PharmaSUG 2015 - Paper SP06

Confidence Intervals Are a Programmer's Friend

Xinxin Guo, Quintiles, Cambridge, MA
Zhaohui Su, Quintiles, Cambridge, MA

ABSTRACT

A confidence interval (CI) is a type of interval estimate of a population parameter and is one of the most common terms statistical programmers face in everyday practice. This paper will present a collection of SAS code to calculate the CI's of a proportion obtained by PROC FREQ (special handling for category with zero count is also considered) based on the assumption that proportion follows a Binomial distribution, and the CI of incidence rate obtained by a handy formula based on the assumption that the number of events occurring in a fixed interval of time follows a Poisson distribution. The paper will also look at PROC GENMOD and PROC PLM and compare the results from these two procedures against the code given.

BRIEF INTRODUCTION ABOUT CONFIDENCE INTERVALS

Confidence Intervals (CIs) can be made around almost any statistics calculated, such as CIs for data summaries, i.e. mean for continuous variable, proportion for categorical variable; and CIs for those parameters resulted from hypothesis testing, such as correlation, regression slope, relative risk, and odds ratio etc.

This paper will focus on CIs for a proportion and for incidence rate.

CONFIDENCE INTERVAL FOR A PROPORTION

There are numerous methods available for constructing a Binomial confidence interval. For this paper discussion we will focus on Clopper-Pearson method, also known as exact method.

Here we begin our story with an example. We are observing a sample of 145 subjects for medication effect. It turned out that 86 of them showed effect of this medication. The proportion of effect could be calculated as 0.59=86/145. This is informative. However, we also need to know its CI, as how this proportion can be generalized to reflect the population parameter.

USING PROC FREQ TO OBTAIN EXACT CONFIDENCE INTERVAL FOR A PROPORTION

With the built dataset one, and some simple SAS code, we can obtain exact confidence interval for this proportion as (0.5085, 0.6738) by executing SAS syntax below.

```
data one;
 input grp response count;

 datalines;
 1 1 61
 1 2 39
 2 1 25
 2 2 15
 3 1  0
 3 2  5
 ;
run;

proc freq data=one;
  tables response/binomial (exact) alpha=0.05 missprint;
  weight count;
run;
```

The dataset and corresponding output looks like this:

	grp	response	count
1	1	1	61
2	1	2	39
3	2	1	25
4	2	2	15
5	3	1	0
6	3	2	5

```
                                Cumulative    Cumulative
response    Frequency     Percent        Frequency      Percent
---------------------------------------------------------------
       1           86       59.31             86        59.31
       2           59       40.69            145       100.00


                    Binomial Proportion
                      for response = 1
                    ----------------------
                    Proportion      0.5931
                    ASE             0.0408

             Type                     95% Confidence Limits

             Clopper-Pearson (Exact)       0.5085    0.6738
```

What we would like to discuss more is that for some cases where the frequency could be zero, for example if we run the same SAS codes for group 3 within data one.

```
proc freq data=one;where grp=3;
  tables response/binomial (exact) alpha=0.05 missprint;
  weight count;
run;
```

```
                                      Cumulative    Cumulative
response    Frequency     Percent     Frequency      Percent
-----------------------------------------------------------
       2           5      100.00             5       100.00


                    Binomial Proportion
                      for response = 2
                    ----------------------
                    Proportion      1.0000
                    ASE             0.0000

             Type                     95% Confidence Limits

             Clopper-Pearson (Exact)       0.4782    1.0000
```

SAS did not show frequency level for response 1 (being efficient). In this case, is it safe to report confidence interval as zero? Many papers showed it's inappropriate to show population confidence interval as zero when observed zero frequency from a study sample.

Here is a simple tweak I would like to present to modify dataset ONE a bit and with using an extra WEIGHT statement, PROC FREQ will be able to display confidence interval for zero frequency category.

```
data two;
 set one;
 if count=0 then wgt=0; else wgt=1;
run;

proc freq data=two;where grp=3;
  tables response/binomial (exact) alpha=0.05 missprint;
  weight count;
  weight wgt/zeroes;
run;
```

	grp	response	count	wgt
1	1	1	61	1
2	1	2	39	1
3	2	1	25	1
4	2	2	15	1
5	3	1	0	0
6	3	2	5	1

Basically after a simple make-up to dataset ONE, by adding an indicator of WGT as zero to zero count category, and an extra WEIGHT statement in PROC FREQ, SAS displayed zero count category and it's confidence interval level as (0.0, 0.5218).

```
                                          Cumulative    Cumulative

response    Frequency     Percent     Frequency      Percent
-----------------------------------------------------------------
       1           0        0.00             0         0.00
       2           5      100.00             5       100.00


                        Binomial Proportion
                          for response = 1
                        ---------------------------------
                        Proportion      0.0000
                        ASE             0.0000

                Type                     95% Confidence Limits

                Clopper-Pearson (Exact)       0.0000    0.5218
```

More Confidence Intervals from Proc Freq

With using the same dataset ONE above, Proc Freq actually can give a full-range of Confidence Intervals by specifying "ALL" in tables statement.

```
proc freq data=one;
  tables response/binomial (ALL) alpha=0.05 missprint;
  weight count;
run;
```

The output looks like this:

```
                         Binomial Proportion
                           for response = 1
                         ---------------------------------
                         Proportion      0.5931
                         ASE             0.0408

                Type                     95% Confidence Limits
                Wald                          0.5131    0.6731
                Wilson                        0.5117    0.6697
                Agresti-Coull                 0.5117    0.6697
                Jeffreys                      0.5120    0.6706
                Clopper-Pearson (Exact)       0.5085    0.6738
```

```
Test of H0: Proportion = 0.5

ASE under H0              0.0415
Z                         2.2422
One-sided Pr >  Z         0.0125
Two-sided Pr > |Z|        0.0249

Sample Size = 145
```

The focus for this paper however is not talking about theoretical difference among all those methods, it is very helpful for statistical programmers to know the basic application for each method. It's said that coverage probability, conservatism and interval width are critical in evaluating such competing methods.

Wald interval was based on the normal approximation to the binomial distribution. It employees the simple asymptotic method. For application, only if sample size is greater than 30 and proportion is not too close to 0 or 1, Wald intervals are close to true proportion. In fact, many literatures suggested not using Wald intervals for scientific publications due to its anti-conservative nature and poor coverage probability.

Wilson Score Interval is based on inverting the normal test. This one has a good property to sample with small size or extreme proportions. Many reviews suggest this method performs better than Wald test and Clopper-Pearson test due to its high conservatism and average coverage probability very close to the nominal value.

Agresti-Coull is an adjusted form of Wald intervals.

Jeffreys Interval is a derivation of Bayesian Credit Interval. The method is based on Beta distribution and having good coverage probability similar to Wilson Score test with the advantage of being equal-tailed.

Clopper-Pearson intervals for binomial proportion were produced by using F-distribution. Due to its restrict conservatism, this method has been regarded as 'Golden standard'.

The decision towards to method selection should be made carefully with study questions, study population and statistical theory fitting.

Confidence Intervals for Incidence Rates

The incidence rate is estimated as the number of events observed divided by the time at risk of event during the observation period. The confidence interval of the incidence rate can be estimated based on the assumption that the number of events occurring in a fixed interval of time follows a Poisson distribution. Below we present four different methods to achieve such goals.

Say that 14 events are observed in 200 people studied for 1 year and 100 people studied for 2 years.
The person time at risk is 200 + 100 x 2 = 400 person years.

1. With PROC GENMOD

We use such SAS codes to build this dataset and see how PROC GENMOD can help us.

```
data x_trans;
   infile cards;
   input n c;
   ln=log(n);
   dummy=1;
cards;
400 14
;
run;

*Standard usage, proc genmod;
proc genmod data=x_trans;
   model c = / offset=ln dist=poisson lrci;
   estimate 'Mean' intercept 1 ;
run;
```

Contrast Estimate Results										
		Mean					L'Beta			
Label	Mean Estimate	Confidence Limits		L'Beta Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
Mean	0.0350	0.0207	0.0591	-3.3524	0.2673	0.05	-3.8762	-2.8286	157.34	<.0001

We got IR as 0.035 and confidence interval as (0.0207, 0.0591) from Proc Genmod.

2. A ready formula for exact CI validated by using EpiSheet.

The formula was developed based on the formula below.

$$Y_l = \frac{\chi^2_{2Y,\alpha/2}}{2}$$

$$Y_u = \frac{\chi^2_{2(Y+1),1-\alpha/2}}{2}$$

Where Y is the observed number of events, Y_i and Y_u are lower and upper confidence limits for Y respectively, $\chi 2v,a$ is the chi-square quantile for upper tail probability on v degrees of freedom.

```
data yourcode;
   set x_trans;
   IR=c/n;
   LCI=quantile('chisq',0.025, c*2)/(n*2);
   UCI=quantile('chisq',0.975, (c+1)*2)/(n*2);
Run;

proc print data=yourcode;
   id n c;
run;
```

n	c	ln	dummy	IR	LCI	UCI
400	14	5.99146	1	0.035	0.019135	0.058724

We got IR as 0.035 and confidence interval as (0.0191, 0.0587) from our handy formula.

3. With PROC FREQ

Can we use PROC FREQ to obtain such incidence intervals? Yes, let's build the dataset in a creative way.

```
data x;
   infile cards;
   input grp count;
cards;
1 14
2 386
;
proc freq data=x;
   tables grp/binomial (all) alpha=0.05 missprint;
   weight count;
run;
```

The output is presented here:

Binomial Proportion	
grp = 1	
Proportion	0.0350
ASE	0.0092

Confidence Limits for the Binomial Proportion		
Proportion = 0.0350		
Type	95% Confidence Limits	
Agresti-Coull	0.0204	0.0584
Clopper-Pearson (Exact)	0.0193	0.0580
Jeffreys	0.0202	0.0565
Wald	0.0170	0.0530
Wilson	0.0210	0.0579

The IR is 0.035 is the same as two other methods. Among bunch of CIs provided, all of them fall in cluster except Wald intervals due to the reason stated earlier that it is not a good fit for rare event.

4. With PROC PLM and STORE statement in PROC GENMOD

PROC PLM was experimental in SAS 9.2 and production from 9.3 onwards. PROC PLM can deal with post-modeling handling for many procedures without re-executing the model. Here we give an example on how PROC PLM is able to get us the rates needed.

```
*From Usage Note 24188, method 1;
proc genmod data=x_trans;
   class dummy;
   model c = dummy / dist=poisson link=log offset=ln;
   store out=insmodel;
run;
proc plm source=insmodel;
   score data=x_trans out=inspred pred stderr lclm uclm / nooffset ilink;
proc print label;
   id n c;
run;
```

The converted model stored a dataset called "insmodel" through "store" statement while running PROC GENMOD, and PROC PLM could exact all parameter information for post-modeling analysis.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.3524	0.2673	-3.8762	-2.8286	157.34	<.0001
dummy	1	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Please note, there is no any contrast statement in PROC GENMOD. And we are able to read the rate and CIs via PROC PLM below.

n	c	ln	dummy	Predicted Value	Standard Error	Lower 95% Confidence Limit	Upper 95% Confidence Limit
400	14	5.99146	1	0.035	.009354143	0.020729	0.059096

5. With OUTPUT statement in PROC GENMOD

We can also look at another example on how to obtain the rate and CIs via output statement.

```
*From Usage Note 24188, method 2;
proc genmod data=x_trans;
   class dummy;
   model c = / dist=poisson link=log offset=ln;
   output out=out p=pcount xbeta=xb stdxbeta=std;
run;

data predrates;
   set out;
   obsrate=c/n;          /* observed rate */
   lograte=xb-ln;
   prate=exp(lograte);  /* predicted rate */
   lcl=exp(lograte-probit(.975)*std);
   ucl=exp(lograte+probit(.975)*std);
   keep n c dummy prate lcl ucl;
proc print data=predrates;
   id n c;
run;
```

And the results matches what from PROC PLM.

n	c	dummy	prate	lcl	ucl
400	14	1	0.035	0.020729	0.059096

CONCLUSION

This paper intended to present some sample codes for programmers and provide some practical references. Besides the regular use of PROC FREQ and PROC GENMOD for developing CIs for binomial proportion, this paper also compared different methods given by PROC FREQ, and presented multiple ways through a handy formula, PROC PLM with STORE statement from PROC GENMOD, and OUTPUT statement from PROC GENMOD to obtain CIs for incidence rate. Hopefully you find it useful to your programming practice.

REFERENCES

Cody, Ronald P, and Smith Jefferey K (1997), Applied Statistics and the SAS® Programming Language, Fourth Edition, Prentice-Hall Inc,.

Attain 100% Confidence in Your 95% Confidence Interval, Indu Nair, Binal Patel, PharmaSUG 2014 - Paper IB05

Peter Langlois, "BOOST YOUR CONFIDENCE (INTERVALS) WITH SAS". http://www.scsug.org/SCSUGProceedings/2010/Langlois/Langlois_conf_intervals_for_SCSUG_Educ_Forum.pdf ,

Newcombe, Robert (1998). "Two-sided confidence intervals for the single proportion: Comparison of seven methods." Statistics in Medicine 17 (2, issue 8): 857–872.

Ken Rothman's Episheet. The spreadsheet for Epidemiologic data. http://krothman.hostbyet2.com/episheet.xls

ACKNOWLEDGMENTS

My sincere appreciation goes to David Franklin, Manager of Statistical Programming at RWLPR of Quintiles who performed a systemic review and gave constructive ideas in shaping this paper. I am also thankful for Jianjun Liu, Senior Statistician at RWLPR of Quintiles who provided meaningful edits.

CONTACT INFORMATION

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

Name: Xinxin Guo
Enterprise: Quintiles Inc.
Address: 201 Broadway
City, State ZIP: Cambridge, MA 02132
Work Phone: 617-715-6840
E-mail: xinxin.guo@quintiles.com