example shows that the choice of method is not purely an academic exercise. (That being said, one may not have the highest confidence in the results if different methods can lead to different conclusions.)

## DISCUSSION

Fundamentally, the difference between the methods is whether or not coverage probabilities need to be guaranteed or not. Starting with approximate methods, the Wald method (Method 1) is the most widely taught. This is primarily due to its computational simplicity. Four of the other methods presented here (Methods 2, 12, 13, and 16) are slight variants on the Wald method, each seeking to overcome some of its known deficiencies.

In his paper, Newcombe (1998) showed that both the Haldane method (Method 3) and Jeffreys-Perks method (Method 4) can still yield aberrations in the confidence intervals. The true profile likelihood method (Method 7) was also shown to be theoretically inferior to the Miettinen-Nurminen method (Method 6). Method 6 also uses an unbiased estimate of the variance of the difference, unlike Mee's method (Method 5), though the two are nearly identical otherwise.

The method of Pradhan-Banerjee (Method 20) was shown to perform better than the true profile likelihood method (Method 7) and comparable to the Newcombe score method (Method 10).

The Miettinen-Nurminen-Brown-Li method (Method 17) proposed by Laud and Dane (2014) with a weighting factor of  $k = 2/3$  is an empirically-based hybrid of two other existing methods that seeks to benefit from each of its components. It may be best to stick with either method rather than arbitrarily weighting the two.

Newcombe (1998) also noted that the profile likelihood methods (Methods 8 and 9) are difficult to implement and require large amounts of computation time unless  $n_1$  and  $n_2$  are small. He goes on to show how well his score methods (Methods 10 and 11) perform relative to the others. One of its most attractive features is its computational simplicity relative to other methods such as the Miettinen-Nurminen method (Method 6). That being said, since Method 6 is now part of SAS®, it is this author's opinion that this is the preferred approximate method.

Shifting focus to exact methods, there are four methods under consideration in this paper. There is general consensus that the Santner-Snell method (Method 14) is overly conservative relative to other exact methods and therefore is not recommended. The Chan-Zhang method (Method 15) and the Wang method (Method 19) are based on constructing two one-sided confidence intervals and taking the intersection while the Agresti-Min method (Method 18) is based on a single two-sided confidence interval.

In their paper, Shan and Wang (2013) give an example (21/23 – 19/32) that shows their exact confidence interval is narrower than the two available in SAS (Methods 14 and 15). However, in the larger sample size comparisons above (eg, 56/70 – 48/80), Wang's method (Method 19) resulted in the widest confidence interval. Additionally, the method is very computationally intensive. The moderate sample size examples took about 1 hour to run on an Intel® i7-2640M CPU@2.80GHz laptop with 4GB of RAM. Thus, it is not clear that Wang's method should be preferred over the other methods.

When performing equivalence testing, the US Food and Drug Administration (FDA) recommends the use of a two one-sided tests (TOST) approach as put forth by Schuirmann (1987). Thus, while the Agresti-Min method (Method 18) may result in smaller confidence intervals than Chan-Zhang (Method 15), it may not be considered appropriate by regulatory agencies. (It is worth noting that the approximate methods usually yield symmetric confidence intervals around the point difference and thus would automatically satisfy the TOST approach.)

Lastly, the Chan-Zhang method (Method 15) is available in SAS® 9.3 and later while the Agresti-Min method (Method 18) is only available in StatXact. For these reasons, it is this author's opinion that the Chan-Zhang method (Method 15) should be used an exact method.

Beyond the recommendation for using a TOST approach, no single method has been recommended by the FDA for use in clinical trials. As the case study above illustrates, the choice of method can be critical when determining the success or failure of a clinical trial and thus should be chosen with care.

## SUMMARY

Multiple methods exist for computing confidence intervals for the difference of two independent binomial proportions. The general consensus is that the most widely taught method (Method 1) is inappropriate in many situations and should not be used. This paper presented 20 such methods, but that list is far from exhaustive.

There are two main classes of confidence interval methods: approximate and exact. In terms of approximate methods, the Miettinen-Nurminen method (Method 6) is recommended. If an exact method is preferred, the Chan-Zhang method (Method 15) should be used.

## REFERENCES

1. Newcombe, R., "Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods", *Statist. Med.* 17, 873-890 (1998).
2. Chan, I. S. F. and Zhang, Z., "Test-Based Exact Confidence Intervals for the Difference of Two Binomial Proportions", *Biometrics*. 55, 1202-1209 (1999).
3. Laud, P. and Dane, A., "Confidence intervals for the difference between independent binomial proportions: comparison using a graphical approach and moving averages", *Pharmaceut. Statist.* 13, 294-308 (2014).
4. Wang, W., "On construction of the smallest one-sided confidence interval for the difference of two proportions", *The Annals of Statistics* 38, 1227-1243 (2010).
5. Shan, G. and Wang, W., "ExactCldiff: an R package for computing exact confidence intervals for the difference of two proportions", *The R Journal*, 5(2), 62-71 (2013).
6. Pradhan, V. and Banerjee, T., "Confidence Interval of the Difference of Two Independent Binomial Proportions Using Weighted Profile Likelihood", *Communications in Statistics - Simulation and Computation*, 37: 4, 645-659 (2008).
7. Schuirmann, D., "A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability", *J. Pharma. and Biopharma.*, 15: 6, 657-680 (1987).
8. Beal, S. L., "Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples", *Biometrics*, 43, 941-950 (1987).
9. Agresti A, Caffo B., "Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures", *American Statistician* 54: 280-288 (2000)
10. Mee RW., "Confidence bounds for the difference between two probabilities", *Biometrics* 40:1175-1176 (1984).
11. Miettinen OS, Nurminen M., "Comparative analysis of two rates." *Statistics in Medicine* 4:213-226 (1985).
12. Brown, L. and Li, X., "Confidence intervals for two sample binomial distribution", *Journal of Statistical Planning and Inference*, 130(1), 359-375 (2005).
13. Agresti A, Min Y., "On small-sample confidence intervals for parameters in discrete distributions", *Biometrics* 57(3): 963-971 (2001).
14. Hauck WW, Anderson S., "A comparison of large-sample confidence interval methods for the difference of two binomial probabilities." *The American Statistician* 40(4): 318-322 (1986).
15. Fagerland, M.W., Lydersen, S. and Laake, P., "Recommended confidence intervals for two independent binomial proportions", *Statistical methods in medical research*, 24(2), 224-254 (2015).
16. Reynes JLA, Pulido F, et al., "Lopinavir/ritonavir combined with raltegravir demonstrated similar antiviral efficacy and safety as lopinavir/ritonavir combined with tenofovir disoproxil fumarate/emtricitabine in treatment-naïve HIV-1 infected subjects." Program and abstracts of the XVIII International AIDS Conference; July 18-23, 2010; Vienna, Austria. Abstract MOAB0101.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Will Garner  
Company: Gilead Sciences, Inc.  
Address: 333 Lakeside Dr  
City, State ZIP: Foster City, CA 94404  
Work Phone: (650) 522-5691  
E-mail: will.garner@gilead.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

**APPENDIX A – SAS® CODE TO GENERATE METHODS 5, 6, 7, AND 13**

Below is SAS® code to implement the methods in SAS® releases prior to SAS® 9.4 (as well as Method 7, which is not available in SAS® 9.4).

```

/* Method 5: Miettinen-Nurminen-Mee (Mee) */
data method_mn_mee;
  set testdat2;
  do j = -0.99999 to 0.99999 by 0.00001; /* increment over all possible differences
(delta) */
    n1 = x1 + y1;      n2 = x2 + y2;      p1 = x1/n1;      p2 = x2/n2;
    n = n1+n2;      theta = n2/n1;      a = 1 + theta;
    b = -(1 + theta + p1 + theta*p2 + j*(theta+2));
    c = j*j + j*(2*p1+theta+1)+p1+theta*p2;      d = -p1*j*(1+j);
    v = (b*b*b)/(27*a*a*a) - (b*c)/(6*a*a) + d/(2*a);
    u = (abs(v)/v)*sqrt((b*b)/(9*a*a) - c/(3*a));
    w = (constant('pi')+acos(v/(u*u*u)))/3;
    p1_mle = 2*u*cos(w) - b/(3*a);      p2_mle = p1_mle - j;
    var = p1_mle*(1-p1_mle)/n1 + p2_mle*(1-p2_mle)/n2;
    var_mn = (n/(n-1))*var;      var_mee = var;
    T_mn = ((p1-p2)-j)/sqrt(var_mn);      T_mee = ((p1-p2)-j)/sqrt(var_mee);
    TPL = x1*log(p1_mle/(x1/n1))+x2*log(p2_mle/(x2/n2))+y1*log((1-
p1_mle)/(y1/n1))+y2*log((1-p2_mle)/(y2/n2));
    output;
  end;
run;

data method_mee_min(keep=trial j rename=(j=LowerCL)) method_mee_max(keep=trial j
rename=(j=UpperCL));
  set method_mn_mee;
  by trial;
  where -z < T_mee < z;
  if first.trial then output method_mee_min;
  if last.trial then output method_mee_max;
run;

data method_mee(keep=trial LowerCL UpperCL method);
  merge testdat2 method_mee_min method_mee_max;
  by trial;
  length method $25.;
  method = " 5. Mee";
run;

/* Method 6: Miettinen-Nurminen (MN) */
data method_mn_min(keep=trial j rename=(j=LowerCL)) method_mn_max(keep=trial j
rename=(j=UpperCL));
  set method_mn_mee;
  by trial;
  where -z < T_mn < z;
  if first.trial then output method_mn_min;
  if last.trial then output method_mn_max;
run;

data method_mn(keep=trial LowerCL UpperCL method);
  merge testdat2 method_mn_min method_mn_max;
  by trial;
  length method $25.;
  method = " 6. Miettinen-Nurminen";
run;

/* Method 7: True Profile Likelihood (if both proportions are not 0 or 1) */
data method_truepro_min(keep=trial j rename=(j=LowerCL)) method_truepro_max(keep=trial
j rename=(j=UpperCL));

```

```

    set method_mn_mee;
    by trial;
    where TPL >= -z*z/2;
    if first.trial then output method_truepro_min;
    if last.trial then output method_truepro_max;
run;

data method_truepro(keep=trial LowerCL UpperCL method);
    merge testdat2 method_truepro_min method_truepro_max;
    by trial;
    length method $25.;
    method = " 7. True profile";
run;

/* Method 13: Agresti-Caffo (AC) */
data method_ac(keep=trial LowerCL UpperCL method);
    set testdat2;
    length method $25.;
    method = "13. Agresti-Caffo";
    p1 = (x1+1)/((x1+y1)+2);
    p2 = (x2+1)/((x2+y2)+2);
    LowerCL = max(-1, (p1-p2)-z*sqrt(p1*(1-p1)/((x1+y1)+2)+p2*(1-p2)/((x2+y2)+2)));
    UpperCL = min(1, (p1-p2)+z*sqrt(p1*(1-p1)/((x1+y1)+2)+p2*(1-p2)/((x2+y2)+2)));
run;

```

## APPENDIX B – SAS® CODE TO GENERATE METHOD 18 USING PROC STATXACT

Below is SAS® code to implement Method 18 (Agresti-Min) using PROC StatXact.

```

proc binomial data = x_data out=exact;
    by trial;
    pd /ex two std;
    po treat;
    ou outcome;
    weight count;
run;

data exactci;
    set exact(keep=item value by_val);
    if item in ('TLWR_CI' 'TUPR_CI ');
    trial = input(by_val,8.);
run;

proc sort data = exactci;
    by trial;
run;

proc transpose data = exactci out = exactci;
    by trial;
    var value;
    id item;
run;

data final(keep=trial LowerCL UpperCL method);
    set exactci;
    label LowerCL = "95% Lower Confidence Limit"
           UpperCL = "95% Upper Confidence Limit";
    LowerCL = -TUPR_CI;
    UpperCL = -TLWR_CI;

    length method $25.;
    method = "18. Agresti-Min";
run;

```