PharmaSUG 2022 - Paper SA - 003

Estimating Differences in Probabilities (Marginal Effects) with Confidence Interval

Jun Ke, Independent Statistician
Kelly Chao, LLX Solutions
Corey Evans, LLX Solutions

ABSTRACT

Fitting logistic models using SAS procedures such as PROC LOGISTIC can obtain log odds for populations in the data, you can also obtain estimates of odds and odds ratios by options prespecified in the procedure (or by exponentiating the log odds). But to get difference in population event probabilities (population means), we need to estimate a nonlinear function of the parameters of the logistic model. Alternatively, you could model the probabilities themselves rather than the log odds. This paper will present the statistical background as well as the modeling process using layman's explanations. SAS Inc has provided a series of macros that could assist with calculating the difference in probabilities, however, in this paper, we will introduce a more robust model that could accommodate uncommon scenarios that could happen during the trial.

INTRODUCTION

Efficacy endpoints in clinical trials are usually summarized by the estimates of rate or proportion of responder, denoted as P(Y=1|Treatment) and P(Y=1|Placebo). However, there's no options available now for confidence intervals and P-values of LSMEAN P(Y=1|T) - P(Y=1|P). This paper provides some solutions to calculate such values.

EXAMPLE

Let us start with a sample response analysis table we typically see.

Suppose our data contains 100 placebo patients and 100 treatment patients in ITT population. We are interested in whether the patient's health condition would hit an improvement target (Y=1) at 48 weeks.

Table 1 Proportion of Patients with Improvement Target

	Placebo	Treatment
Number of subjects in ITT Population	100	100
Number of subjects with Y=1 at week 48	25(25%)	32(32%)
Estimated proportion of subjects with Y=1[a]	0.232 P(Y=1|P)	0.289 P(Y=1|T)
Odds ratio (Treatment vs. Placebo) [a]		1.35 P(Y=1|T)/P(Y=1|P)
95% CI[a]		(0.78, 2.33)
p-value[a]		0.2908

[a] Based on logistic regression adjusted for category1, category2, covariate1, covariate2 and covariate3.

Sometimes we are interested in marginal effects $P(Y = 1|T) - P(Y = 1|P)$ adjusted by covariates, per above example, the marginal effect now becomes

$$P(Y = 1|T) - P(Y = 1|P) = 0.289 - 0.232 = 0.057$$

We will discuss how we shall get the difference, confidence intervals, and p-value in the next section.

MULTIPLE APPROACHES

APPROACH 1: LINEAR REGRESSION (IDENTITY LINK)

We can estimate the probability of the responder by:

$$E[P(Y=1)] = \beta_0 + \beta_1 \times I(Treatment) + X_2\beta_2 + \cdots + X_p\beta_p$$

Where $I(Treatment)$ is the indication function for treatment (here the indication function becomes 0 for placebo and 1 for treatment).

Therefore,

$$E[P(Y=1|T)] - E[P(Y=1|P)] = \beta_1$$

Model the rate directly and easy to implement in SAS

```
proc genmod data=adeff descending;
      class category1 category2 trtpn;
      model aval = category1 category2 covariate1 covariate2 covariate3
      trtpn/dist=binomial link=identity;
      lsmeans trtpn/ diff cl;
run;
```

Using linear regression approach as above is straightforward but it may run into convergence issue when P(Y=1) is close to 0 or 1. Moreover, linear regression is designed for a linear trend with range from negative infinity to infinity, and logistic is more for proportions with range from 0 to 1, results from linear regression may be inconsistent with the ones from logistic regression.

APPROACH 2: LOGISTIC REGRESSION (LOGIT LINK)

Logistic regression is one of the generalized linear model with a logit link to model a binary dependent variable. Under logit setting, we can model our response as below:

$$log\frac{E[P(Y=1)]}{1-E[P(Y=1)]} = \beta_0 + \beta_1 \times I(Treatment) + X_2\beta_2 + \cdots + X_p\beta_p$$

Let L_p and L_T be the linear predictor for placebo and treatment, respectively.

$$L_P = \beta_0 + \beta_1 + \beta_{P2}X_2 \ldots + X_{Pp}\beta_p,$$

$$L_T = \beta_0 + \beta_{T2}X_2 \ldots + X_{Tp}\beta_p$$

The difference between two rates can be considered as a function of β:

$$E[P(Y=1|T)] - E[P(Y=1|P)]$$

$$= h(\beta) = \frac{1}{1+exp(-L_T)} - \frac{1}{1+exp(-L_P)}$$

where $\beta = (\beta_0, \ldots, \beta_p)^T$.

Then the standard error for rate difference can be derived by Delta method. For large sample size,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma)$$

The distribution of $h(\beta)$ can be approximate by

$$\sqrt{n}(h(\hat{\beta}) - h(\beta)) \xrightarrow{d} N(0, (h')^T \Sigma h)$$

where $h'(\beta) = (\frac{\partial h}{\partial \beta_0}, \ldots, \frac{\partial h}{\partial \beta_p})^T$ and $\frac{\partial h}{\partial \beta_i} = \frac{exp(-L_T)X_{Ti}}{[1+exp(-L_T)]^2} - \frac{exp(-L_P)X_{Pi}}{[1+exp(-L_P)]^2}$.

The following SAS scripts fit a logistic regression with proc genmod.

```
proc genmod data = adeff descending;
      class category1 category2 trtpn (desc);
      model aval = category1 category2 covariate1 covariate2 covariate3
      trtpn/dist=binomial link=logit;
      lsmeans trtpn /ilink diff exp e cl;
      ods output coef=c;
run;
```

The below SAS output provides almost everything except confidence intervals of the rate difference.

TRTPN Least Square Means														
Planned Treatment (N)	Estimate Log Odds	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper	Mean Rate	Standard Error of Mean	Lower Mean	Upper Mean	Exponentiated Odds	Exponentiated Lower	Exponentiated Upper
1	-0.9016	0.2096	-4.30	<.0001	0.05	-1.3125	-0.4908	0.2887	0.04305	0.2121	0.3797	0.4059	0.2692	0.6121
0	-1.1980	0.2225	-5.38	<.0001	0.05	-1.6341	-0.7620	0.2318	0.03962	0.1633	0.3182	0.3018	0.1951	0.4667

Differences of TRTPN Least Square Means										
Planned Treatment (N)	Planned Treatment (N)	Estimate Difference in Log Odds	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper	Exponentiated Odds Ratio	Exponentiated Lower	Exponentiated Upper
1	0	0.2964	0.2806	1.06	0.2908	0.05	-0.2536	0.8464	1.3450	0.7760	2.3312

Below is SAS output from logistic.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.5046	0.6705	-1.8187	0.8095	0.57	0.4517
Catogory1	xxx	1	-0.3656	0.3218	-0.9961	0.2649	1.29	0.2557
Catogory1	yyy	1	-0.6376	0.3821	-1.3865	0.1113	2.78	0.0952
Catogory1	www	0	0.0000	0.0000	0.0000	0.0000		
Catogory2	zzz	1	0.1804	0.3052	-0.4177	0.7785	0.35	0.5544
Catogory2	vvv	0	0.0000	0.0000	0.0000	0.0000		
Covariate1		1	0.0655	0.1448	-0.2182	0.3493	0.20	0.6509
Covariate2		1	0.0117	0.0579	-0.1018	0.1253	0.04	0.8396
Covariate3		1	-0.0392	0.0312	-0.1004	0.0220	1.58	0.2089
TRTPN	1	1	0.2964	0.2806	-0.2536	0.8464	1.12	0.2908
TRTPN	0	0	0.0000	0.0000	0.0000	0.0000		
Scale		0	1.0000	0.0000	1.0000	1.0000		

To get the confidence interval, we will fit logistic regression by PROC NLMIXED.

```
proc nlmixed data=adeff;
      parms Intercept -0.5046 b1 -0.3656 b2 -0.6376 b3 0.1804 b4 0.0655
      b5 0.0117 b6 -0.0392 b7 0.2964; → set initial values obtained from logistic
      regression

      p = logistic(Intercept + b1*(category1=xxx) + b2*(category1=yyy)
      + b3*(category2=zzz) + b4*covariate1 + b5*covariate2 +
      b6*covariate3 + b7*(trtpn=1));

      model aval ~ binary(p);
```

```
      estimate 'T-P' logistic(Intercept + 0.3333*b1 + 0.3333*b2 +
      0.5*b3 + 3.0418*b4 + 6.1137*b5 + 22.088*b6 + b8) -
      logistic(Intercept + 0.3333*b1 + 0.3333*b2 + 0.5*b3 + 3.0418*b4 +
      6.1137*b5 + 22.088*b6); → design matrix from lsmean e option
run;
```

Rate difference from PROC NLMIXED.

Additional Estimates								
Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
T - P	0.05689	0.05372	263	1.06	0.2906	0.05	-0.04888	0.1627

APPROACH 3: APPLY MACROS DIRECTLY PREPARED BY SAS INC

Luckily SAS Inc. has prepared several macros to support the need of estimating differences in probabilities. Details can be found by using the link: https://support.sas.com/kb/37/228.html. Several methods are illustrated in this website showing how the difference in population probabilities can be estimated and tested in logistic models with either a binary, or a multinomial response.

CONCLUSION

By using the logistical regression (logit link) and PROC NLMIXED, we are able to find the confidence interval and p-value of the difference in probability. The other approach, such as linear regression (identity link), could find the result within one procedure, but this method might cause the convergence issue when P(Y=1) is close to 0 or 1. Macro from SAS Inc. gives us the result of the differences in probabilities without the phase of processing. Therefore, when estimating the confidence interval of differences in probability, logistical regression (logit link) and PROC NLMIXED would give us the accurate value, also with more flexibility to the uncommon scenarios during the study.

REFERENCES

SAS Institute. “Usage Note 37228: Estimating differences in probabilities (marginal effects) with confidence interval” https://support.sas.com/kb/37/228.html.

ACKNOWLEDGMENTS

The authors would like to acknowledge the following contributors to this body of work for their support: Jin Shi and Hao-Wun Chen.

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- *SAS® For Dummies®*

CONTACT INFORMATION

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

Jun Ke, PhD
statker@gmail.com