Surviving your Statistician: A Programmer's Guide to Survival Analysis

James Diserens, Statistical Programmer, Veramed Limited

veramed

1. Introduction

Working alongside statisticians, statistical programmers can be required to produce or validate statistical outputs. Effective programmers understand the purpose of the outputs and the meaning of the results.

Survival Analysis is a commonly used statistical method in clinical trials that involves the investigation of time-to-event data. The key variable of interest is time and the event of interest must be a clearly measurable, pre-defined, distinct and recordable point, such as death, the first relapse of cancer or the failure of a component in engineering. This poster introduces the main concepts of survival analysis from a programming viewpoint.

2. Kaplan-Meier

The Kaplan-Meier Estimate product limit is the most widely used method for representing time-to-event data and to estimate the survival over time:

$$\hat{S}(t) = \prod \frac{n_i - n_e}{n_i}$$

$\hat{S}(t) = $ Survival Estimate
$n_i = $ number of subjects at the start
$n_e = $ number of subjects experiencing event

The Kaplan-Meier Estimate product limit is usually presented as a graph or Kaplan-Meier plot, as a series of steps representing each small time interval. This shows the estimated probability of surviving for a given amount of time.

For the purpose of the calculation of the survival function it is necessary to assume that the events occurred at the end of the time interval. This means that it is important to choose an appropriately short time interval, often days in clinical trials.

EXAMPLE by hand: 10 subjects reviewed daily, all subjects followed up for 10 days (no censoring takes place, see (3.) below):

Time (t)	Number of Subjects Experiencing Event	Number of Subjects Remaining n	Survival Probability n end /n start	Survival Function Calculation	Survival Estimate S(t)
0	0	10	10/10=1.0		1.00
1	1	9	9/10=0.90	= 0.90 x 1.00	0.90
4	1	8	8/9=0.89	= 0.89 x 0.90	0.80
5	2	6	6/8=0.75	= 0.75 x 0.80	0.60
8	1	5	5/6=0.83	= 0.83 x 0.60	0.50
10	1	4	4/5=0.80	= 0.80 x 0.50	0.40
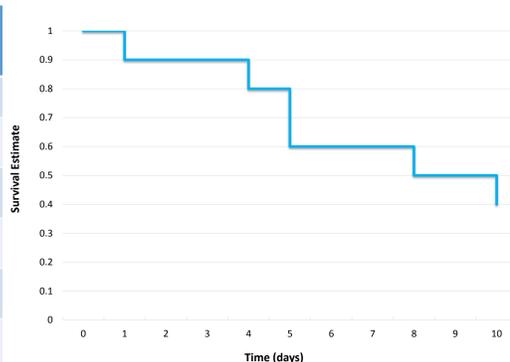
3. Censoring

Complete follow up of a subject is not always possible; a subject may withdraw from a trial early and before the event of interest has been observed. In survival analysis, this is called censoring. These subjects still provide useful information and should not be ignored as removing them could introduce a bias into the study data.

In clinical trials right censoring is most common. This is where we know that an event has not occurred by a certain time point (time of withdrawal), but we do not know what might have happened, had the subject been followed up for longer. We assume that censored subjects have the same risk of an event as those who continue to be followed up.

There could several reasons why a subject is censored: A study ends before the subject could experience the event, the subject withdrew from the study or 'lost to follow-up' or the subject has another event that stops the event of interest from happening (e.g. dying before experiencing a different event of interest, known as a competing event).

EXAMPLE by hand: As above, but this time two subjects withdraw part way through at day 3 and 7.

Time (t)	Number of Subjects Experiencing event	Subjects Censored	Number of Subjects Remaining n	Survival Probability n end / n start after censoring	Survival Function Calculation	Survival Estimate S(t)
0	0	0	10	10/10=1.0		1.000
1	1	0	9	9/10=0.90	= 0.90 x 1.00	0.900
4	1	1	7	7/8=0.875	= 0.875 x 0.900	0.788
5	2	0	5	5/7=0.714	= 0.714 x 0.788	0.563
8	1	1	3	3/4=0.750	= 0.750 x 0.563	0.422
10	1	0	2	2/3=0.667	= 0.667 x 0.422	0.281
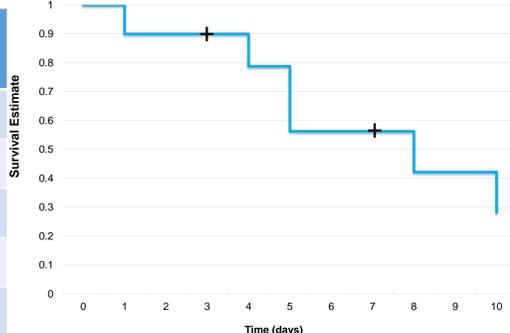
Note that the steps down in the survival estimate after censoring are bigger compared to above. This is because the censored subjects have been removed from the calculation.
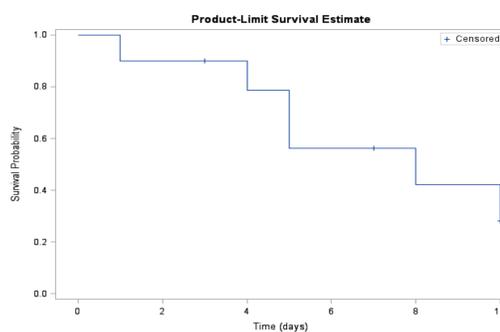
SAS® Proc Lifetest

To produce Kaplan-Meier plots in SAS, a subject level dataset is needed containing a time variable (e.g. aval) and a censor variable (e.g. cnsr) which should be set to 0 for subjects who experienced an event and 1 for subjects who were censored. The PROC LIFETEST code below will output the plot to the right:

```
proc lifetest data=ex1 outsurv=km1 plots=s;
    time aval*cnsr(1);
run;
```

Most clinical trials will compare treatments. To produce Kaplan-Meier plots comparing these, include a `strata` statement (see (4)).

EXAMPLE in SAS: As above plotted with PROC LIFETEST

4. Improvements

The plots option on PROC LIFETEST has limitations. Often, it is better to output the survival tables and then use PROC SGPLOT to customize the display. Using PROC LIFETEST in combination with ODS PRODUCTLIMITESTIMATES gives a dataset of the Kaplan-Meier Estimates and from this we build the plot manually.

Using a PROC SGPLOT and a SERIES statement, we can manually plot the Kaplan-Meier as survival vs aval (Time). From there we can start to make changes to improve the plot:

- Split by treatment group to compare effect.
  `strata trtp`
- Remove censoring marks that obscure the trend.
  `cs=none`
- Focus on relevant area of the plot by changing the axis scale and choosing appropriate tick marks
  `proc sgplot`
- Add the number of subjects at risk for each group over time, at relevant time intervals.
  `xaxistable`
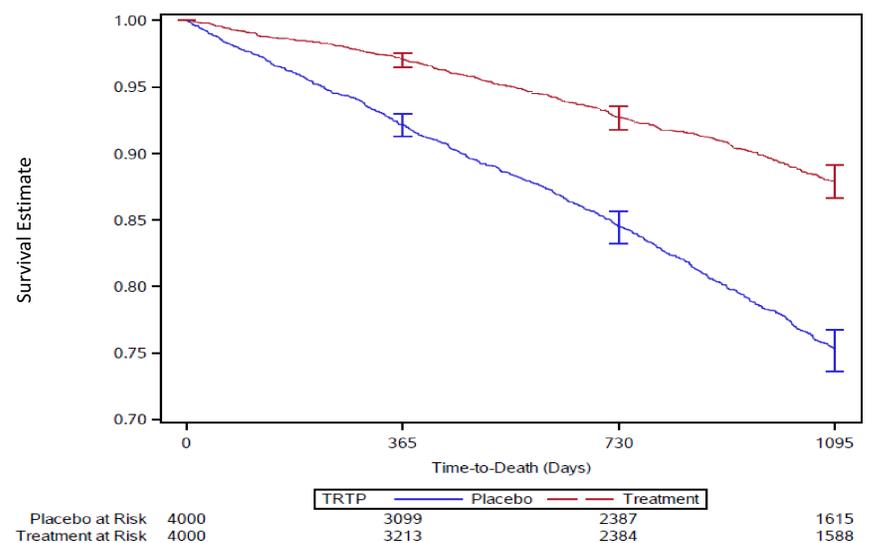- Add confidence intervals at relevant time intervals to compare groups
  `scatter x=aval y=survival`

EXAMPLE in SAS: 8,000 subjects randomized to Placebo or Treatment (trtp) for 3 years with a primary endpoint of death. Subjects can have event (die), complete (censored at 3 years) or withdraw (censor before 3 years)

```
ods output productlimitestimates=surv1;
proc lifetest data=adtte outsurv=outsurv cs=none;
    time aval*cnsr(1);
        strata trtp;
run;
ods output close;

proc sgplot data=surv2;
    series x=aval y=survival / group=trtp;
    scatter x=aval y=survival / group=trtp yerrorlower=lower yerrorupper=upper;
    xaxis label='Time-to-Death(Days)' min=0 values=(0 to 1095 by 365) max=1095;
    yaxis label='Survival Estimate' values=(0.7 to 1 by 0.05);
    xaxistable plarisk trtrisk;
run;
```

See accompanying paper to see dataset structure and manipulation of surv1 to surv2

TRTP	Placebo	Treatment
Placebo at Risk 4000	3099	2387	1615
Treatment at Risk 4000	3213	2384	1588

5. Assumptions and Interpretation

- The Kaplan-Meier plot shows an estimate of the probability of surviving to a given time.
- A higher survival estimate suggests a longer time to death, not that the death is less likely to occur overall.
- Can represent as survival (going down) or time to event (going up) – however curve can only go in one direction.
- Assumes censored subjects have the same probability of survival as those who continue.
- Before any subjects are censored, the estimate is an exact representation of the data.
- More frequent follow-up of subjects increases accuracy of estimate.
- Due to censoring, the left hand side of the graph will have greater precision.
- Small subsets can be misleading due to high effect of a single event.
- Kaplan-Meier curves are limited in that multiple predictors cannot be taken into account. Theoretically, randomisation reduces this issue, however use of a Cox PH model (see (6)) with covariates may be preferred.

6. Cox PH Analysis

Cox proportional hazards is a regression model for investigating the effect of variables on time-to-event data, using survival times and censoring. When the assumptions of Cox regression are met it provides a better estimate of survival probabilities than Kaplan-Meier.

Cox PH analysis outputs 'hazards' which can be thought of as instantaneous risks. A Hazard Ratio is used to show the difference between groups. It is equivalent to the ratio of the probability of an event between the two groups:

EXAMPLE in SAS: PROC PHREG using the Placebo group as reference

- Hazard Ratio = 0.5: half as many subjects in the treatment group are experiencing an event compared to the control group.
- Hazard Ratio = 1: event rates are the same in both groups
- Hazard Ratio = 2: twice as many subjects in the treatment group are experiencing an event compared to the control group.
- Reduction in risk = 1 – HR (usually converted to %).

Covariates

As with linear regression modelling, Cox PH models allows the use of covariates – variables that could also be associated with the outcome of interest. By including covariates our estimate of the treatment effect is improved.

In the example here we have two possible covariates: Age and Sex. However this data comes from a randomised clinical trial, so we would not expect to see much change in the hazard ratios.

EXAMPLE in SAS: PROC PHREG with covariates

Results

Using the example from above, without Covariates, comparing Treatment against the reference group of Placebo, gives a hazard ratio of 0.441 (0.388, 0.502). This means that there is a 56% reduction in the risk of dying for subjects in the treatment group compared to the placebo group. With a p-value <0.0001 this can be considered significant.

When the covariates Age and Sex are included, as expected, there is little change in the hazard ratio. The reduction in risk is again 56% for subjects in the treatment group compared to the placebo group.

SAS® Proc PHREG

```
proc phreg data=adtte;
    class trtp(ref='Placebo');
    model aval*cnsr(1)=trtp / risklimits;
    hazardratio trtp;
run;
```

	Hazard Ratio (95% CI)	Reduction in risk of Death	p-value
Treatment vs. Placebo	0.441 (0.388, 0.502)	55.9% (49.8%, 61.2%)	<.0001

SAS® Proc PHREG – Covariates

```
proc phreg data=adtte;
    class sex age trtp(ref='Placebo');
    model aval*cnsr(1)=trtp sex age / risklimits;
    hazardratio trtp;
    ods output ParameterEstimates=coxph;
run;
```

	Hazard Ratio (95% CI)	Reduction in risk of Death	p-value
Treatment vs. Placebo	0.436 (0.383, 0.496)	56.4% (50.4%, 61.7%)	<.0001