Introduction to Survival Analysis and Creating Kaplan-Meier Curves
Ojesh Upadhyay, PPD Inc., Hamilton, NJ

ABSTRACT
A non-technical approach will be employed to introduce SAS programmers to the basic statistical concepts underlying survival analysis and to discuss common issues encountered when creating Kaplan-Meier curves. Programmers will become familiar with how the Kaplan-Meier estimates are derived at each timepoint, how to use the ANNOTATE option to build customized graphs and the OVERLAY option in PROC GPLOT to mark only censored subjects and not subjects with events.

INTRODUCTION
Most programmers are aware that they can use PROC LIFETEST to generate Kaplan-Meier estimates, but they may not be aware of exactly how these estimates are being generated. Some may be wary of statistical jargon and of statistics and mathematics in general. This presentation is designed for just such a person. The goal of this presentation will be to close this “knowledge gap” in the most non-technical manner possible.

The presentation will be divided into two parts. The first part will focus on basic statistical concepts and calculations involved in generating Kaplan-Meier estimates. Attendees will understand why the Kaplan-Meier is such a widely used and preferred method in survival analysis, how it deals with censored subjects, what its limitations are and how we should interpret its results. Simple, fictional datasets created by the author will be used to assist in illustrating these concepts.

If time permits, the second part will focus on the actual creation of the Kaplan-Meier Curve using PROC GPLOT in SAS. Attendees will become familiar with the ANNOTATE option to create customized features, such as boxes/legends containing survival estimates at particular timepoints or hazard ratios when comparing two treatment groups. The use of the OVERLAY option to mark only censored subjects and not subjects with events will be illustrated.

THE PROBLEM
You are granted funds to conduct a “clinical trial” to determine the “efficacy” of a new online dating site, Harmonize_Me.com versus that of an existing site, Match_Me.com. Assumptions are as follows:

- The goal of everyone who registers for the sites is to find a partner – the sooner the better. Thus, the company with the least average interval between registration and finding a partner is considered to be the most “efficacious”.
- Not everybody who registers “completes” the study (i.e., they never log in after a certain timepoint and their status cannot be confirmed after this timepoint).
- Subject enter the study at the time of their choosing. To keep the analysis simple, we will choose the next five subjects that register for each site, for a total of ten subjects in the study. For simplicity, we will assume that these subjects were randomly selected. Analysis will be done one year after the first subject registers. Thus, some subjects will be in the study for a longer period than others.

One year after the first subject registers, you have the following dataset “DATE”:

```sas
data DATE;
  input SUBJ $ TRT $ DUR EVENT;
datalines;
01 HARMO 0.5 1
02 HARMO 4.5 1
03 HARMO 8.5 1
04 HARMO 1.5 0
05 HARMO 2.5 0
06 MATCH 0.5 1
07 MATCH 4.5 1
08 MATCH 8.5 1
09 MATCH 5.5 0
```
SUBJ is the subject; TRT is the “treatment” group (MATCH = Match_Me.com, HARMO = Harmonize_Me.com); DUR is duration in months from registration to finding a partner or from registration to last known follow up date; EVENT is the subject’s status as of their last known follow up date (1 means found partner, 0 means did not find partner).

Let’s take a closer look at the data:

<table>
<thead>
<tr>
<th>SUBJ</th>
<th>TRT</th>
<th>EVENT</th>
<th>DUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>HARMO</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>02</td>
<td>HARMO</td>
<td>1</td>
<td>4.5</td>
</tr>
<tr>
<td>03</td>
<td>HARMO</td>
<td>1</td>
<td>8.5</td>
</tr>
<tr>
<td>04</td>
<td>HARMO</td>
<td>0</td>
<td>1.5</td>
</tr>
<tr>
<td>05</td>
<td>HARMO</td>
<td>0</td>
<td>2.5</td>
</tr>
<tr>
<td>06</td>
<td>MATCH</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>07</td>
<td>MATCH</td>
<td>1</td>
<td>4.5</td>
</tr>
<tr>
<td>08</td>
<td>MATCH</td>
<td>1</td>
<td>8.5</td>
</tr>
<tr>
<td>09</td>
<td>MATCH</td>
<td>0</td>
<td>5.5</td>
</tr>
<tr>
<td>10</td>
<td>MATCH</td>
<td>0</td>
<td>9.5</td>
</tr>
</tbody>
</table>

At first glance, it looks like the average time to find a partner is the same for both groups, 4.5 months=(0.5+4.5+8.5)/3. From this, one might conclude that there is no significant difference between the two dating sites. However, note that the two subjects either lost to follow up or still single at the one year anniversary of the study in each group have significantly different DUR. Subjects lost to follow or those who have not had an event as of the analysis date are referred to as censored subjects. The two censored subject in the “HARMO” group were single after 1.5 and 2.5 months after registration, while the two censored subjects in the “MATCH” group were single after 5.5 and 9.5 months. This opens up the possibility that at least some of the “HARMO” censored subjects married soon after being lost to follow up (or if they signed up recently, will marry soon afterwards). If they did, then it would seem that Harmonize_Me.com might be the superior dating site. However we do not know with certainty what these censored subjects did. It is just this sort of situation where the Kaplan-Meier method comes in handy.

THE SOLUTION

The Kaplan-Meier method calculates the cumulative proportion single for Harmonize_Me.com as follows:

<table>
<thead>
<tr>
<th>Interval (Months)</th>
<th>At Risk</th>
<th>Censored</th>
<th>Found Partner</th>
<th>Proportion Single</th>
<th>Cumulative Proportion Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>4/5 = 0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>1 – 2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>3/3 = 1.00</td>
<td>0.80 * 1.00 = 0.80</td>
</tr>
<tr>
<td>2 – 3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2/2 = 1.00</td>
<td>0.80 * 1.00 = 0.80</td>
</tr>
<tr>
<td>3 – 4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2/2 = 1.00</td>
<td>0.80 * 1.00 = 0.80</td>
</tr>
<tr>
<td>4 – 5</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1/2 = 0.50</td>
<td>0.80 * 0.50 = 0.40</td>
</tr>
<tr>
<td>5 – 6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1/1 = 1.00</td>
<td>0.40 * 1.00 = 0.40</td>
</tr>
<tr>
<td>6 – 7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1/1 = 1.00</td>
<td>0.40 * 1.00 = 0.40</td>
</tr>
<tr>
<td>7 – 8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1/1 = 1.00</td>
<td>0.40 * 1.00 = 0.40</td>
</tr>
<tr>
<td>8 – 9</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0/1 = 0.00</td>
<td>0.40 * 0.00 = 0.00</td>
</tr>
<tr>
<td>9 – 10</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

0.00
For Match_Me.com, the calculations are as follows:

<table>
<thead>
<tr>
<th>Interval (Months)</th>
<th>At Risk</th>
<th>Censored</th>
<th>Found Partner</th>
<th>Proportion Single</th>
<th>Cumulative Proportion Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>4/5 = 0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>1 – 2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4/4 = 1.00</td>
<td>0.80 * 1.00 = 0.80</td>
</tr>
<tr>
<td>2 – 3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4/4 = 1.00</td>
<td>0.80 * 1.00 = 0.80</td>
</tr>
<tr>
<td>3 – 4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4/4 = 1.00</td>
<td>0.80 * 1.00 = 0.80</td>
</tr>
<tr>
<td>4 – 5</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>3/4 = 0.75</td>
<td>0.80 * 0.75 = 0.60</td>
</tr>
<tr>
<td>5 – 6</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2/2 = 1.00</td>
<td>0.60 * 1.00 = 0.60</td>
</tr>
<tr>
<td>6 – 7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2/2 = 1.00</td>
<td>0.60 * 1.00 = 0.60</td>
</tr>
<tr>
<td>7 – 8</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2/2 = 1.00</td>
<td>0.60 * 1.00 = 0.60</td>
</tr>
<tr>
<td>8 – 9</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1/2 = 0.50</td>
<td>0.60 * 0.50 = 0.30</td>
</tr>
<tr>
<td>9 - 10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

The following code outputs Figure 1 below:

```
proc lifetest data=date plots=(s);
time dur*event(0);
strata trt;
run;
```

![Figure 1. Proportion Single vs. Time Since Registration](chart)

Figure 1. Proportion Single vs. Time Since Registration
The above calculations and corresponding survival curve suggest that subjects enrolled in MATCH have greater chance of “surviving”, which in our case is actually a bad thing since surviving really means remaining single (no doubt that many would argue this point). Let’s examine the reason that this happens:

- At one month, subjects in both sites have 4/5 or 80% chance of remaining single. Nobody was censored and one subject married during the first month after registration for both studies. At this point, four subjects are “at risk” of finding a partner.

- Between one and two months, one subject is censored for the HARMO group, while none are censored for the MATCH group. Neither group has anyone finding a partner during this interval. The censored subject is removed from the at risk pool and the proportion remaining single in the HARMO group is 3/3 or 100%; for the MATCH group its 4/4 or 100%. The cumulative proportion is calculated by multiplying the cumulative proportion from the previous interval with the proportion remaining single during the current interval.

- The cumulative proportion is also referred to as the product limit estimator. It is important to note that these numbers are only estimates of “survival” (singlehood in our case). We cannot be certain of what the actual survival rates are until we can confirm when all of the subjects “died” (found a partner in our case).

- Between the two and three month interval, we again have one subject censored in the HARMO group, thus 2 are at risk starting the next interval. Four are still at risk in the MATCH group.

- Between the four and five month interval, one subject in each group is married. The effect of censoring is seen clearly during this interval. In the HARMO group, only two subject were at risk at the beginning of the interval - one subject finding a partner means that the chance of remaining single is 1/2 or 50% during this interval. For the MATCH group, nobody has been censored thus far, thus the one subject marrying means a chance of 3/4 or 75% of remaining single during this period. Thus, at this point, only 40% (0.80 X 0.50) of the HARMO group is single, while 60% (0.80 X 0.75) of the MATCH group is still single.

- From the above we see that censoring decreases the numerator and thus an event will lead to an increased impact on the survival curve, as seen by more steep drops in the curve. We also note that censoring does not lead to a step down on the survival curve – this occurs when an event occurs. The only way we can know when a subject was censored is if there are tick marks on the curve to indicate censorship.

- Censoring subjects is a “necessary evil” to deal with real world situations – it is not desired as it leads to a decrease in the sample size. A study with a high number of censored patients and especially a high number of early censored patients will lead to an unreliable survival curve. This is why many study designs incorporate a minimum follow up for all patients and the analysis is not done until all patients are followed up for the defined minimum length of time.

As seen above in Figure 1, PROC LIFETEST is able to output a basic survival curve. However, the user has little control if he/she wants to customize the graph. The code below shows how using the ANNOTATE and OVERLAY option in SAS can output a highly customized graph:

```sas
ods listing close;
ods output productlimitestimates=est;
ods output quartiles=quart;
proc lifetest data=DATE;
time DUR*EVENT(0);
strata TRT;
run;
ods listing;

data est2;
set est;
retain surv2;
if survival ne . then surv2=survival;
run;

data est3;
set est2;
by stratum;
if stratum=1 then do;
  if censor=1 then c1=surv2;
  if censor=0 then s1=surv2;
```

As seen above in Figure 1, PROC LIFETEST is able to output a basic survival curve. However, the user has little control if he/she wants to customize the graph. The code below shows how using the ANNOTATE and OVERLAY option in SAS can output a highly customized graph:
if last.stratum then s1=surv2;
end;
if stratum=2 then do;
   if censor=1 then c2=surv2;
   if censor=0 then s2=surv2;
   if last.stratum then s2=surv2;
end;
run;

data q2;
set quart;
where percent=50;
*drop stratum;
length est2 lo2 up2 $8 cnum $20;
if estimate=. then est2='NE';
else est2=compress(put(estimate,8.1));
if lowerlimit=. then lo2='NE';
else lo2=compress(put(lowerlimit,8.1));
if upperlimit=. then up2='NE';
else up2=compress(put(upperlimit,8.1));
cnum=compress(est2) || '(' || compress(lo2) || ',' || compress(up2) || ')';
call symput("med"||put(_n_,1.),cnum);

%let med1=%trim(&med1);
%let med2=%trim(&med2);
%put &med1 &med2;

data date2;
set date;
if trt='HARMO' then trtn=1;
else trtn=2;
run;
ods listing close;
ods output parameterestimates=haz;
proc phreg data=date2;
model dur*event(0) = trtn / rl;
run;
ods listing;

data haz2;
set haz;
length hr2 lo2 up2 $20 cnum $20;
hr2=compress(put(hazardratio,8.2));
lo2=compress(put(hrlowercl,8.2));
up2=compress(put(hruppercl,8.2));
cnum=compress(hr2) || '(' || compress(lo2) || ',' || compress(up2) || ')';
call symput("haz"||put(_n_,1.),cnum);
%put &haz1;

data anno1;
length function $8 text $200;
function='label'; position='3'; x=24.00; y=2.00; style='markere'; text='U';
size=1;
output;
function='label'; position='3'; x=27.00; y=2.00; size=0.7; style='';
text="Censored";
output;

data anno2;
set est2;
retain xsys ysys;
length function $8 text $200;
xsys='2'; ysys='2';
x=dur; y=surv2;
function='label'; position='3'; x=3.7; y=0.25; text="HARMO"; output;
function='label'; position='3'; x=6.6; y=0.25; text="MATCH"; output;
function='label'; position='3'; x=0.8; y=0.2; text="Median (95% CI)"; output;
function='label'; position='3'; x=3.4; y=0.2; text="&med1"; output;
function='label'; position='3'; x=6.3; y=0.2; text="&med2"; output;
function='label'; position='3'; x=0.8; y=0.15; text="Hazard Ratio (95% CI)"; output;
function='label'; position='3'; x=6.3; y=0.15; text="&haz1"; output;
function='move'; position='3'; x=0.6; y=0.12; output;
function='draw'; position='3'; x=8.3; y=0.12; output;
function='draw'; position='3'; x=8.3; y=0.30; output;
function='draw'; position='3'; x=0.6; y=0.30; output;
function='draw'; position='3'; x=0.6; y=0.12; output;
run;

data anno2a;
length function $8;
function='move'; x=4;y=7.0; output;
function='draw'; x=70;y=7.0; output;
function='draw'; x=70;y=10.3; output;
function='draw'; x=4;y=10.3; output;
function='draw'; x=4;y=7.0; output;
run;

data anno3;
set anno1 anno2 anno2a;
run;

ods listing close;
ods output productlimitestimates=test3;

proc lifetest data=date method=km
TIMELIST=0 2 4 6 8 10;
time dur*event(0);
strata trt;
run;

ods listing;
data test4;
set test3(keep=stratum trt timelist left);
if _n_ lt 10 then do;
  call symput("bign"||put(_n_,1.),compress(put(left,4.))); 
end;
else do;
  call symput("bign"||put(_n_,2.),compress(put(left,4.))); 
end;
run;
%let bign1=%trim(&bign1);
%let bign2=%trim(&bign2);
%let bign3=%trim(&bign3);
%let bign4=%trim(&bign4);
%let bign5=%trim(&bign5);
%let bign6=%trim(&bign6);
%let bign7=%trim(&bign7);
%let bign8=%trim(&bign8);
%let bign9=%trim(&bign9);
%let bign10=%trim(&bign10);
%let bign11=%trim(&bign11);
%let bign12=%trim(&bign12);
%put &bign1 &bign6 &bign12;

data _null_; 
call symput("datef",put(datetime(),datetime19.)); 
run; 
%put &datef; 
%let lspace=0.5;
filename test1 "C:\Documents and Settings\upadhyor\My Documents\PharmaSUG 2010\graph.cgm";
goptions reset=all 
   rotate=landscape 
   device=cgmo97l 
   gsfmode=replace 
   gsfname=test1 
   htext=0.6 
   htitle=0.6 
   ftext=hwcgm009 
   hsize = 11 in 
   vsize = 7.5 in 
; 
axis1
   minor=none 
   label=(angle=90 h=0.6 "Survival Distribution Function"); 
axis2
   minor=none 
   order = 0 to 10 by 2 
   value=(tick=1 '0' tick=2 '2' tick=3 '4' tick=4 '6' tick=5 '8' tick=6 '10') 
   label=(h=0.6 justify=center 'Months' 
     justify=left 'Subjects at Risk:   Month 0    Month 2    Month 4    Month 6    Month 8    Month 10' 
     justify=left 'HARMO:                &bign1          &bign2          &bign3 &bign4 &bign5 &bign6" 
     justify=left 'MATCH:                &bign7          &bign8          &bign9 &bign10 &bign11 &bign12"
     justify=left');
symbol1
proc gplot data=est3;
plot s1*dur s2*dur c1*dur c2*dur/ overlay vaxis=axis1 haxis=axis2 hminor=0
      nolegend noframe anno=anno3;
run;
quit;
The above code outputs a CGM file which can be copied and pasted onto a RTF or DOC file. The resulting graph follows:

**Figure 1:**
Duration of Singlehood (Months) by Treatment
All Randomized Population

**CONCLUSION**
Survival analysis, also called time to event analysis, is used widely not only in the life sciences but also in the social sciences (for example, length of non-recidivism for subjects receiving alcohol counseling) and engineering (for example, how long before machine fails). Readers interested in the topic are encouraged to explore concepts related to survival analysis such as the hazard rate, relative risk and the odds ratio.

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

Name: Ojesh Upadhyay  
Enterprise: PPD Inc.  
Address: 3575 Quakerbridge Rd, Suite 201  
City, State ZIP: Hamilton, NJ 08619  
Work Phone: 609-528-8051  
Fax: 609-584-0294  
E-mail: ojesh.upadhyay@ppdi.com  
Web: www.ppd.com