Are you computing Confidence Interval for binomial proportion using PROC FREQ? Be Careful!!

Sandeep Sawant, i3 Statprobe, Mumbai, MH
Gary Allen, i3 Statprobe, Ann Arbor, MI

ABSTRACT

PROC FREQ is the most commonly used procedure for the analysis of categorical data. However, in some situations output generated by this procedure needs special attention. One such case is, computing the confidence interval for responders using binomial proportion. In many situations it is possible that the data do not contain any responders, however, the summary table still needs the confidence interval for responder.

Does PROC FREQ calculate the confidence interval for responders when the data has only non-responders? No, currently there is no functionality in PROC FREQ to handle this situation.

This paper will discuss the limitations of PROC FREQ for the above situation and provides the solution for the same.

INTRODUCTION

In oncology trials, we find sometimes that subjects do not achieve complete response (Disappearance of all evidence of disease) or partial response (Regression of measurable disease and no new sites) during the trial. We often need to provide the computation of 95% CI for the proportion of subjects achieving complete response or partial response.

In this note, we will be explaining how Proc Freq can be misleading when we come across a situation where in neither subject provides partial nor complete response. This fact is depicted through discussions of two examples, for the purpose of comparison, the first one being discussing a case where there are complete responses from two subjects and, the second example being of no partial responses from any of the subjects recruited.

CASE 1: Let us consider a special case where in two subjects in the trial provide complete responses. The table shall will typically look like the one described below.

<table>
<thead>
<tr>
<th>Patients with:</th>
<th>Trt A (N=xx)</th>
<th>Trt B (N=xx)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Response</td>
<td>n (%)</td>
<td>xx (xx.x%)</td>
</tr>
<tr>
<td></td>
<td>95% CI [1]</td>
<td>xx.x%, xx.x%</td>
</tr>
<tr>
<td>Partial Response</td>
<td>n (%)</td>
<td>xx (xx.x%)</td>
</tr>
<tr>
<td></td>
<td>95% CI [1]</td>
<td>xx.x%, xx.x%</td>
</tr>
</tbody>
</table>

[1] CI computed using Exact Binomial Test.

Consider the data as follows.

DATA

Following is the data for 10 subjects who were randomly assigned to treatment A or B. Subjects 3 and 4 achieve the complete response while no subject attains the partial response; the corresponding binary flags (1=Yes, 2=No) are presented below;
SAS CODE

The SAS code to get the required count, percentages and corresponding 95% CI for the subjects achieving complete response is as follows:

```sas
PROC FREQ DATA=status ;
   BY trt;
   TABLES cr/OUT=cnt binomial;
   EXACT binomial;
   ODS OUTPUT binomialprop=bin(where=(name1 in ('XL_BIN', 'XU_BIN')));
RUN;
```

The required count and percentages are stored in the data set CNT. The required 95% CI are stored in the data set BIN.

**OUTPUT DATA SET CNT:**

<table>
<thead>
<tr>
<th>trt</th>
<th>cr</th>
<th>Frequency Count</th>
<th>Percent of Total Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 a</td>
<td>1</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>2 a</td>
<td>2</td>
<td>4</td>
<td>80</td>
</tr>
<tr>
<td>3 b</td>
<td>1</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>4 b</td>
<td>2</td>
<td>4</td>
<td>80</td>
</tr>
</tbody>
</table>

So from the above data set we can conclude that 1 subject from treatment A achieve complete response and 1 subject from treatment B achieve complete response.

**OUTPUT DATA SET BIN:**

<table>
<thead>
<tr>
<th>trt</th>
<th>Table</th>
<th>Name</th>
<th>Label</th>
<th>cValue</th>
<th>nValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 a Table cr</td>
<td>XL_BIN</td>
<td>95% Lower Conf Limit</td>
<td>0.0051</td>
<td>0.05051</td>
<td></td>
</tr>
<tr>
<td>2 a Table cr</td>
<td>XU_BIN</td>
<td>95% Upper Conf Limit</td>
<td>0.7164</td>
<td>0.716418</td>
<td></td>
</tr>
<tr>
<td>3 b Table cr</td>
<td>XL_BIN</td>
<td>95% Lower Conf Limit</td>
<td>0.0051</td>
<td>0.05051</td>
<td></td>
</tr>
<tr>
<td>4 b Table cr</td>
<td>XU_BIN</td>
<td>95% Upper Conf Limit</td>
<td>0.7164</td>
<td>0.716418</td>
<td></td>
</tr>
</tbody>
</table>

PROC FREQ always computes the CI for the lowest level of the responder variable (Ref: SAS OnlineDoc, V8). The lowest level for variable CR is 1 i.e responder so corresponding 95% CI for proportion of subjects achieving complete response for treatment A and B are (0.51%, 71.64%) and (0.51%, 71.64%) respectively.
**CASE 2:** None of the subjects in the trial provide partial responses.

**SAS CODE:**

```
PROC FREQ DATA=status ;
   BY trt;
   TABLES pr/OUT=cnt binomial;
   EXACT binomial;
   ODS OUTPUT binomialprop=bin(where=(name1 in ('XL_BIN', 'XU_BIN')));
RUN;
```

**OUTPUT DATA SET CNT:**

<table>
<thead>
<tr>
<th>trt</th>
<th>pr</th>
<th>Frequency Count</th>
<th>Percent of Total Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

Since none of the subjects attains partial response, values under PR column are 2 which mean no subjects have achieved partial responses in both the treatment arms.

**OUTPUT DATA SET BIN:**

<table>
<thead>
<tr>
<th>trt</th>
<th>Table</th>
<th>Name1</th>
<th>Label1</th>
<th>cvale1</th>
<th>n/value1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>pr</td>
<td>XL_BIN</td>
<td>95% Lower Conf Limit</td>
<td>0.4782</td>
<td>0.479176</td>
</tr>
<tr>
<td>2</td>
<td>pr</td>
<td>XU_BIN</td>
<td>95% Upper Conf Limit</td>
<td>1.0000</td>
<td>1.003000</td>
</tr>
<tr>
<td>3</td>
<td>pr</td>
<td>XL_BIN</td>
<td>95% Lower Conf Limit</td>
<td>0.4782</td>
<td>0.479176</td>
</tr>
<tr>
<td>4</td>
<td>pr</td>
<td>XU_BIN</td>
<td>95% Upper Conf Limit</td>
<td>1.0000</td>
<td>1.003000</td>
</tr>
</tbody>
</table>

Remember that the CI always gets computed for the lowest level of responder variable (Ref: SAS OnlineDoc V8). Therefore, 95% CI gets computed for proportion of subjects NOT achieving partial response as none of the subjects achieve partial response.

When the output data set doesn’t contain any variable to identify for which level of the variable these CI’s are computed there is higher probability/chance that one considers the CI calculated by Proc Freq as the one for Partial Respondents where it is actually NOT. One needs to be really careful while using this feature of Proc Freq.

We use the algorithm used by Proc Freq to compute CI in this set up (The algorithm essentially computes CI for the lowest level of the variable). In our example mentioned previously, we show how the level identifier/variable can be added in the BIN data set by using the following simple data steps.
PROC SORT DATA= cnt;
   BY trt pr;
RUN;

DATA cntx;
   SET cnt;
   BY trt pr;
   IF first.trt;
   KEEP trt pr;
RUN;

PROC SORT DATA=bin;
   BY trt;
RUN;

DATA comb;
   MERGE bin cntx;
   BY trt;
RUN;

DATA SET COMB:

<table>
<thead>
<tr>
<th>trt</th>
<th>Table</th>
<th>Name</th>
<th>Label 1</th>
<th>c.Value</th>
<th>n/EXacted</th>
<th>#r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>Table pr</td>
<td>1-L_BIN</td>
<td>0.4782</td>
<td>0.478766</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>Table pr</td>
<td>1-U_BIN</td>
<td>1.0000</td>
<td>1.000000</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>Table pr</td>
<td>1-L_BIN</td>
<td>0.4782</td>
<td>0.478766</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>b</td>
<td>Table pr</td>
<td>1-U_BIN</td>
<td>1.0000</td>
<td>1.000000</td>
<td>2</td>
</tr>
</tbody>
</table>

Observe that, the COMB data set now contains the identifier variable. It facilitates to conclude that the CI’s are computed for non-responder i.e. for proportion of subjects not-achieving the partial response.

The next step is to get the CI for responders using the CI of non-responders. To calculate the CI for the responder we follow the logic as described below:

p=proportion of subjects achieving partial response
q= proportion of subjects not achieving partial response
p and q being the proportions satisfies the equation  \( p + q = 1 \) \( \cdots \cdots \) (\(^*\))

Since we have the CI for non-responders that is for \( q \), probability that the value of \( q \) lies between upper and lower limit is 1-alpha, where alpha being the level of significance.

\[ \text{Prob}( \text{LCL} < q < \text{UCL} ) = 1 - \alpha \]
\[ \therefore \text{Prob}( -\text{LCL} > -q > -\text{UCL} ) = 1 - \alpha \]
\[ \therefore \text{Prob}( 1-\text{LCL} > 1-q > 1-\text{UCL} ) = 1 - \alpha \]
\[ \therefore \text{Prob}( 1-\text{UCL}<p<1-\text{LCL} ) = 1 - \alpha \] \[\cdots \text{using } \cdots(\ast)\]
From the above equation, it is clear that $1-UCL$ is the lower limit and $1-LCL$ is the upper limit for the proportion of responder where LCL and UCL are the lower and upper limit of non-responders, respectively.

Let's apply this logic to COMB data set.

```plaintext
DATA final;
  SET comb;
  if pr=2 then do;
    pr=pr-1;
    nvalue2=1-nvalue1;
    if name1='XL_BIN' then do;
      name2='XU_BIN';
      label2='95% Upper Conf Limit';
    end;
    else if name1='XU_BIN' then do;
      name2='XL_BIN';
      label2='95% Lower Conf Limit';
    end;
  end;
  else if pr=1 then do;
    name2=name1;
    label2=label1;
    nvalue2=nvalue1;
  end;
  DROP name1 label1 nvalue1 cvalue1;
RUN;
```

<table>
<thead>
<tr>
<th>b</th>
<th>Table</th>
<th>pr</th>
<th>nvalue2</th>
<th>name2</th>
<th>label2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>1</td>
<td>0.5201757601</td>
<td>XU_BIN</td>
<td>95% Upper Conf Limit</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>1</td>
<td>0.5201757601</td>
<td>XL_BIN</td>
<td>95% Lower Conf Limit</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>1</td>
<td>0.5201757601</td>
<td>XU_BIN</td>
<td>95% Upper Conf Limit</td>
</tr>
<tr>
<td>4</td>
<td>b</td>
<td>1</td>
<td>0.5201757601</td>
<td>XL_BIN</td>
<td>95% Lower Conf Limit</td>
</tr>
</tbody>
</table>

Using the above code, we get the CI for proportion of subjects achieving partial response which is (0.00%, 52.18%) and (0.00%, 52.18%) for treatment A and B respectively.
CONCLUSION

If Proc Freq would have facilitated the lowest level of variable as an output while generating CI for the same variable, a major source of mistake could have been avoided. This paper identifies this gap and proposes a solution.

ACKNOWLEDGMENTS

I thank Dr. Prashant Kirkire, Sr. Director of Biostatistics and Statistical Programming, i3 Statprobe for guiding this work and carefully reviewing the paper with comments and suggestions.

CONTACT INFORMATION

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

Sandeep Sawant
i3 Statprobe, 7th Floor, Corporate Center,
Opp. To VITS Hotel,
Andheri-Kurla Road,
Andheri (E)- 400059
Mumbai, India
Work Phone: +91-22-30554032
E- mail: sandeep.sawant@i3global.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.