PharmaSUG2011 – Paper PO08

FAQ: Issues with Efficacy Analysis of Clinical Trial Data Using SAS

Sandeep Sawant, i3 Statprobe (INGENIX Pharmaceutical Services), India
Neha Mohan, i3 Statprobe (INGENIX Pharmaceutical Services), India

ABSTRACT

In pharmaceutical/CRO industry, role of statistical programmers is to support the statistical activities of the clinical trial by means of performing safety and/or efficacy analysis. It is recommended that the statistical programmers should have the knowledge about the statistical methods used especially while working on the efficacy analysis. This however, is not mandatory because detailed specifications are provided by the Biostatisticians. Despite of well written specifications, there is a variety of questions from the programmers regarding the analysis or the outputs produced by SAS® procedures. There are many situations where SAS doesn't provide the output directly and hence some kind of manipulations is required. This paper is designed to impart a better understanding of efficacy analysis using SAS, as well to communicate the related complexities in the execution and application.

INTRODUCTION

This paper is in a question - answer format and is a collection of questions raised by statistical programmers while performing the efficacy analysis.

Q1) Is there any procedure in SAS using which will compute the Geometric Statistics such as Geometric Mean, Geometric SD and Geometric CV?

Ans: There is no such procedure in SAS which gives the Geometric statistics directly. However, making use of the below formulae, we can compute the required statistics

Geo CV	The geometric coefficient of variation is calculated as: $\text{Geo CV} = \sqrt{\exp^{(\ln(\text{GeoSD}))^2} - 1} \times 100\%$
Geo Mean	The geometric mean is calculated as: $\text{Geo Mean} = \sqrt[N]{y_1 \times y_2 \times y_3 \ldots y_N} = \exp[\frac{\ln(y_1)+\ldots\ln(y_n)}{n}]$
Geo SD	The geometric standard deviation is calculated as: $\text{Geo SD} = \exp^{(SD(\ln(y_1),\ln(y_2),\ln(y_3)\ldots\ln(y_N)))}$

For instance, if you want to calculate the Geometric Statistics for variable X proceed as follows:

- Derive the variable Y as Y=log(X).
- Calculate the mean and SD for variable Y using PROC MEANS.
- Using the EXP function, calculate the antilog of mean and SD obtained from PROC MEANS. The resultant values are the Geometric Mean and Geometric SD respectively for the variable X.
- Once the geometric SD is obtained, make use of the aforementioned formula to get the Geometric CV.

Q2) I am calculating 95% CI for the mean using PROC MEANS. In my dataset, all the values are identical. The output from PROC MEANS has missing values for both upper and lower limits. In this scenario, what values should be presented for the upper and lower limits in the output?

Ans: When the dataset contains all identical values, this value becomes the point estimate for mean. Here both the upper as well as the lower limit for CI of the mean is the same as that identical value. This fact has not been taken into consideration while computing CI in PROC MEANS. Hence, it is recommended to use PROC UNIVARIATE with CIBASIC options whenever 95% CI for means is required. PROC UNIVARIATE takes into account this special data case while calculating the same.

Q3) How do I check if my data follows normal distribution?

Ans: Use PROC UNIVARIATE with NORMAL option.
SAS produces p-value for four normality tests viz Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-Von Mises and Anderson-Darling. Consult with the statistician for any preference for a particular test. If there is no preference, all the four tests should be used for conclusion. If the p-values from all the four tests are greater than 0.05 (or desired level of significance), you can conclude that the data follows normal distribution. If at least one p-value is less than 0.05 (or desired level of significance) , it would indicate that your data doesn't follow normal distribution.

Q4) While computing the Pearson's correlation coefficient I get a missing value. What may be the reason and what value should be presented on the table?

Ans: This happens when one of the variables contains identical values (which mean zero variance).
Correlation is defined as covariance divided by product of SD of the two variables for which correlation is needed. Hence, no variance (SD=0) for one of the variable results in missing correlation. In this situation, display the correlation as NC (Not Computable).

Q5) I want to calculate the CI for proportion of subject achieving Complete Response (CR) for treatment arm A. However in my dataset, out of 10 subjects receiving treatment A, no subject has achieved CR. I am using PROC FREQ with BINOMIAL option. I get the CI for proportion of subjects not achieving CR but I need the CI for proportion of subjects achieving CR. How should I calculate it?

Ans: You can make use of ZEROS option in PROC FREQ.
Create a dataset for counts and add a dummy record in your dataset to indicate that no subjects have achieved CR as follows

Treatment	Response	Count
A	CR	0
A	No-CR	10

Use PROC FREQ with WEIGHT statement and ZEROS option (i.e. WEIGHT count/ ZEROS). This will provide the CI for the subjects achieving CR despite of data containing No-CR.

Q6) My dataset contains both the responses Yes/No. I want the CI for binomial proportion, for responders i.e. for Yes. However, PROC FREQ is providing the CI for NO.

Ans: PROC FREQ computes the CI for the lower level of the variable (when sorted alphabetically). In the given scenario, CI will get computed for the lowest level i.e. No. If you need the CI for the level 'Yes' then please correct the BINOMIAL option to BINOMIAL(LEVEL=2). This will compute the CI for the second level of the variable holding Yes/No values, i.e. CI for responders (Yes).

Q7) When should I use Chi-Square and when to use Fisher's test?

Ans: Both these tests are used for analysis of 2X2 (or more) contingency tables. Fishers test is most appropriate, if the expected frequency of any of the cell group is less than 5. In order to get the expected frequency use OUTEXP option in TABLE statement. If all the expected frequencies are more than 5 use Chi-square test.

Q8) I am using Mc-Nemar test. However, corresponding p-value is not getting calculated and there is a note in the log indicating that one of the variable has less than 2 non-missing levels. What should I do?

Ans: Mc-Nemar's test is used for analysis of aXb tables where both a and b are greater than 1 i.e. table should be at least 2X2. If some of the cell frequencies are missing, add dummy records with zero counts as follows to complete the 2 X 2 table.

Treatment	Response	Count
A	CR	0
A	No-CR	10
B	CR	0
B	No-CR	10

Use PROC FREQ with WEIGHT statement and ZEROS option (i.e. WEIGHT count/ ZEROS). This will create 2X2 tables and the p-value will get calculated.

Q9) I am using CMH test and have been asked to use alternative hypothesis as "Row Mean Scores Differ". However, I get a different p-value than the validation programmer. We both are using the same data for PROC FREQ?

Ans: The reason for the difference in the p-values could be due to the differences in the order in which the variables are listed in the table statement.
When you use CMH with "Row Mean Scores Differ" as alternative hypothesis, make sure that the desired variable appears in rows in the 2X2 table. Suppose we want to find out the association between treatment and response across the country. Make sure that the treatment variable appears in rows and response in column for the country. This can be obtained as Tables Country*Treatment*Response/CMH in PROC FREQ.

Q10) We are checking association between occurrence of AE and Treatment using Chi-Square test. 10 Subjects have received Treatment A and 10 subjects have received Treatment B. In AE dataset, 9 subjects from treatment arm A have reported AE and 8 subjects from treatment B have reported AE. I am creating the binary variable in AE dataset to indicate occurrence of AE. However, I get different p-value than validation programmer.

Ans: Make sure that missing responses are set as non-responders i.e. 1 subject from treatment A who has not experience an AE should be treated as non-responder and 2 subject from treatment B who has not experience an AE should be treated as non-responder. Basically, all the treated subjects should have the binary flag instead of creating the binary flags in AE. This data then passed to PROC FREQ will give the p-value.

Q11) What are Semi-Log plots?

Ans: In semi-log plot, the data is plotted on logarithmic scale. However, the data is not log-transformed. The axis can be converted to logarithmic scale by using LOGBASE=10 option in AXIS statement.

Q12) What are Sphegetti plots?

Ans: Sphegetti plots are 3D plots with data for different individuals (subjects/patients) plotted on the same page. For e.g. Plasma concentration plotted on y-axis, time-points on x-axis for 10 subjects will constitute a sphegetti plot. This can be plotted using the statement PLOT plasma*time=subjid in the PROC GPLOT.

Q13) Is there a difference between Box plot and Box-Whisker Plot?

Ans: In box plot, the upper and lower end of the box is extended to the minimum and maximum values of the data. On the other hand in a box-whisker plot, the upper and lower ends are extended till 1.5 inter-quartiles range and the values beyond those points are denoted by a specified symbol. Box plots can be plotted using I=BOX00 option and Box-whisker plot can be plotted using I=BOX option in the SYMBOL statement.

Q14) I have been asked to use PROC TTEST for comparing means of two treatment groups. The resultant TTEST generates two p-values - one with equal variance and the other with unequal variance. Which p-value should be used?

Ans: You need to first check if the variances of the two groups are the same. This can be checked using the p-value from "Equality of Variance" section. If p<0.05 (or desired level of significance), we can say that variances are unequal for these two groups. Otherwise we can conclude that the variances for these two groups are equal. Once this decision is taken, you can choose the appropriate p-value from the section "T-Test" based on equal/unequal variance.

Q15) I am getting different value for the odds ratio than the validation programmer despite of using the same syntax in PROC LOGISTIC. What could be the reason?

Ans: One should take care of the below points while calculating the odds ratio. The value of the ratio would differ if there is inconsistency in any of these.
1) Check if missing responses need to be set as non-responders.
2) Check if the model is getting fitted for the correct level of response. If the response variable has values Yes/No then the model should be fitted for Yes. Check the "Response Profile" section to see for which level model is fitted. If the model is fitted for No, use DESCENDING option after PROC LOGISTIC DATA=<datatest> statement.
3) There are two ways of getting odds ratio. By default SAS provides Wald odds ratio. Another method is to use CONTRAST statement with appropriate contrast coefficients. If you are using CONTRAST statement make sure you are using correct contrast coefficient and using ESTIMATE=EXP option.

Q16) While using PROC LOGISTIC there is a log message that reads "More coefficients than the levels specified for the effect. Some coefficients will be ignored." How could this warning be avoided?

Ans: Use PARAM=GLM option in the CLASS statement.

Q17) I am using PROC GENMOD to fit the logistic regression model. How do I get the odds ratios from it?

Ans: Remember to use the option DIST=BIN and LINK=LOGIT in the MODEL statement. Make sure you are taking the antilog (using EXP function) of the estimates obtained from ESTIMATE statement which will provide you odds ratio and corresponding CI.

Q18) Is a value of Odds ratio such as 1.2833E14 correct?

Ans: It is ok to have odds ratio value too high or low.
This usually happens in cases when there is an unequal distribution of responses. For instance, out of 100 responses 95 are responders and the remaining are non-responders. There isn't anything that we can do here.

Q19) I am fitting a generalised linear model (GLM) with treatment and country as fixed effects and the interaction between country and treatment. I get some of the estimates for LSMean as 'Non-Est' when using estimate statement?

Ans: Whenever an interaction term is fitted in the model, it is advisable to use the interaction term in the TABLES statement of PROC FREQ to examine the missing data. If any of the cell frequencies contain zero observation, estimates would be non-estimable. In this situation, either wait for more data or consult with statistician regarding removal of the interaction term.

Q20) I am performing survival analysis using PROC LIFETEST. I have been asked to display the survival probabilities at 6, 12, 18 months. However LIFETEST is providing the probabilities at censored values?

Ans: Make use of TIMELIST option in PROC LIFETEST. This will provide you the survival probabilities at the required time points.

Q21) To generate KM plot using PROC GPLOT, I am using the survival probabilities from OUTSURV dataset generated through PROC LIFETEST. I expect the plot to match the one generated through plot=(S) option in PROC LIFETEST. However this is not the case. What could be the reason?

Ans: If you look at the OUTSURV dataset, you will notice that some of the time values have missing probabilities. If you use this data directly for plotting KM plot, time points with missing probabilities will be dropped; hence the difference. To avoid this, sort the data by time variable within the strata and retain the previous non-missing probability values in place of missing ones. Use these derived probabilities to produce KM plot.

CONCLUSION

I have tried covering some of the questions asked by statistical programmers which they have faced while performing the statistical analysis because of various data cases or the way SAS produces the output. Though the question list is not exhaustive, this paper should serve as a basis for understanding the topic of efficacy analysis and the related complexities.

ACKNOWLEDGMENTS

I thank all my colleagues 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, , a division of Ingenix Pharmaceutical Services, Inc.
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