Paper AS16

Comparing Confidence Intervals for a Difference in Adjusted Predicted Probabilities

Cralen Davis, Cytel Inc., Massachusetts, USA
Trevano Dean, Cytel Inc., Massachusetts, USA

ABSTRACT

As a statistical method, calculating adjusted predicted probabilities (APPs) is a common and straight-forward analysis when modeling a binary outcome. The procedure, PROC LOGISTIC (PL) within the analysis software package, SAS® (Cary, NC) is one of the most used approaches to accomplish this task. However, when a goal is to compute a valid confidence interval (CI) of the difference between mean APPs for 2 groups, the analysis is less straight-forward; this is evident by the fact that a CI for the difference of mean APPs is not directly provided within SAS. For a given dataset with 2 groups and a binary outcome, we used the estimates provided by PL and computed the difference between mean APPs for the 2 groups; computed a 95% CI under asymptotic normal approximation (ANA); and compared it to 95% CIs constructed using bootstrap sampling. We conclude that ANA is a practical approach that can be valid and perform adequately as a method for calculating CIs on differences between mean APPs for 2 groups.

Keywords: predicted probabilities, logistic regression, confidence interval, bootstrap sampling, bootstrapping, asymptotic normal approximation

INTRODUCTION

In the world of research, data, and statistical analysis, one may face a problem that requires a solution which falls outside of the norm. A real-life example of this situation served as the inspiration for the work presented in this paper. For a given clinical trial or any other data-related study where it may be appropriate to calculate adjusted predicted probabilities (APPs), one may have a hypothesis or research question that requires evaluating a confidence interval (CI) for the difference between mean APPs for 2 groups. While there are multiple approaches and software packages available to calculate APPs, getting a difference between mean APPs and a CI for that difference is not as straight-forward. The work presented here is an approach to compute the difference between mean APPs for 2 groups and to calculate and compare different methods for creating a CI for the difference. Given the familiarity and practicality of asymptotic normal approximation (ANA), it was the primary method applied; and an empirical bootstrap sampling method was employed for comparison and validation.

ANALYSIS APPROACH

The main analysis used for the paper models a discreet binary outcome to produce adjusted predicted probabilities. There are multiple approaches and software packages available to perform such an analysis. For generality, a very common approach and software was selected for this analysis; the logistic regression modeling procedure (PROC LOGISTIC) and the SAS® software package (Cary, NC).

PROC LOGISTIC (PL) within SAS is a well-known tool for generating group mean APPs and confidence intervals. While PL provides group mean APPs, it does not directly provide a pair-wise difference in mean APPs nor any CI for that mean. What is given by the procedure are estimates of the group mean APPs and the corresponding standard error (SE) for each mean. In this analysis, the given estimates are used to calculate a pair-wise difference in mean APPs ($p_{diff} = p_B - p_A$) and an asymptotic normal approximated 95% CI. The CI is based on p_{diff} and the pooled variance of p_B and p_A ($Var_{pooled} = SE_B^2 + SE_A^2$). The ANA based 95% CI is calculated as follows:

$$95\%\ CI = p_{diff} \pm 1.96 \times \sqrt{Var_{pooled}}$$

An empirical method was selected as the alternative approach for generating 95% CIs. Bootstrap sampling or bootstrapping was the selected empirical method used. The approach involved generating replicate datasets of the same size as the original dataset by random sampling records, with replacement, from the original dataset. The same PL model run on the original dataset is run on each replicate dataset. Using the mean APPs provided by PL, p_{diff} is calculated for each model. Therefore, k replicated datasets result in k calculated values of p_{diff}. A 95% confidence interval is then created by ordering the k p_{diff} values from smallest to largest and selecting the 2.5th percentile and 97.5th percentile values as the lower and upper confidence interval bounds respectively. For example, at k = 100, the 2nd and 3rd values among the 100 total smallest to largest ordered

p_{diff} values would be selected and averaged for the lower bound of the 95% CI and the 97th and 98th values would be selected and averaged for the upper bound such that the 95% CI = ({$p_{diff2}/2 + p_{diff3}/2$},{$p_{diff97}/2 + p_{diff98}/2$}). A total of three 95% CIs were created using bootstrap sampling, one each at *k* = 100, 500, and 1000.

DATASET DETAILS

The data used in this analysis is a de-identified simulated dataset with 263 rows representing 263 unique subjects who have been randomized 1:1 into 2 treatment groups, A and B. There are 5 total variables in the dataset: ID (unique numeric identifier); TRT_GRP (Group A/Group B); AGE (numeric); SEX (M/F); and Severe_Event (0/1). The outcome variable is Severe_Event which equals 1 if a subject experienced any severe event over the treatment period and 0 otherwise. Figure 1 contains summary statistics on the variables in the dataset.

Figure 1.

The MEANS Procedure

Analysis Variable : AGE				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
263	42.4	13.0	18.0	76.0

The TTEST Procedure

		Analysis Variable : AGE				
TRT_GRP	**N**	**Mean**	**Std Dev**	**Std Err**	**Minimum**	**Maximum**
Group A	133	41.8	12.9	1.1	18.0	75.0
Group B	130	43.0	13.2	1.2	20.0	76.0

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	261.0	-0.79	0.4290
Satterthwaite	Unequal	260.5	-0.79	0.4291

Table of SEX by TRT_GRP

	TRT_GRP		
SEX(Sex)	**Group A**	**Group B**	**Total**
F	107 80.45	107 82.31	214
M	26 19.55	23 17.69	49
Total	133	130	263

Statistics for Table of SEX by TRT_GRP

Statistic	DF	Value	Prob
Chi-Square	1	0.1495	0.6990

Table of Severe_Event by TRT_GRP

	TRT_GRP		
Severe_Event	**Group A**	**Group B**	**Total**
0	80 60.15	59 45.38	139
1	53 39.85	71 54.62	124
Total	133	130	263

Statistics for Table of Severe_Event by TRT_GRP

Statistic	DF	Value	Prob
Chi-Square	1	5.7521	0.0165

Of the 263 total subjects, 133 (51%) were randomized to Treatment Group A and 130 to Treatment Group B (49%). All subjects were between the ages of 18 to 76 years old with an overall average age of 42.4 years. The majority of the sample was female (81%), with Group A being only 19% male and Group B, only 18% male. The sample is balanced as the T- and Chi-square tests showed there was no statistically significant difference on age (p-value = 0.429) or sex (p-value = 0.699) by treatment, respectively. As for the outcome variable, the Chi-square test showed there was a statistically significant difference by group (p-value = 0.017) with those assigned to Group B having a higher probability or percent of experiencing a severe event than those assigned to Group A (55% vs. 40%).

ANALYSIS DETAILS

The main analyses and results detailed in the paper were performed using logistic regression and bootstrap sampling on the data described above. All analyses were conducted using processes and procedures of SAS version 9.4.

Logistic Regression Model

The primary analysis involved calculating the difference in adjusted predicted probabilities between 2 groups and generating 95% confidence intervals for that difference. To accomplish this, the SAS procedure for performing logistic regression modeling, PROC LOGISTIC, was employed as a first step. A binary 0/1 measure (Severe_Event) was used as the dependent variable, where 1 indicated the occurrence of a severe event, and treatment group (TRT_GRP) was the independent variable. AGE and SEX were included in the model as covariates. The SAS code used for this part of the analysis is as follows:

```
proc logistic order=data;
class TRT_GRP SEX / order=data param=glm;
model Severe_Event = AGE SEX TRT_GRP;
lsmeans TRT_GRP / cl ilink diff;
run;
```

The least-square mean option (lsmeans) within PL was employed to generate the model-based group mean APPs as well as the corresponding standard errors for those means. The mean estimates were used to calculate a pair-wise difference in mean APPs (p_{diff}), and the SEs were used to calculate a pooled variance for the difference (Var_{pooled}). The asymptotic normal approximated 95% CI was calculated using:

$$95\%\ CI = p_{diff} \pm 1.96 \times \sqrt{Var_{pooled}}$$

Bootstrap Sampling

To assess the validity and performance of the ANA approach in the primary analysis, an alternative empirical approach was used to get 95% CIs. Bootstrap sampling was the approach selected for creating the comparator CIs.

The first step in the bootstrapping process was to decide how many replicate datasets would be created and how many rounds of bootstrapping there would be. The a priori decision was to have 3 total rounds of bootstrapping, one round each at 100, 500, and 1000 replicate datasets. The repeated sampling technique was performed using the PROC SURVEYSELECT procedure within SAS. The code was as follows:

```
proc surveyselect data=tmp method=balboot outhits
          reps=100 seed=1142 out=out100;
run;
```

Note that, “tmp” represents the original 263 record dataset; “balboot” indicates the bootstrapping method is to be used; “outhits” is an option to preserve all repeatedly selected records; “100” are the k number of replicates; “1142” is a user provided seed for random sample process initiation (can be any integer from 1 to 2 x 10^9); and “out100” captures all the replicated datasets of 263 records each. In this case of k = 100 replicates, the “out100” dataset would contain 26300 rows. Similar code was employed for the k = 500 and 1000 replicate rounds where seeds “2131425” and “520271” replaced “1142” respectively, and “100” was replaced in the code by the corresponding replicate count being generated.

Bootstrap Confidence Intervals

Three total rounds of bootstrapping were completed, each round employing k total models and generating k total p_{diff} values. Within each round, the k values were order from smallest to largest, and the corresponding 2.5th percentile and 97.5th percentile values were selected to create an empirical 95% CI. Those 95% CI lower and upper bound values were: ($\{p_{diff2}/2 + p_{diff3}/2\},\{p_{diff97}/2 + p_{diff98}/2\}$) for k = 100; ($\{p_{diff12}/2 + p_{diff13}/2\},\{p_{diff488}/2 + p_{diff489}/2\}$) for k = 500; and ($p_{diff25}, p_{diff976}$) for k = 1000.

ANALYSIS RESULTS

The results for the APPs, difference in mean APPs, and CIs are presented below. Unadjusted and adjusted analyses are presented.

Predicted Probability Estimates

The following tables display the unadjusted and adjusted predicted probability estimates by treatment group and select summary statistics for the estimates. The bootstrapping results are reported in the “Samples” rows.

Table 1a.

Source	Group A p	$\sqrt{Var}$	Median	Min	Max
Unadjusted Data	0.3985	0.0425	--	--	--
Logistic Modeling	0.4291	0.0495	--	--	--
100 Samples WR	0.4295	0.0496	0.4333	0.3067	0.5519
500 Samples WR	0.4287	0.0500	0.4289	0.2846	0.5873
1000 Samples WR	0.4292	0.0513	0.4301	0.2497	0.5938

Table 1b.

Source	Group B p	$\sqrt{Var}$	Median	Min	Max
Unadjusted Data	0.5462	0.0437	--	--	--
Logistic Modeling	0.5847	0.0505	--	--	--
100 Samples WR	0.5858	0.0556	0.5906	0.3756	0.7188
500 Samples WR	0.5856	0.0509	0.5876	0.4426	0.7939
1000 Samples WR	0.5857	0.0543	0.5867	0.4009	0.7675

The tables show that the predicted probability estimates by treatment group were similar between the unadjusted and adjusted measures. The predicted probabilities were significantly different between treatment groups A and B in both the unadjusted ($p = 0.0165$) and logistic model ($p = 0.0129$) estimation, with treatment B having the greater probability for a severe event occurring. The measure of variance by group is consistent amongst the APPs, but given that the randomized sample was equally balanced, the adjusted APPs variance was consistently larger than the unadjusted APPs variance. Note that the bootstrap sample-based estimates are taken over the k samples of a given round; therefore, the round-specific predicted probability estimate displayed is the mean of the k APPs and the variance displayed corresponds to that mean. The median, minimum (MIN), and maximum (MAX) are also calculated over the k values. The bootstrapping summary statistics indicate a symmetric and stable distribution.

Predicted Probability Difference Estimates and Confidence Intervals

The primary analysis was to obtain an estimate of the difference between mean APPs for 2 groups and to compute CIs for that difference and compare them for performance and validity. Table 2 shows the unadjusted and adjusted predicted probability difference estimates, select summary statistics for those estimates, and the corresponding 95% CIs for the differences. The bootstrapping based estimates are taken over the k p_{diff} values calculated from the k samples of a given round.

Table 2.

Source	p_{diff} (p_B - p_A)	$\sqrt{Var}$	Median	Min	Max	95% C.I. LCL	95% C.I. UCL	C.I. Width
Unadjusted Data	0.1477	0.0609	--	--	--	0.0283	0.2670	0.2387
Logistic Modeling	0.1556	0.0707	--	--	--	0.0170	0.2942	0.2773
100 Samples WR	0.1563	0.0647	0.1492	-0.0132	0.3283	0.0098	0.2730	0.2632
500 Samples WR	0.1569	0.0622	0.1591	-0.0265	0.3665	0.0384	0.2903	0.2519
1000 Samples WR	0.1565	0.0618	0.1583	-0.0251	0.3560	0.0323	0.2755	0.2432

The difference estimates of the mean APPs are similar across the multivariate approaches but are consistently greater than the unadjusted difference estimate. The measure of variance is greatest in the logistic model approach and similar in magnitude amongst the unadjusted and bootstrapping values. The bootstrapping difference estimates are the means of the k p_{diff} values computed for each specific round and the variances reported correspond to each accompanying mean. The median, minimum (MIN), and maximum (MAX) are also calculated over the k p_{diff} values. The bootstrapping summary statistics suggest a well-balanced distribution providing confidence that the sampling method is appropriate for the analysis.

A 95% CI was calculated for each p_{diff} estimate. The CIs for the unadjusted and logistic modeling method were computing using ANA as described earlier, and the bootstrapping CIs were created using the empirical approach also described earlier. None of the CIs contained zero; therefore, all methods provided a p_{diff} estimate that was statistically different from zero at a significance level of 0.05. Additionally, each 95% CI contained all the p_{diff} estimates given by the different approaches. The difference between the upper and lower CI bound, or width, measures the tightness of the interval. A tighter CI is generally associated with a higher quality estimate. The smallest width was for the unadjusted 95% CI and the largest width was for the logistic model CI. The width of the bootstrap sample CIs decreased as the number of replicates for a bootstrapping round increased; therefore, the second smallest 95% CI width was when k = 1000.

SUMMARY AND CONCLUSION

Generating adjusted predicted probabilities for 2 groups from logistic regression is straightforward; however, the aim of this work was to obtain the difference between mean APPs for 2 groups and to calculate and compare CIs for that difference to assess performance and validity. The desired measures are generally not provided using the most common procedures and statistical software packages. The solution described here uses methods and measures provided from common procedures (PROC LOGISTIC, PROC SURVEYSELECT) within a well-known statistical software package (SAS).

The following steps were conducted for the analysis. A difference in mean APPs was calculated from PL based estimates and compared to similarly adjusted estimates from summarized bootstrap samples. The adjusted estimates were also compared to the unadjusted predicted probability difference. The individual treatment group predicted probability estimates as well as the difference in mean APPs estimates were comparable and consistent among the unadjusted and adjusted estimates. The asymptotic normal approximated and bootstrapping 95% CIs were also comparable and consistent with all having a lower bound greater than zero and containing all the reported p_{diff} estimates produced by the 5 different approaches. Additionally, the CIs were relatively similar in performance as measured by CI width; and although the PL based CI was the widest, the magnitude of the width is not great enough to conclude that the performance is not adequate relative to the other CIs.

In conclusion, asymptotic normal approximation is a practical approach that can be valid and perform adequately as a method for calculating CIs on differences between mean APPs for 2 groups. The conclusion is supported by the comparability to and consistency with similar CIs created using an empirical approach, bootstrap sampling in this case. Performance as measures by CI width was also comparable; however, the PL based CI tended to be wider which indicates ANA results in a more conservative CI.

Although PL within SAS was used to perform these analyses, the steps and methods described in this paper can be applied in many other software packages. It is also important to note that while we showed validity in using ANA to compute a CI for a difference in mean APPs for 2 groups, several factors can influence the analysis including, but not limited to: sample size; covariates; missing data; data imbalance; etc.. Therefore, it is important to know and understand the data being analyzed to determine if the steps described here are appropriate to use.

ACKNOWLEDGMENTS

The authors would like to acknowledge and thank Cytel Leadership for affording the opportunity and time to complete this work, as well as Cytel colleague, Bastian Mengelle, who helped develop this work. The authors would also like to thank PHUSE 2025 Leadership for allowing this work to be published and presented. Last, but certainly not least, thank you to Everyone who reads this paper or attended the presentation.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Author Name: Cralen Davis
Company: Cytel, Inc.
Email: cralen.davis@cytel.com