Paper SP10-2009

Confidence Intervals for the Binomial Proportion with Zero Frequency

Xiaomin He, ICON Clinical Research, North Wales, PA
Shwu-Jen Wu, Biostatistical Consultant, Austin, TX

ABSTRACT

Estimating confidence interval for the binomial proportion is a challenge to statisticians and programmers when the proportion has zero frequency. The most widely used method based on Wald asymptotic statistics gives a degenerate interval, that is, (0, 0), in this case. This paper reviews the statistical methods used for estimating confidence intervals which are available in SAS version 9.2. In practice, when calculating the frequency and intervals, SAS by default does not present the missing categorical level; this level has zero frequency but is no less important than other categorical levels. This paper also builds a macro to share tips on how to create confidence intervals with zero frequency.

KEY WORDS

Binomial Proportion, Confidence Intervals, Zero Frequency, Wilson (Score) Confidence Interval, SAS Macro.

INTRODUCTION AND BACKGROUND

In a clinical trial, assume one observation has several levels and the proportion of observations in the first variable level is your primary interest. A binary response is a typical example, which has 0 (non-response) and 1 (response). Define n_1 as the frequency of the first (or designated) level and n as the total frequency of the one-way table. The binomial proportion is computed as

$$\hat{p} = n_1 / n \,.$$

Denote by $z_{\alpha/2}$ the $100(1-\alpha/2)$ th percentile of the standard normal distribution. Several methods to estimate the confidence interval for the binomial proportion (we focus on two-sided intervals here) are as follows:

- Wald asymptotic confidence interval:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right).$$

- Agresti-Coull confidence interval:

$$\left(\tilde{p} - z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/n},\ \tilde{p} + z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/n}\right),$$

  where $\tilde{n}_1 = n_1 + z_{\alpha/2}/2$, $\tilde{n} = n + z_{\alpha/2}^2$, and $\tilde{p} = \tilde{n}_1 / \tilde{n}$.

- Jeffreys confidence interval:

$$\left(\beta(\alpha/2, n_1 + 1/2, n - n_1 + 1/2),\ \beta(1-\alpha/2, n_1 + 1/2, n - n_1 + 1/2)\right),$$

  Where $\beta(\alpha, b, c)$ is α th percentile of the beta distribution with shape parameters b and c . The lower bound is set to 0 when $n_1 = 0$, and the upper bound is set to 1 when $n_1 = n$.

- Exact (Clopper-Pearson) confidence interval:

$$\left(\left(1 + \frac{n - n_1 + 1}{n_1 F(1-\alpha/2, 2n_1, 2(n - n_1 + 1))}\right)^{-1},\ \left(1 + \frac{n - n_1}{(n_1 + 1)F(\alpha/2, 2(n_1 + 1), 2(n - n_1))}\right)^{-1}\right),$$

  Where $F(\alpha, b, c)$ is α th percentile of the *F* distribution with b and c degrees of freedom. The lower bound is set to 0 when $n_1 = 0$, and the upper bound is set to 1 when $n_1 = n$.

- Wilson (score) confidence interval:

$$\left(\begin{array}{l} \hat{p} + z_{\alpha/2}^2/(2n) - z_{\alpha/2}\sqrt{\left(\hat{p}(1-\hat{p}) + z_{\alpha/2}^2\right)/(4n)}\Big/\left(1 + z_{\alpha/2}^2/n\right), \\ \quad \hat{p} + z_{\alpha/2}^2/(2n) + z_{\alpha/2}\sqrt{\left(\hat{p}(1-\hat{p}) + z_{\alpha/2}^2\right)/(4n)}\Big/\left(1 + z_{\alpha/2}^2/n\right) \end{array}\right).$$

The literature (see Brown, Cai & DasGupta (2001), Fleiss, Levin & Paik (2003) for details) has compared these confidence intervals (and more). The coverage probability of the interval covering the true binomial proportion is the one of criteria of comparison. Generally, two aberrations of two-sided confidence intervals estimators for the binomial proportion were considered: *overshoot* and *degeneracy*. For zero frequency, $n_1 = 0$, the simplest and most widely used Wald asymptotic gives a degenerate interval, that is, (0, 0), which has the poorest coverage probability among all of these methods. Note that a continuity correction, $1/(2n)$, was suggested to adjust for the difference between the normal approximation and the binomial distribution, which improves the Wald asymptotic interval in some respects but is still very inadequate. The Agresti-Coull confidence interval is another adjusted Wald asymptotic interval that adds 2 successes and 2 failures ($z_{\alpha/2}$ is close to 2 for $\alpha = 0.05$). Jefferys confidence interval is an equal-tailed interval based on noninformative Jeffreys prior to a binomial proportion. Exact (Clopper-Pearson) confidence interval is constructed by inverting the equal-tailed test based on the binomial distribution. Due to the discrete property of binomial distribution, the exact (Clopper-Pearson) confidence interval is not exactly $(1-\alpha)$ but is at least $(1-\alpha)$, so it is conservative. Wilson (score) confidence interval is constructed by inverting the normal test that uses the null proportion in the variance (the score test). The bounds are the roots of $| p - \hat{p} | = z_{\alpha/2}\sqrt{p(1-p)/n}$.

Except the Wald asymptotic method, all other four methods are recommended for calculating confidence intervals for the binomial proportions. In the case of binomial proportion with zero frequency, Agresti-Coull always gives the longest confidence interval, while Jeffreys gives the shortest. From the anti-conservative and coverage consideration standpoint, we would recommend using the Wilson (score) confidence interval because it has been shown to have better performance than the exact (Clopper-Pearson) confidence interval.

Prior to SAS 9.1.3, PROC FREQ procedure only provides the Wald asymptotic and exact (Clopper-Pearson) confidence intervals for the binomial proportion. In SAS 9.2, when you specify the BINOMIAL (ALL) option in the TABLES statement, then all of five confidence interval mentioned in this paper will be presented. You can also specify one or more types of binomial confidence intervals instead of ALL. The choices are AC (Agresti-Coull), EXACT (Clopper-Pearson), J (Jeffreys), W (Wilson score) and WALD (Wald asymptotic).

SAS APPLICATION

When using PROC FREQ to calculate the frequency and estimate confidence intervals, SAS by default doesn't include missing observations in the analysis. In this sense, the observations with zero frequency will be treated as missing and not presented in the output. However, as far as we know, observations with zero frequency are as important as other observations. A comprehensive summary including all categorical levels should be created.

We will use the following sample data set to illustrate, and how to avoid, the problem when deriving the confidence interval for the binomial proportion with zero frequency.

```
/* Group A has all binary levels of observations; Group B has the zero frequency */
/* at response=1; and Group C has zero frequency at response=0. */
data temp;
    do i = 1 to 20; group='A'; response=0; output; end;
    do i = 1 to 80; group='A'; response=1; output; end;
    do i = 1 to 100; group='B'; response=0; output; end;
    do i = 1 to 100; group='C'; response=1; output; end;
run;

ods select BinomialProp;
proc freq data=temp;
    by group;
    tables response / binomial;
run;
```

Below is the output based on the PROC FREQ statement above.

```
group=A

The FREQ Procedure

      Binomial Proportion for response = 0
_______________________________
```

```
Proportion                0.2000
ASE                       0.0400
95% Lower Conf Limit      0.1216
95% Upper Conf Limit      0.2784

Exact Conf Limits
95% Lower Conf Limit      0.1267
95% Upper Conf Limit      0.2918


group=B

The FREQ Procedure

      Binomial Proportion for response = 0
_____________________________________
Proportion                1.0000
ASE                       0.0000
95% Lower Conf Limit      1.0000
95% Upper Conf Limit      1.0000

Exact Conf Limits
95% Lower Conf Limit      0.9638
95% Upper Conf Limit      1.0000


group=C

The FREQ Procedure

      Binomial Proportion for response = 1
_____________________________________
Proportion                1.0000
ASE                       0.0000
95% Lower Conf Limit      1.0000
95% Upper Conf Limit      1.0000

Exact Conf Limits
95% Lower Conf Limit      0.9638
95% Upper Conf Limit      1.0000
```

Note that unlike Groups A and B, the binomial proportion for Group C was calculated for response=1 because there is 0 observation for response=0. SAS by default reports the binomial proportion in the first non-missing variable level; or you can specify the variable level to be calculated, but it must be non-missing. Therefore, the exact (Clopper-Pearson) confidence intervals for Groups B and C are "identical": (0.9638, 1.000). Also, SAS did not issue notes/warnings in the log, because there is not any algorithm error in programming. But this is not what we intended to derive.

In this paper, we share a tip about how to correctly calculate and present the confidence interval for the binomial proportion with zero frequency by using SAS. The algorithm is as follows:

- Suppose the discrete variables in the raw dataset include response, group and other stratification;
- Define a weight variable from the raw dataset. If the raw dataset doesn't have a weight variable, then generate one and assign all weights to be 1;
- Create a dummy dataset which contains all categorical levels of discrete variables;
- Merge the dummy dataset with the raw dataset by discrete variables;
- For the observation of discrete variables with zero frequency, the value of weight variable is missing. Hard-code it as 0, such as the COALESCE function in PROC SQL;
- When using PROC FREQ, add the ZEROS option in the WEIGHT statement;
- If necessary, specify the response variable level for which to compute the proportion.

Note that response could be not limited to a binary variable. It may have there or more categorical levels, which

makes our method be more generally and robustly used in practice. A SAS macro is shown here to illustrate the strategy:

```
/*------------------------------------------------------------------------------*/
/*- Macro Name:     CI_BP
/*- Description:    Calculate confidence intervals for the binomial proportion of
/*-                 response variables.
/*- SAS Version:    9.2
/*- Data Input:     indata   = The input data. (Required)
/*- Data Output:    None or user defined
/*- Parameters:     group    = Grouping or block variable. (Required)
/*-                 response = Response or event variable for which to compute the
/*-                        proportion. (Required)
/*-                 weight   = Frequency or weight variable. If it is not specified,
/*-                        the default value is 1. (Optional)
/*-                 by       = Stratification variable. (Optional)
/*-                 level    = Formatted value of the response level in which the
/*-                        confidence interval is calculated. By default the confidence
/*-                        interval is for the first response level. (Optional)
/*-                 method   = Method used for calculating confidence intervals.
/*-                        The default value is ALL, which requests all types of
/*-                        confidence intervals introduced in the paper. The
/*-                        alternatives are AC(Agresti-Coull), EXACT (Clopper-Pearson),
/*-                        J(Jeffreys), WALD(Wald), and W (Wilson score) confidence
/*-                        intervals. (Optional)
/*------------------------------------------------------------------------------*/
%macro CI_BP(indata=, group=, response=, weight=1, by=, level=, method=ALL);
proc sql noprint;
  create table indata_ as
    select %if %length(&by) > 0 %then %do; &by as by, %end;
    &group as group, &response as response, &weight as weight
    from &indata;

  %if %length(&by) > 0 %then %do;
  create table by_ as
    select distinct(&by) as by from &indata;
  %end;

  create table group_ as
    select distinct(&group) as group from &indata;

  create table response_ as
    select distinct(&response) as response from &indata;

  create table dummy_ as
    select %if %length(&by) > 0 %then %do; a.by, %end; b.group, c.response
    from %if %length(&by) > 0 %then %do; by_ as a, %end;
         group_ as b, response_ as c;

  create table outdata_ as
    select %if %length(&by) > 0 %then %do; a.by, %end;
           a.group, a.response, coalesce(b.weight, 0) as weight
    from dummy_ as a left join indata_ as b
    on %if %length(&by) > 0 %then %do; a.by=b.by and %end;
       a.group=b.group and a.response=b.response;
quit;

proc freq data=outdata_;
  by %if %length(&by) > 0 %then %do; by %end; group;
  tables response / binomial (&method
   %if &level= %then level=1; %else level="&level";);
  weight weight /zeros;
run;
%mend CI_BP;

%CI_BP(indata=temp, group=group, response=response);
```

Note that the macro was written on SAS 9.2. In SAS 9.0 and 9.1, the options after BINOMIAL in the TABLES statement should be removed; in that case only the Wald Asymptotic and exact (Clopper-Pearson) confidence intervals were calculated. For SAS version prior to 9.0, the algorithm doesn't work because ZEROS option in

WEIGHT statement is unavailable. Therefore, algorithm of using 1 minus the confidence interval of total compensative categorical levels is a more reasonable choice. We didn't build the macro using for early SAS versions due to limited use.

You may specify any formatted value of the response level in which the confidence interval is calculated; and the macro can also be applied to one more stratification variable other than group variable. A more general example is as follows:

```
data temp2;
  do i=1 to 20; b=1; g='A'; r=0; output; end;
  do i=1 to 40; b=1; g='A'; r=1; output; end;
  do i=1 to 40; b=1; g='A'; r=2; output; end;
  do i=1 to 100; b=1; g='B'; r=0; output; end;
  do i=1 to 30; b=1; g='C'; r=1; output; end;
  do i=1 to 70; b=1; g='C'; r=2; output; end;
  do i=1 to 80; b=2; g='A'; r=0; output; end;
  do i=1 to 20; b=2; g='A'; r=1; output; end;
  do i=1 to 90; b=2; g='B'; r=0; output; end;
  do i=1 to 10; b=2; g='B'; r=2; output; end;
  do i=1 to 50; b=2; g='C'; r=1; output; end;
  do i=1 to 50; b=2; g='C'; r=2; output; end;
run;

* Calculate the exact confidence interval for the proportion of strongest response;
%CI_BP(indata=temp2, group=g, response=r, by=b, level=2, method=EXACT);
```

CONCLUSION

This paper gives an overview on the statistical methods used for estimating confidence intervals, which are all available in SAS 9.2. We recommend using the Wilson (score) confidence interval from the consideration of anti-conservative and coverage when the binomial proportions have zero frequency. We also notice that confidence intervals generated from the PROC FREQ procedure by default have unexpected results for the binomial proportions with zero frequency. By creating a dummy dataset with all categorical levels and hard-coding the missing (or of zero frequency) observations, we build a macro to share the tips on how to create confidence intervals under this situation.

REFERENCES

SAS/STAT® 9.2 User's Guide, The FREQ Procedure.

Agresti ,A. and Coull ,B.A.(1998), "Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions", The American Statistician ,52,119–126.

Clopper,C.J.,and Pearson,E.S.(1934),"The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial", Biometrika 26, 404–413.

Collett ,D.(1991), Modelling Binary Data, London: Chapman & Hall.

Wilson, E.B.(1927), "Probable Inference, the Law of Succession, and Statistical Inference", Journal of the American Statistical Association, 22, 209–212.

Brown, L.D., Cai, T.T., & DasGupta, A. (2001). "Confidence intervals for a binomial proportion (with discussion)", Statistical Science, 16, 101–133.

Fleiss, J.L., Levin, B., and Paik, M.C. (2003), Statistical Methods for Rates and Proportions, Third Edition, New York: John Wiley & Sons.

ACKNOWLEDGMENTS

We are grateful to Dr. Liz Morgenthien and Hima Bhatia from ICON Clinical Research and Robert Schechter from Octagon Research Solutions for their valuable suggestions and reviewing of this paper.

CONTACT INFORMATION

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

Name: Xiaomin He, Ph.D.
Enterprise: ICON Clinical Research
Address: 1700 Pennbrook Parkway
City, State ZIP: North Wales, PA 19454
Work Phone: 215-616-6406
Fax: 215-616-8685
E-mail: Xiaomin.he@iconplc.com
Web: www.iconplc.com