Survival Analysis Translation from Statistician to Programmer

Amanda Yi, Roche, Shanghai, China

Abstract

Survival analysis is very common in clinical trials. Programmers are often requested to program tables and figures that are often referred to Kaplan Meier plots simply. There can be a huge gap between the language of a statistician and a programmer. This paper will provide a basic understanding of survival analysis by using example data and code to compare what the programmer sees to what the statistician would describe. The output data from PROC LIFETEST will be explained.

Introduction

In most pharmaceutical company, statistician and programmer work together for statistical analysis tables as separate roles. Statistician focuses more on concept and is the final decision maker for the analysis scope and content, and programmer focus more on the programming part to make the tables available based on complex data scenarios. The problem occurs when statisticians want to use their language with programmers that are unfamiliar with the statistical terms. This issue is compounded when statisticians throw out these terms without explanation. This paper will aim to build relations between statistician's language and the understanding from programmer.

Statistical Terms

Here are a few of the most-used terms with their basic definitions and information:

Event: What we are calculating the time to. And it's also referred to as 'Event of Interest'. This can include such things as death, recurrence of disease/symptoms.

Censor: What we use for subjects that do not experience an event. We cannot follow subjects indefinitely, but still need to be able to use them in the analysis. The most common form of censoring is right censoring. Think of a timeline, where there is a line marking the end of the study follow-up period. If the event ever did occur it would happen to the right of that line. With the CENSOR variable, 1 is commonly used to indicate that an event occurred and 0 indicates that the event did not occur (the subject was censored).

Failure: Used to indicate subjects that experienced the event. This comes from the traditional use of survival analysis - analyzing death. If a subject experienced death, you would consider that subject and the treatment a failure in avoiding death.

Survival: All of this is part of 'Survival Analysis', another name for time-to-event analysis. This is often used to calculate, model, and derive inference about the time to death. Hence the term 'survival' may be used even when we are analyzing recurrence or recovery.

Survival Function: S(t)=Pr[T>t]. It is the probability that the actual time of event (if we could follow subjects indefinitely) is greater than the current time that we are evaluating.

LIFETEST: The name of the SAS® procedure used. This procedure produces survival functions (estimates on how subjects are experiencing events), plots, and other pieces of survival analysis.

Data To Be Used in Analysis

There are three required variables for input data sets: the subject variable (USUBJID), censor indicator variable (STATUS), and time to event variable (DUR). The subject variable can take any value to differentiate subjects. It should be one record per subject. If this is not the case, we will not have reliable results. The censor indicator variable will indicate whether the subject experienced the event or not. The time to event variable will take the value of the time the event occurred or the last observed time for censored subjects. This variable can be in any unit, for example hours, days, weeks, months, or years. Grouping variables such as treatment or stratifications are often included when the object of the analysis is to determine the effect of group differences in survival.

Below is the input data set that will be used to create output. USUBJID is our subject variable, ARM is our treatment grouping variable, and DUR (in months) is our time to event variable. The censor variable STATUS is the indicator of an event, where 1 indicates that death occurred and 0 indicates that it did not.

USUBJID	ARM	DUR	STATUS
1	1	1	1
2	2	2	0
3	1	3	1
4	2	5	1
5	1	6	1
6	2	9	0
7	2	10	1
8	1	11	1
9	2	12	0
10	1	13	1
11	2	14	1
12	1	17	1
13	1	18	0
14	2	19	1
15	1	21	0
16	2	23	1
17	1	24	1
18	2	25	1

Program Code To Be Used for Output

```
proc lifetest DATA=tte method=km plots=(survival) OUTSURV=ttesurv;
strata ARM;
time DUR*STATUS(0);
run;
```

As we have the required data to analyze, then we can run the LIFETEST procedure as above. This example code illustrates the capabilities of PROC LIFETEST. Your statistician may throw a sentence to you 'Give me a Kaplan Meier Plot!' Before you jump into programming, well understanding below things can help you get more clear and confident to the output you will deliver to the team.

METHOD: Specifies the method of estimation used to create the survival function. The Kaplan Meier (KM) method is the default and most common. This method is best when event times are measured with precision, especially if the number of observations is small. The other option is the Life-Table (LT) method and is better for large data sets or when measurement of event times is crude.

STRATA: Indicates the variable to use when comparing survival functions. Usually we are comparing different groups, sometimes we add stratification variables as well for comparison between different stratifications groups.

TIME: Assigns the time and censor indicator variables. Time variable * Censor indicator variable (0/1), (0/1) sets the value of the censor variable that indicates that no event occurred within the study. Making sure the correct indicator used is very important to get the right results.

PLOTS: Specifies the output plot. The most common one is the Kaplan-Meier plot (=survival).

OUTSURV: Names the output data set that will contain survival estimates and confidence limits.

Interpreting the Output

PROC LIFETEST produces a lot of information. The key part is interpreting the output. Let's break the information into sections and explore the meaning behind the labels, values, and SAS notes.

The first section, Table 1, is the Product-Limit Survival Estimates.

Table 1

Product-Limit Survival Estimates

DUR	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	9
2.0000*	.	.	.	0	8
5.0000	0.8750	0.1250	0.1169	1	7
9.0000*	.	.	.	1	6
10.0000	0.7292	0.2708	0.1650	2	5
12.0000*	.	.	.	2	4
14.0000	0.5469	0.4531	0.2006	3	3
19.0000	0.3646	0.6354	0.2001	4	2
27.0000	0.1823	0.8177	0.1632	5	1
29.0000	0	1.0000	.	6	0

NOTE: The marked survival times are censored observations.

This section is output for both values of STRATA. The first column is the time variable DUR, where all values of DUR in that stratum are displayed in order. There is an extra observation for time 0.00 that is always included in the output even if no event or censoring occurred at that time. If a time value is for a censored subject an asterisk (*) will be displayed to the right of the time. The second column is the survival estimate at each time. As mentioned previously, survival can be defined as the probability that the actual time to event is greater than the current time. When a subject is censored, the survival estimate takes a missing value in the output. People are often confused by this output because when discussing survival, we treat the estimate as constant throughout censored times. For example, in treatment group 2, the survival estimate at 2 months would still be 1, because no event had occurred. This estimate would not drop until 5 months when the first event occurred, lowering the survival estimates to 0.8750. The third column (failure) is simply (1-survival), or the probability of the event occurring by that time point. The fourth column shows the survival standard error, which can used to create confidence intervals. SAS will also produce a reliable confidence interval in our output data set. The fifth column, labeled Number Failed, will only increase when a subject experiences the event. The sixth column, Number Left, will start with the total number of subjects in that stratum and will decrease with each observation. A subject that is censored cannot be considered as part of the treatment group after the time of their last observation.

The next section, Table 2, Summary Statistics for Time Variable DUR, is where you can acquire the median survival time within the stratum. Percent indicates the quartile which can be then translated into survival values. 25 percent marks at the estimated time point where survival would drop to 0.75. The Point Estimate column displays that time point value if it available. The transformation type is LOGLOG as default and you can customize it to Linear or Log…etc. by using option CONFTYPE.

Table 2

```
           Summary Statistics for Time Variable DUR

                     Quartile Estimates

                   Point           95% Confidence Interval
         Percent   Estimate   Transform     [Lower     Upper)

            75      27.0000    LOGLOG       14.0000    29.0000
            50      19.0000    LOGLOG        5.0000    29.0000
            25      10.0000    LOGLOG        5.0000    19.0000


                      Mean      Standard Error

                    18.3073         3.5830


        Summary of the Number of Censored and Uncensored Values

                                                         Percent
     Stratum         ARM      Total  Failed   Censored   Censored

        1             1         9       7         2        22.22
        2             2         9       6         3        33.33
     -------------------------------------------------------------
      Total                    18      13         5        27.78
```

The median survival time is often requested in the time to event table summary. The information can be found under the 'Point Estimate' column where the 'Percent' column is 50. If there is no event time where the survival estimate exactly equals 0.5, we would use the smallest observed survival time where the survival estimate is less than 0.5, which in the case of ARM 2 is 19 months. If there is a time where the survival estimate is exactly 0.5 we would take the average of the time where the estimate equals 0.50 and the following time.

Confidence intervals are also produced for the quartile estimates. If you get missing upper limit, don't worry, it might due to the right-censoring in the data and not a programmatic error. This right-censoring also caused SAS to generate a NOTE stating that the mean and standard error of survival time were underestimated.

The Summary of the Number of Censored and Uncensored Values section in the last part of table 2 gives basic frequencies of number of subjects, number of failures, number of censored, and the percent censored. Note that having more censored subjects does not necessarily mean that the treatment was better. If all the subjects were censored before the end of study due to drop outs or loss of follow-up this is more of an indication that there was an issue with follow-up methods or possibly study design than an issue with the effectiveness of one treatment over another.

We get the Product-Limit Survival Estimates, Quartile Estimates and Mean/Standard Error Estimates for both treatment groups. Then let's go to the comparison. Table 3 is usually the most interest part to statisticians. We can find the p-value in the Test of Equality over Strata section. The output contains the default test. Other tests can be added using the TEST option in

the STRATA statement. Our sample output does not show significant difference between two treatment groups at the alpha=0.05 level using any of the default tests.
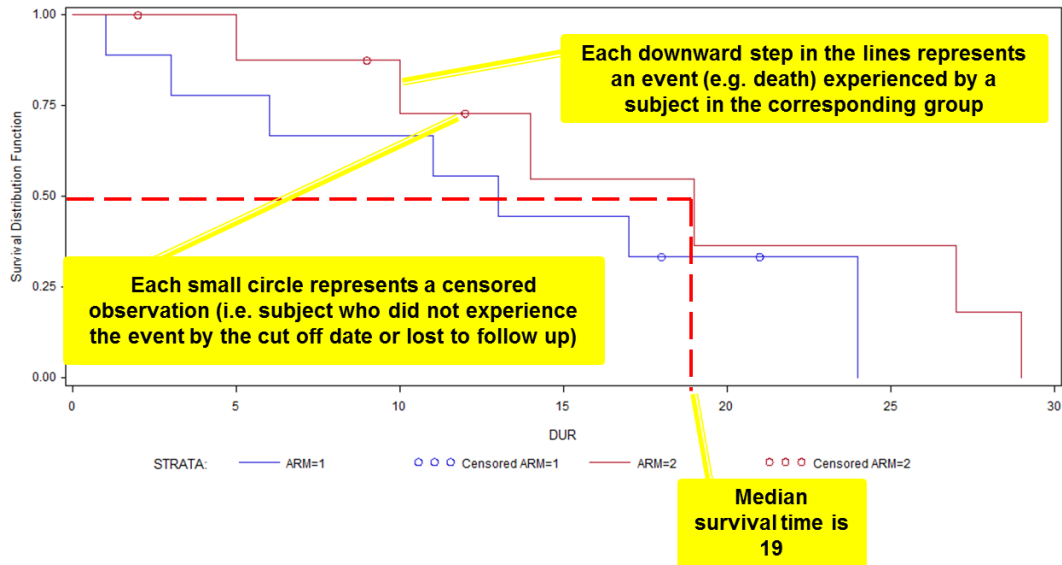
Table 3

```
       Test of Equality over Strata

                                    Pr >
Test         Chi-Square    DF    Chi-Square

Log-Rank       0.9808       1      0.3220
Wilcoxon       0.9704       1      0.3246
-2Log(LR)      0.2229       1      0.6369
```

Interpreting the Plot

A picture is worth a thousand words. One of the quickest ways to analyze the data is to review the Kaplan Meier plot. Below figure 1 is the default plot produced. A higher survival probability at each time is preferred when the event is something undesirable, such as death. From the plot we can see treatment 2 stays above treatment 1 throughout the plot. In the plot, the circles represent a censored time and each drop means an event (death) happen at the time of X axis. We can also estimate the probability that a subject survives past 15 months by locating 15 months on the X axis and reading up to the Y axis. The proportion of subject s surviving past 15 months is over 50% for treatment group 2. The median survival is estimated by locating 0.5 on the Y axis and reading over and down to the X axis. The median survival is approximately 19 months for treatment group 2.

There are many ways to plot the survival data. In the default one, there is less flexibility on how the information is displayed. If you need control of the axes, colors, labels, etc. you will have to use the OUTSURV data and your preferred plotting method.

Figure 1

Each downward step in the lines represents an event (e.g. death) experienced by a subject in the corresponding group

Each small circle represents a censored observation (i.e. subject who did not experience the event by the cut off date or lost to follow up)

Median survival time is 19

Survival Distribution Function

DUR

STRATA: —— ARM=1   ○ ○ ○ Censored ARM=1   —— ARM=2   ○ ○ ○ Censored ARM=2

Something to consider before using PROC LIFETEST

This section is simply some tips for using proc lifetest. Hope this can save you some time when encounter problems.

1) Consult your statistician and make sure use the right method (KM/LT)
2) Multiple observations per subject will be counted twice in LIFETEST. LIFETEST is for one occurrence-only events.
3) 0 is not always censored indicator, be careful of passing it into program.
4) Missing confident limit values can come from the data and censoring, so it is not always a program error.
5) Default transformation type in SAS 9 or above is LOGLOG and can be changed by CONFTYPE option. In SAS 8, there is no flexibility to change it and forced to LINEAR.

References
SAS User Guide Chapter 37, The LIFETEST Procedure.

Contact Information

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

Name: Amanda (Chun) Yi
Enterprise: Roche
City, State ZIP: Shangai, China
E-mail: chun.yi@roche.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.