SAS ® Macros CORR_P and TANGO: Interval Estimation for the Difference Between Correlated Proportions in Dependent Samples

6 authors, including:

Patricia Rodriguez de Gil
University of South Florida
29 PUBLICATIONS   **146** CITATIONS

SEE PROFILE

Thanh Pham
University of South Florida
27 PUBLICATIONS   **95** CITATIONS

SEE PROFILE

Diep Nguyen
University of South Florida
36 PUBLICATIONS   **232** CITATIONS

SEE PROFILE

Eunsook Kim
University of South Florida
104 PUBLICATIONS   **2,122** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Intensive Longitudinal Data View project

Paper SD-03

SAS® Macros CORR_P and TANGO: Interval Estimation for the Difference Between Correlated Proportions in Dependent Samples

Patricia Rodríguez de Gil, Jeanine Romano, Thanh Pham, Diep T. Nguyen,
Jeffrey D. Kromrey, Eun Sook Kim
University of South Florida, Tampa, FL

ABSTRACT

Two proportions from the same sample of observations or from matched-pair samples are correlated. A number of studies proposed interval estimation for the difference in correlated proportions (e.g., Bonett & Price, 2011; Newcombe, 1998; Tango, 1998). Considering that confidence intervals (CI) are more informative than point estimates but the CI for the difference in correlated proportions is not readily available in SAS, the purpose of this paper is to provide a SAS macro for three types of confidence intervals suggested in the literature: Wald CI, adjusted Wald CI, and approximate CI proposed by Tango. The results from a simulation study comparing these three confidence intervals are also presented.

INTRODUCTION

The difference in correlated or dependent proportions is often of interest in studies such as pretest-posttest designs, matched-pair designs, and rater-agreement designs (Bonett & Price, 2011). Currently, McNemar's test (1947) is commonly conducted to test the equivalence of two correlated proportions. McNemar's test is also implemented in SAS (the AGREE option in PROC FREQ). Two dichotomous outcomes measured from dependent samples are summarized in a 2x2 contingency table. For example, the results of two diagnostic tests of dyslexia (Tests A and B) administered to a sample of pre-school children are summarized with the cell frequency of the cases in which diagnosis is correct in both tests (correct-correct), correct in Test A but incorrect in Test B (correct-incorrect), incorrect in Test A but correct in Test B (incorrect-correct), and incorrect in both tests (incorrect-incorrect) as illustrated in Table 1.

		Test B		
		Correct	Incorrect	Total
Test A	Correct	a (π_{11})	b (π_{12})	a + b (π_{1+})
	Incorrect	c (π_{21})	d (π_{22})	c + d (π_{2+})
	Total	a + c (π_{+1})	b + d (π_{+2})	n (1)

Table 1. Contingency Table of Dyslexia Diagnosis from Tests A and B

McNEMAR TEST Q STATISTIC

In this example, the null hypothesis is that the population proportion of correct diagnosis on dyslexia using Test A (π_{1+} = [a + b]/n) equals the population proportion of correct diagnosis using Test B (π_{+1} = [a + c]/n). Testing the null hypothesis π_{1+} - π_{+1} = 0 is equivalent to testing π_{12} - π_{21} = 0 because π_{11} is common to both π_{1+} and π_{+1}. Given the null hypothesis McNemar's test Q statistic is computed as

$$Q = \frac{(b-c)^2}{b+c} \ .$$

where

b is the cell frequency of cases in which diagnosis is correct in test A but incorrect in test B, and
c is the cell frequency of cases in which diagnosis is incorrect in test A but correct in test B

Under the null hypothesis, the Q statistic follows an asymptotic chi-square distribution with one degree of freedom when $b + c$ is greater than 10 (McNemar, 1947).

WALD CONFIDENCE INTERVAL

A 100(1 – α)% Wald confidence interval for the difference in the population proportions (π_{12} - π_{21}) can be estimated as

$$\hat{\pi}_{12} - \hat{\pi}_{21} \pm z_{\alpha/2}\sqrt{[\hat{\pi}_{12} + \hat{\pi}_{21} - (\hat{\pi}_{12} - \hat{\pi}_{21})^2]/n} \cdot \tag{1}$$

where

$z_{\alpha/2}$ is a critical value at $\alpha/2$ from the standard normal distribution

However, when Newcombe (1998) examined the performance of existing methods to compute the confidence intervals for the difference in correlated proportions, the Wald confidence interval (Equation 1) showed inadequate performance.

ADJUSTED WALD CONFIDENCE INTERVAL

Recently, Bonett and Price (2011) proposed an alternative CI by making an adjustment to the Wald interval shown in Equation 1,

$$\hat{\pi}_{12} - \hat{\pi}_{21} \pm z_{\alpha/2}\sqrt{[\hat{\pi}_{12} + \hat{\pi}_{21} - (\hat{\pi}_{12} - \hat{\pi}_{21})^2]/n + 2} \cdot \tag{2}$$

where each cell proportion is computed by adding one to the cell frequency and two to the total n, for example, $\hat{\pi}_{12} = (b + 1)/(n + 2)$ and $\hat{\pi}_{21} = (c + 1)/(n + 2)$. Bonett and Price reported that the adjusted Wald interval performs as well as an approximate CI proposed by Tango (1998) but the computation is simpler than Tango's.

TANGO CONFIDENCE INTERVAL

On the other hand, the confidence interval for the difference in two correlated proportions ($\lambda = \pi_{1+} - \pi_{+1} = \pi_{12} - \pi_{21}$) developed by Tango is estimated by solving the following two equations iteratively until the change in estimation is infinitesimal below the predetermined cutoff.

$$\frac{b - c - n\lambda}{\sqrt{n(2\hat{\pi}_{21} + \lambda(1-\lambda))}} = \pm z_{\alpha/2} , \tag{3}$$

where

λ is the difference between two correlated proportions, and

$\hat{\pi}_{21}$ in Equation 3 is estimated as

$$\hat{\pi}_{21} = \frac{\sqrt{(B^2 - 4AC) - B}}{2A} \cdot \tag{4}$$

where

A = $2n$,
B = -b – c + $(2n – $ b + c$)\lambda$, and
C = -c$\lambda(1 – \lambda)$.

Although the computational procedures for Tango's CI are more complex than Wald and adjusted Wald intervals, the upper and lower limits are easily found through the secant method with empirically good coverage probabilities (Tango, 1999) and can be applied to small samples with off-diagonal zero cells (Tango, 1998).

SAS MACROS

Two SAS macros are used to estimate the endpoints of the confidence intervals. The macro TANGO estimates endpoints of the Tango (1998) interval. The macro CORR_P estimates the two Wald estimates and calls the macro TANGO. Both macros are written in BASE SAS.

A SAS macro for confidence intervals suggested by Tango (1998) is presented below. Equations 3 and 4 are solved iteratively by evaluating the change in the estimated λ (i.e., the difference in two correlated proportions).

```
%macro Tango (a, b, c, d, n, X0, X1,Z);

iteration = 1; * Initializing iteration counter;
a = &a;b = &b;c = &c;d = &d;n = &n;X0 = &X0;x1 = &X1;Z = &Z;
do until (abs(change) < .000001);
```

```
       * Evaluate x0;
       lambda = x0;
       AA = 2*n;
       BB = -1*b-c+(2*n - b + c)*lambda;
       CC = -1*c*lambda*(1-lambda);
       q21_hat = (SQRT(BB**2 - 4*AA*CC) - BB)/(2*AA);
       if (n*(2*q21_hat+lambda*(1-lambda))>=0) then fx0 = b-c-n*lambda -
           Z*SQRT(n*(2*q21_hat+lambda*(1-lambda)));

       if (n*(2*q21_hat+lambda*(1-lambda))<0) then fx0 = b-c-n*lambda;
      *+-------------------------------------------------+
       * Evaluate x1;
       +-------------------------------------------------+;
       lambda = x1;
       AA = 2*n;
       BB = -1*b-c+(2*n - b + c)*lambda;
       CC = -1*c*lambda*(1-lambda);
       q21_hat = (SQRT(BB**2 - 4*AA*CC) - BB)/(2*AA);
       if (n*(2*q21_hat+lambda*(1-lambda))>=0) then fx1 = b-c-n*lambda -
           Z*SQRT(n*(2*q21_hat+lambda*(1-lambda)));
       if (n*(2*q21_hat+lambda*(1-lambda))<0) then fx1 = b-c-n*lambda;

      *+-------------------------------------------------+
       * Evaluate change in lambda;
       +-------------------------------------------------+

       x2 = x1 - (fx1 * (x1-x0)/(fx1 - fx0)); * update value of lambda to
                                                  evaluate;
       change = x2 - x1;

  *+-------------------------------------------------------------------------+
  * Set x0 and x1 for next iteration of loop, and increment iteration counter;
  +-------------------------------------------------------------------------+;
       x0 = x1;
       x1 = x2;
       iteration = iteration + 1;
  end;
%mend Tango;
```

The first part of the macro CORR_P runs PROC FREQ with the raw data given below as an input and yields the cell frequencies in the contingency table (i.e., a through d in Table 1) as an output. Next, the output data are restructured to be used in the computation of three types of confidence intervals. In the third part of the macro, the Wald CI (denoted by Wald_old_up and _down) and adjusted Wald CI (denoted by Wald_new_up and _down) are computed using Equations 1 and 2, and the macro for Tango CI is called. A FILE PRINT statement is used to print the obtained confidence interval limits of Tango CI, Wald CI, and adjusted Wald, respectively.

```
%macro CORR_P (indata=, row_var =, col_var =, percent_CI = 95);
* output with count and percent;
* print suppressed;
PROC FREQ /*noprint*/;
 TABLES &row_var*&col_var / norow nocol out=a;
run;
*+-------------------------------------------------------------------------+
* Restructure output data set to a single observation with four frequencies
   and four percents;
+-------------------------------------------------------------------------+;
DATA two;
 set a;
 row = _n_;
 array pct[4] pct1 - pct4;
```

```
 array frq[4] freq1 - freq4;
 retain pct1 - pct4 freq1 - freq4;
 pct[row] = PERCENT;
 frq[row] = COUNT;
 if row = 4 then output;
 keep pct1 - pct4 freq1 - freq4;
proc print data = two;
run;

DATA CI;
set two;
a = Freq4;
b = Freq2;
c = Freq3;
d = Freq1;
* These are four cell frequencies;
*a = 18;
*b = 4;
*c = 12;
*d = 5;
*Proc print data = uplim;
*run;
n = a+b+c+d;
lambda_hat = (b-c)/n; * Sample difference in proportions;

CI_prop = &percent_CI/100;
ZL = PROBIT(1 - (1 - CI_prop)/2); * Upper critical value of Z;
ZU = PROBIT((1 - CI_prop)/2); * Lower critical value of Z;


*+---------------------------------------------+
* Endpoints of orignal Wald interval;
+---------------------------------------------+;

Wald_old_up = lambda_hat + ZL*sqrt ((((c+b)/n)-lambda_hat**2)/n);
Wald_old_low = lambda_hat + ZU*sqrt ((((c+b)/n)-lambda_hat**2)/n);

*+---------------------------------------------+
    * Endpoints of adjusted Wald interval;
 +---------------------------------------------+;
newb = (b +1)/(n+2);
newc=(c+1)/(n+2);
Wald_new_up = newb-newc + ZL *sqrt((newb+newc-(Newb-newc)**2)/(n+2));
Wald_new_low = newb-newc + ZU *sqrt((newb+newc-(Newb-newc)**2)/(n+2));

x0 = -.9999; * starting values of lambda;
x1 = .9999;
%Tango (a, b, c, d, n, X0, X1,ZU);
Tango_up = X1;
x0 = -.9999; * starting values of lambda;
x1 = .9999;
%Tango (a, b, c, d, n, X0, X1,ZL);
Tango_Low = X1;
run;

DATA_null_;
      set CI;
```

```
        CI_Level = CI_prop*100;
file print header = h notitles;
put    @1 'Original Wald' @20 wald_old_low 8.5 @30 wald_old_up 8.5 //
       @1 'Revised Wald'  @20 wald_new_low 8.5 @30 wald_new_up 8.5 //
       @1 'Tango'         @20 Tango_low 8.5 @30 Tango_up 8.5;
return;
h: put      @1 'Confidence Intervals for Correlated Proportions' //
            @1 'Level of Confidence:' @31 CI_Level @33 '%' //
            @1 'Sample Proportion Difference:' @30 lambda_hat 8.5 ///
            @1 'CI Method' @20 '  Lower' @30 '  Upper' /
            @1 '-------------' @20 '--------' @30 '--------';

run;

%mend CORR_P;
```

MACRO EXECUTION

As an example of the use of the macro CORR_P, the following data step creates a data set called "study", containing 39 observations. Each observation corresponds to **pass** or **fai**l in Biology and Algebra. The Macro CORR_P is called to illustrate the estimation of confidence intervals for correlated proportions for the SAS user's provided data.

```
DATA study;
INPUT bio $ alg $ ;
cards;
P P
P P
P P
P P
P P
P P
P P
P P
P P
P P
P P
P P
P P
P P
P P
P P
P P
P P
P F
P F
P F
P F
F P
F P
F P
F P
F P
F P
F P
F P
F P
F P
```

```
F  P
F  P
F  F
F  F
F  F
F  F
F  F
;
run;

%CORR_P(indata=study, row_var =bio, col_var = alg, percent_CI = 95);
```

OUTPUT EXAMPLE OF MACRO CORR_P

The output table from the macro CORR_P is presented in Output 1. The simple table presents the sample difference between the two proportions, the level of confidence requested in the macro call, and the endpoints of the three confidence intervals. In this sample, the upper endpoints of the three intervals are nearly identical, but greater differences are evident in the lower endpoints.

```
Confidence Intervals for Correlated Proportions

Level of Confidence:          95%

Sample Proportion Difference: 0.20513


CI Method            Lower     Upper
-------------      --------  --------
Original Wald       0.01469   0.39556

Revised Wald        0.00130   0.38894

Tango               0.00443   0.39263
```

Output 1: Wald and Tango Confidence Intervals

SIMULATION RESEARCH

The accuracy and precision of the three confidence interval estimation methods were investigated using Monte Carlo methods. In this simulation study three design factors were manipulated: (a) sample size (n = 10, 50, 100, 500, and 1000), (b) magnitude and direction of population correlation between the two proportions (φ = -.40 to +.40, in increments of .10), and (c) difference between population proportions (Δ = -.30, -.25, -.10, -.05, .00, .05, .10, .25, and .30). These factors in the Monte Carlo study were completely crossed, yielding 405 conditions. For each condition, 100,000 replications were conducted. The use of 100,000 estimates provides adequate precision for the investigation of confidence interval coverage and width. For example, 100,000 replications provide a maximum 95% confidence interval width around an observed proportion that is $\pm$ .0031 (Robey & Barcikowski, 1992).

The data for the simulation were generated using uniform random numbers on the zero to one interval (the SAS RANUNI function). The values of the random numbers were used to assign observations to cells in the contingency table. In each sample, the three confidence interval methods were applied to provide estimates of 90%, 95% and 99% intervals.

The distributions of 95% confidence interval coverage estimates across all conditions simulated are presented in Figure 1. These results show that the adjusted Wald method provides the best interval coverage overall among the three methods. The original Wald method provides lower coverage than the nominal confidence level in many conditions. Both the adjusted Wald and the Tango methods provide the higher coverage for most of the sample conditions, but the Tango method produces substantial under coverage in a few conditions. The distribution of interval coverage of the adjusted Wald method is closer to the expected confidence level than that of Tango method. Analogous results were found for the 90% and 99% confidence intervals.
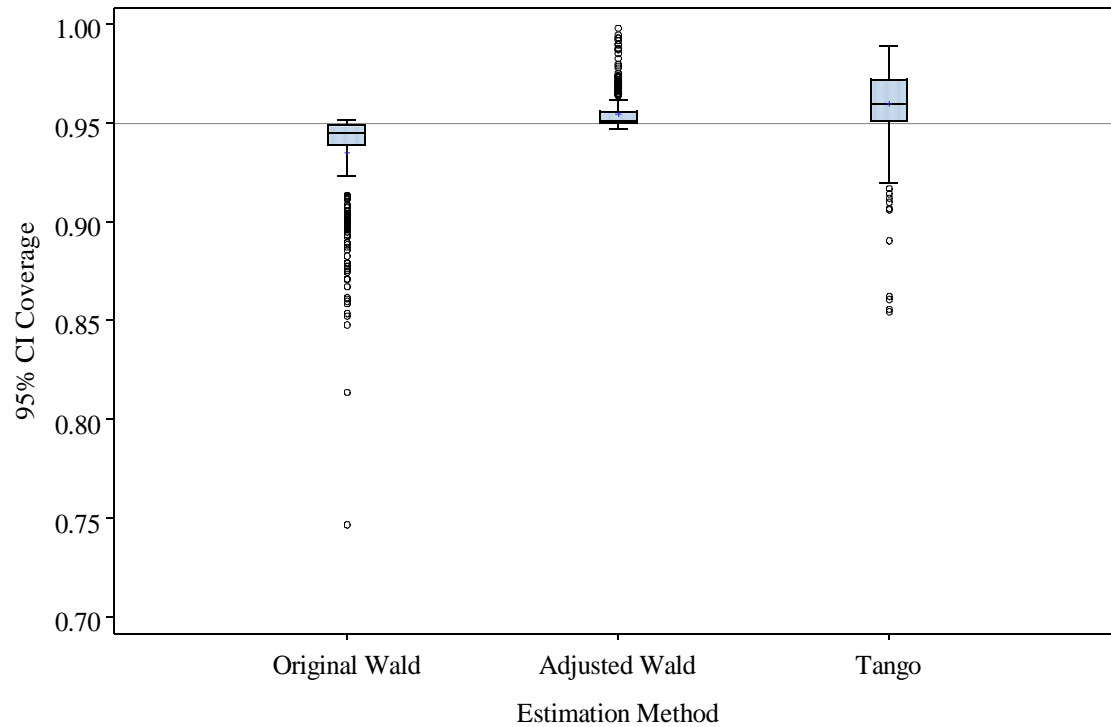
Figure 1. Distributions of Interval Coverage at 95% Confidence Level

Figure 2 presents the distributions of confidence interval widths across all conditions simulated. The results show that these distributions are nearly identical across all three estimation methods, indicating that the methods provide the same degree of precision. The distributions of interval widths for 90% and 99% confidence levels were also nearly identical.
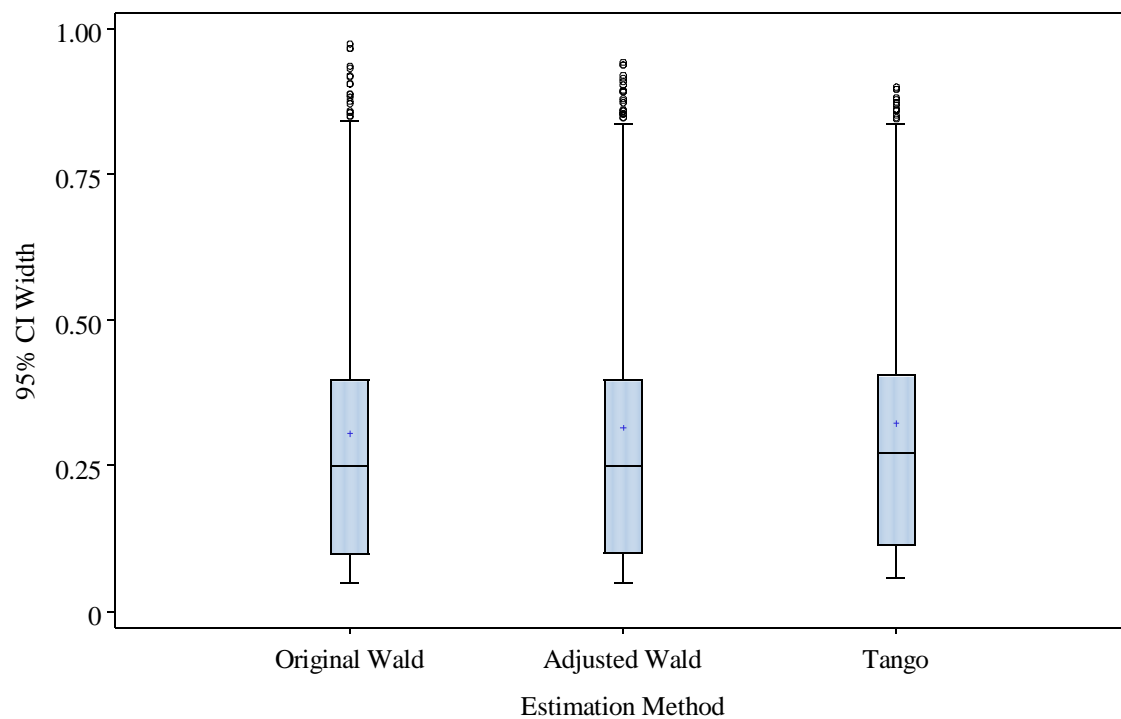

Figure 2. Distributions of Interval Widths at 95% Confidence Level

The influence of sample size on confidence interval coverage is presented in Figure 3. This figure shows that Tango method provides the most accurate coverage at the smallest sample size examined (n = 10) and provides a slight amount of over coverage with large sample sizes. The original Wald method produces substantial under coverage at the small sample sizes and gets closer to the expected confidence interval coverage when sample size increases. The adjusted Wald method provides slight over coverage at the smallest sample size but quickly converges to the nominal .95 coverage with larger samples.
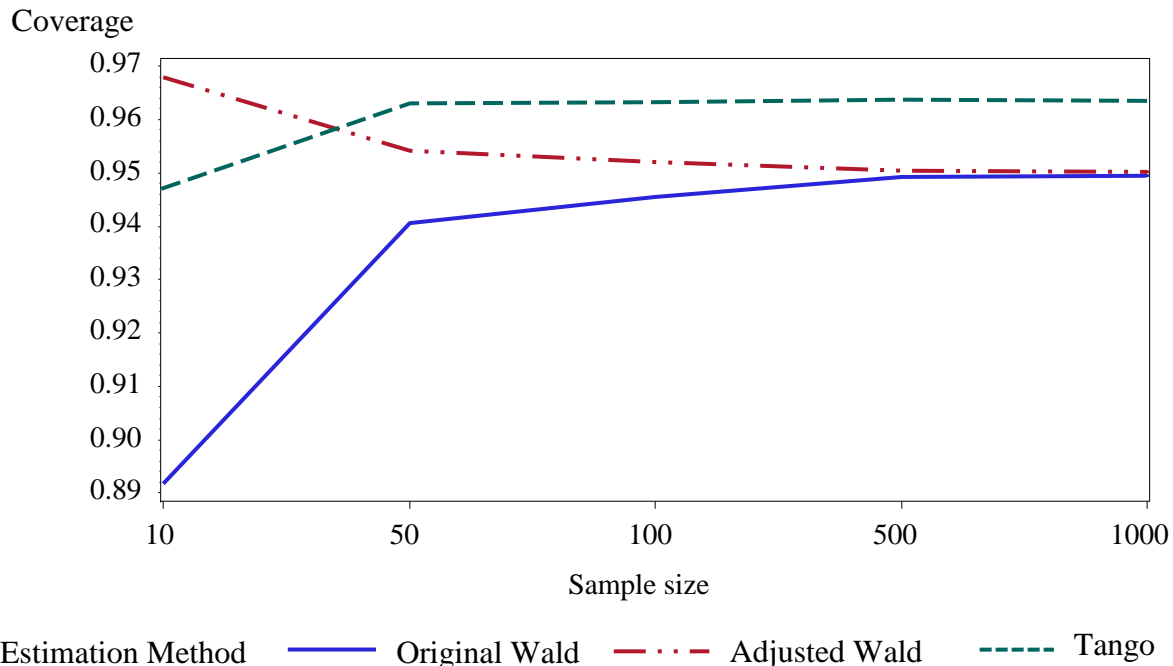

Figure 3. Mean Interval Coverage by Sample Size for 95% Confidence Level

CONCLUSION

The macros TANGO and CORR_P facilitate the computation of confidence intervals for correlated proportions, making the estimation of these intervals easy for applied researchers. The macros are written in BASE SAS and do not require additional components for execution. The macros may be used in the present form or may be easily modified. For example, the data set CI produced in the macro CORR_P may be used with SAS/GRAPH to graphically illustrate the obtained confidence intervals.

REFERENCES

Bonett, D. G., & Price, R. M. (2012). Adjusted wald confidence interval for a difference of binomial proportions based on paired data. *Journal of Educational and Behavioral Statistics, 37*(4), 479-488. doi: 10.3102/ 1076998611411915

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153-157.

Newcombe, R. G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine, 17*(22), 2635-2650. doi: 10.1002/(SICI)1097-0258(19981130)17:22<2635:: AID-SIM954>3.0.CO;2-C

Tango, T. (1999). Improved confidence intervals for the difference between binomial proportions based on paired data by robert G. newcombe, statistics in medicine, 17, 2635-2650 (1998). *Statistics in Medicine, 18*(24), 3511-3513. doi: 10.1002/(SICI)1097-0258(19991230)18:24<3511::AID-SIM303>3.0.CO;2-A

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine, 17*(8), 891-908. doi: 10.1002/(SICI)1097-0258(19980430)17:8<891::AID-SIM780>3.0.CO;2-B

CONTACT INFORMATION

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

Patricia Rodriguez de Gil
University of South Florida
4202 E. Fowler Ave., EDU 105
Tampa, FL 33620
Work Phone: 813 974-3220
Fax: 813 974-4495
prodrig6@usf.edu