

## LI Analysis Training Series

### Survival Analysis/Life Tables

(Last Revised: 4/17/2007 (2/17/2000))

Melissa Ives, Rodney Funk & Michael Dennis  
Chestnut Health Systems  
Bloomington IL 61701  
309-827-6026  
[www.chestnut.org](http://www.chestnut.org)

**Acknowledgement:** This document was developed under contract #270-2003-00006 from the Substance Abuse and Mental Health Services Administration (SAMHSA's) Center for Substance Abuse Treatment (CSAT). Any opinions about this data are those of the authors and do not represent official positions of the government or individual grantees.

**Purpose:** Survival analysis, including Life Tables and Hazard analysis, is a way of examining the time to an event (e.g., discharge, admission/readmission, symptom onset, task completion, death) for one or more groups. Survival analysis considers not simply whether an event might occur, but also considers when it might occur (See Willett and Singer, 1991). It accounts for cases where data about the terminal event is unavailable (censored) due to client attrition or study cut-off dates. The procedure provides two main pieces of information: the survival function—the proportion of the population 'surviving' a given time interval (i.e., who have not reached the terminal event), and the hazard function—the proportion of the surviving population who are likely to reach the event within the interval (i.e., risk of the terminal event occurring). Survival analysis has also been referred to as event history analysis or failure-time analysis.

**Background:** Originally used to study physical illnesses and medical outcomes (Cox, 1972), survival/hazard analysis has also been used in substance abuse and mental health to consider length of stay (CITATION—Bill F. has several), time until relapse/readmission (Bill F. again), onset of symptoms (Martin, Langenbucher, Kaczynski and Chung (1996), and number of events until follow-up (Dennis, et. al., 1999). For more information on the history and uses of survival analysis, see Luke (1993). For this project, there are two primary reasons to use survival analysis, retention in treatment (length of stay) and time until relapse. Each of these may be examined by site, by case mix or by selected demographics. Procedures mentioned below may be found in Chapters 8-11 and 21-23 of *SPSS Advanced Statistics 7.5*. Syntax guides may be found in the same volume on pages 319-331, 395-403 and 514-526. Examples given in this paper use the Survival commands (Chapter 8 and pp. 514-526). More recent information may be found in *SPSS Advanced Models 10.0* (SPSS, 1999), Chapters 19-21

**Data Requirements:** There are two required elements to conduct a survival analysis: a time variable, and a status variable. The time variable indicates the time between two events (e.g. discharge and readmission, drinking and abuse symptoms) in days, months, years, minutes or seconds. Time variables should contain continuous, interval level data.

The status variable indicates whether the terminal event has occurred. Status variables may also include indicators of reasons for attrition without reaching the terminal event (censoring). Status variables are generally categorical and frequently dichotomous, but may be continuous when using Kaplan-Meier (KM) or Cox Regression (COXREG) analyses.

Survival procedures assume that censored and uncensored cases do not differ in terms of the likelihood of the terminal event. It is also generally assumed that the likelihood of the terminal event does not depend upon when the time of the entry event. When this assumption is violated (for example, spelling ability and age for elementary school children) survival analysis may be performed using Cox Regression with Time Dependent Covariates (Cox W/Time Dep Cov).

### **Procedures:**

Life Tables (Survival): Use this procedure for larger samples where the time intervals are large enough to be broken down into smaller units.

Kaplan-Meier (KM) Survival Analysis (also known as product-limit (Luke, 1993)): Use this procedure for smaller samples with discrete lengths of time.

Cox Regression (COXREG): Use this procedure for analyses with covariates. (A variation also exists for Cox Regression with time-dependent factors.)

### **Example 1:**

This example will present a Life Table analysis.

To run a survival analysis using the menus, choose the following:

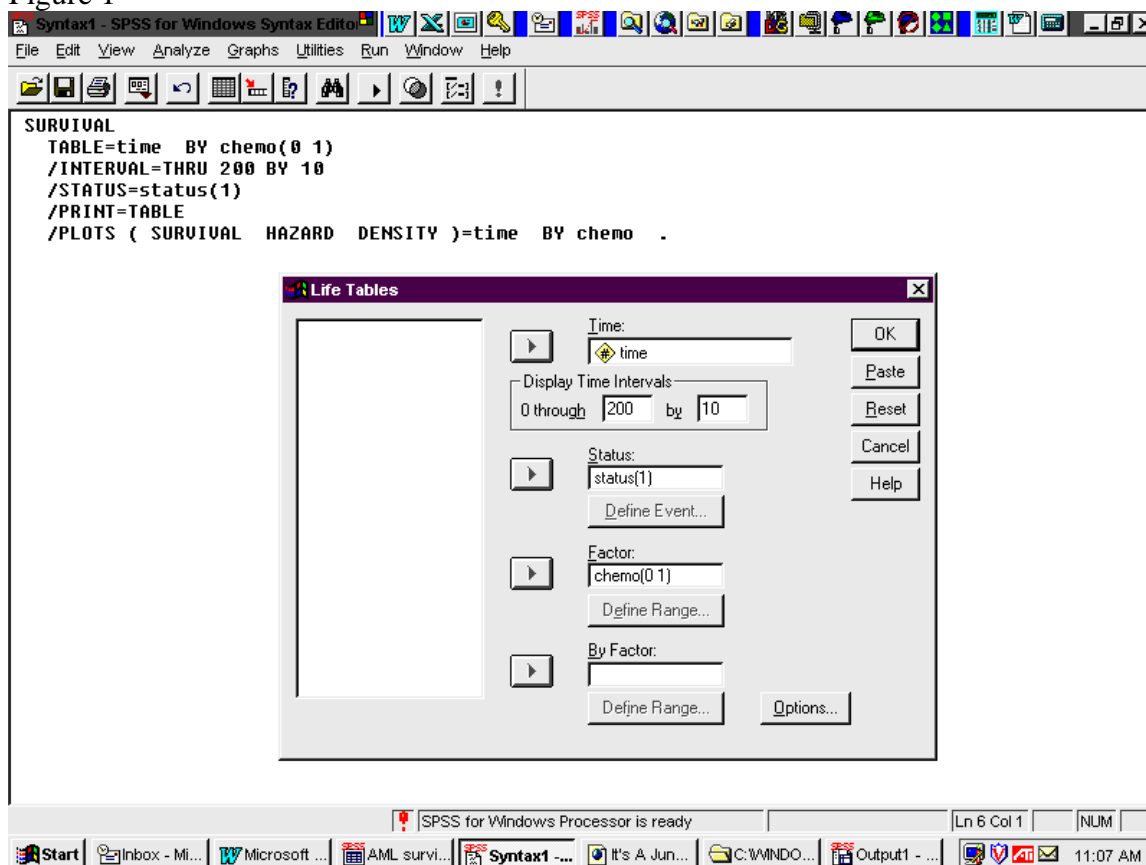
Analysis

Survival

Life Tables (or one of the other choices)

Figure 1 shows the dialog box and pasted syntax for a sample data file (SPSS's sample data set AML survival.sav). In this example, the dataset consisted of only the three variables, time, status and 'chemo'.

Figure 1



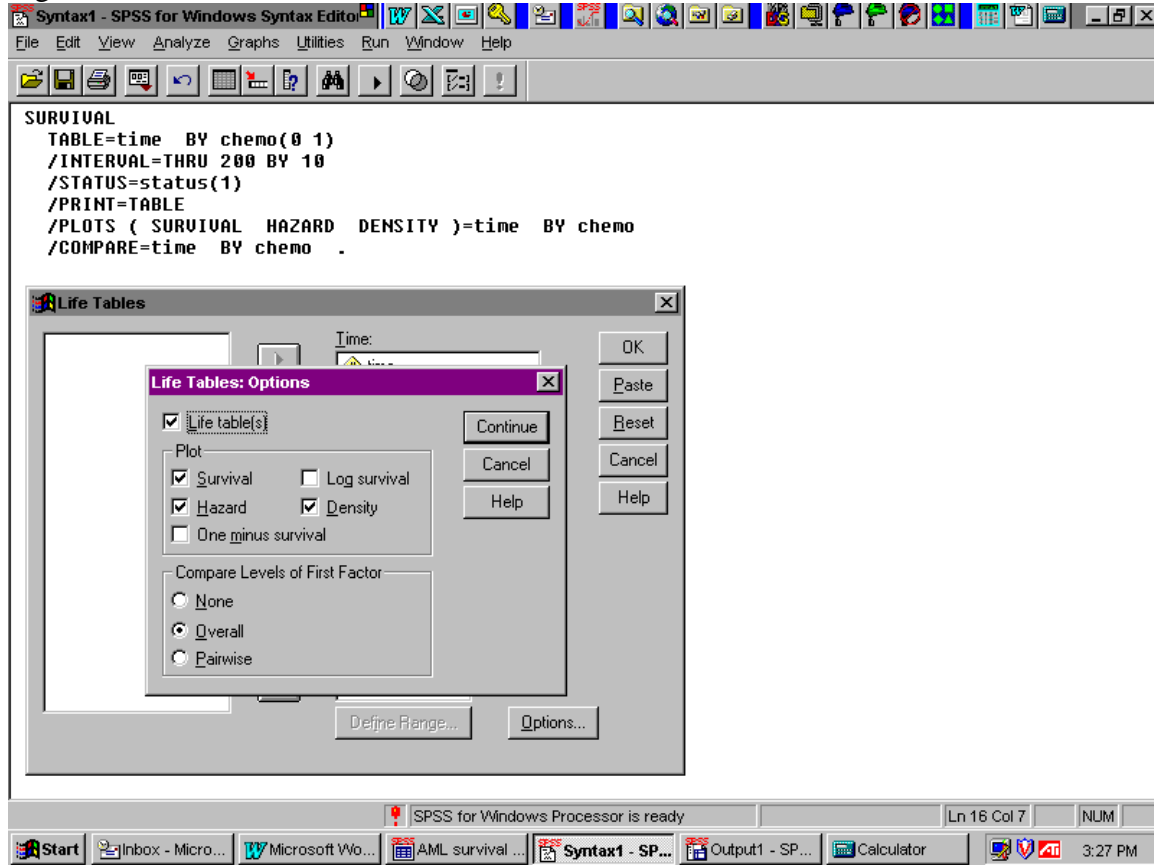
Within the dialog box, the steps are:

- 1) Select the variable containing the time data (e.g., days, months, minutes) and move it to the “Time” box by clicking the arrow to the left of the box.
- 2) In the “Display Time Intervals” boxes below time, enter the upper limit of the time variable and the interval at which to compute the statistics. For example, if the time variable is a count of days, to show monthly information for one year, enter 360 and 30. The interval upper limit must be evenly divisible by the interval. Although it is not available through the dialog boxes, you may divide the time by different intervals by adding a second ‘THRU’ command. For example, to consider daily rates for the first month and monthly for one year thereafter, change the /INTERVAL line to read: /INTERVALS=THRU 30 BY 1 THRU 360 BY 30.
- 3) Select the status variable and move it into the status field by clicking on the arrow to the left of the box. Then click on ‘Define Range’ and enter the value (or range of values) that represents the occurrence of the terminal event.

That is all that is required to run a survival analysis. However, if you wish to specify one (or two) comparison variable, select and move it to the Factor box(es) in the same manner as the time and status variables. You will need to specify the range of individual values to include in each factor. If a factor variable contains variables from 1 to 5 and you specify a range of 1 to 3, cases with values of 4 and 5 will be ignored.

Options In the dialog box, there is an Options button. Figure 2 shows the Options box requesting Survival, Hazard and Density plots and a comparison of the two chemo groups. (along with the pasted syntax in the background).

Figure 2



The comparison is important since it tests the significance of any difference between the survival functions for your factor levels. It will give you a Wilcoxon (Gehan) statistic (chi-square), degrees of freedom, and a p-value. (see SPSS, 1999, page 270 for further information.) The following is a sample of the output (note, this output was adjusted to a maximum time of 100 in intervals of 10). Note that there are two life tables, one for each value of the variable 'chemo'.

Available workspace allows for exact comparisons of 21845 observations

-

This subfile contains: 23 observations

```

Life Table
Survival Variable TIME      Time (weeks)
                for CHEMO   Chemotherapy
=                0 No

Number  Number  Number  Number  Cumul
Intrvl  Entrng  Wdrawn  Exposd   of
Start   this   During  to      Termnl  Propn  Propn  Propn  Proba-
Time   Intrvl  Intrvl  Risk    Events  Termi- Sur-  Surv  bility  Hazard
-----  -----  -----  -----  -----  -----  -----  -----  -----  -----
    
```

.0	12.0	.0	12.0	4.0	.3333	.6667	.6667	.0333	.0400
10.0	8.0	1.0	7.5	1.0	.1333	.8667	.5778	.0089	.0143
20.0	6.0	.0	6.0	2.0	.3333	.6667	.3852	.0193	.0400
30.0	4.0	.0	4.0	2.0	.5000	.5000	.1926	.0193	.0667
40.0	2.0	.0	2.0	2.0	1.0000	.0000	.0000	.0193	.2000

The median survival time for these data is 24.04

Intrvl Start Time	SE of Cumul Surviving	SE of Probability Densty	SE of Hazard Rate
.0	.1361	.0136	.0196
10.0	.1441	.0085	.0142
20.0	.1469	.0121	.0277
30.0	.1211	.0121	.0444
40.0	.0000	.0121	.0000

Life Table

Survival Variable	TIME	Time (weeks)							
for	CHEMO	Chemotherapy	= 1 Yes						
Intrvl Start Time	Number Entrng this Intrvl	Number Wdrwn During Intrvl	Number Exposed to Risk	Number of Termnl Events	Propn Terminating	Propn Surviving	Cumul Propn Surv at End	Proba-bility Densty	Hazard Rate
.0	11.0	.0	11.0	1.0	.0909	.9091	.9091	.0091	.0095
10.0	10.0	1.0	9.5	2.0	.2105	.7895	.7177	.0191	.0235
20.0	7.0	1.0	6.5	1.0	.1538	.8462	.6073	.0110	.0167
30.0	5.0	.0	5.0	2.0	.4000	.6000	.3644	.0243	.0500
40.0	3.0	1.0	2.5	1.0	.4000	.6000	.2186	.0146	.0500
50.0	1.0	.0	1.0	.0	.0000	1.0000	.2186	.0000	.0000
60.0	1.0	.0	1.0	.0	.0000	1.0000	.2186	.0000	.0000
70.0	1.0	.0	1.0	.0	.0000	1.0000	.2186	.0000	.0000
80.0	1.0	.0	1.0	.0	.0000	1.0000	.2186	.0000	.0000
90.0	1.0	.0	1.0	.0	.0000	1.0000	.2186	.0000	.0000
100.0+	1.0	1.0	.5	.0	.0000	1.0000	.2186	**	**

\*\* These calculations for the last interval are meaningless.

The median survival time for these data is 34.42

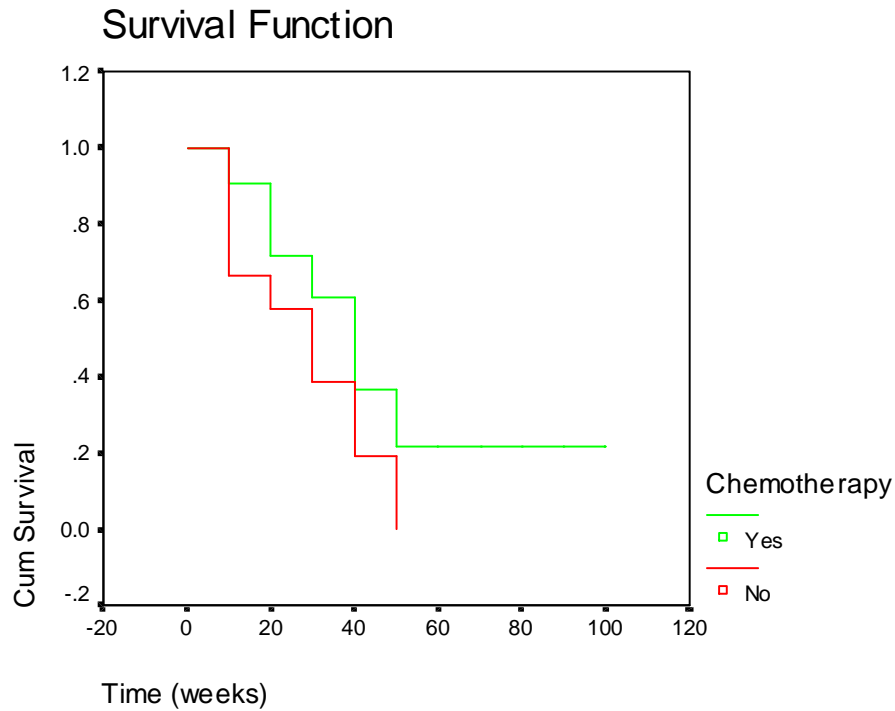
Intrvl Start Time	SE of Cumul Surviving	SE of Probability Densty	SE of Hazard Rate
.0	.0867	.0087	.0095
10.0	.1384	.0122	.0165
20.0	.1550	.0104	.0166
30.0	.1623	.0147	.0342
40.0	.1491	.0130	.0484
50.0	.1491	.0000	.0000
60.0	.1491	.0000	.0000
70.0	.1491	.0000	.0000
80.0	.1491	.0000	.0000
90.0	.1491	.0000	.0000
100.0+	.1491	**	**

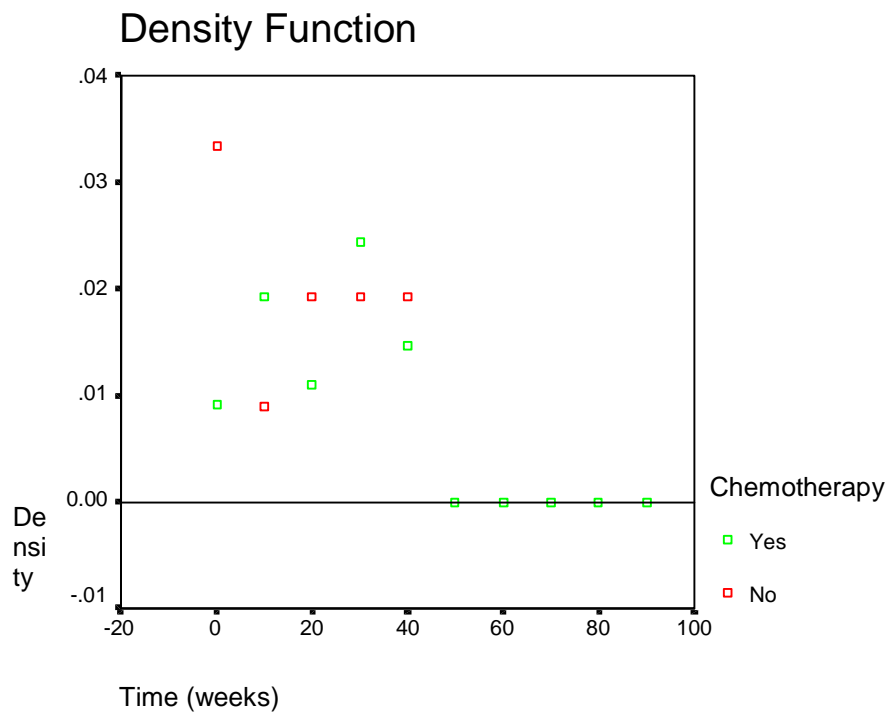
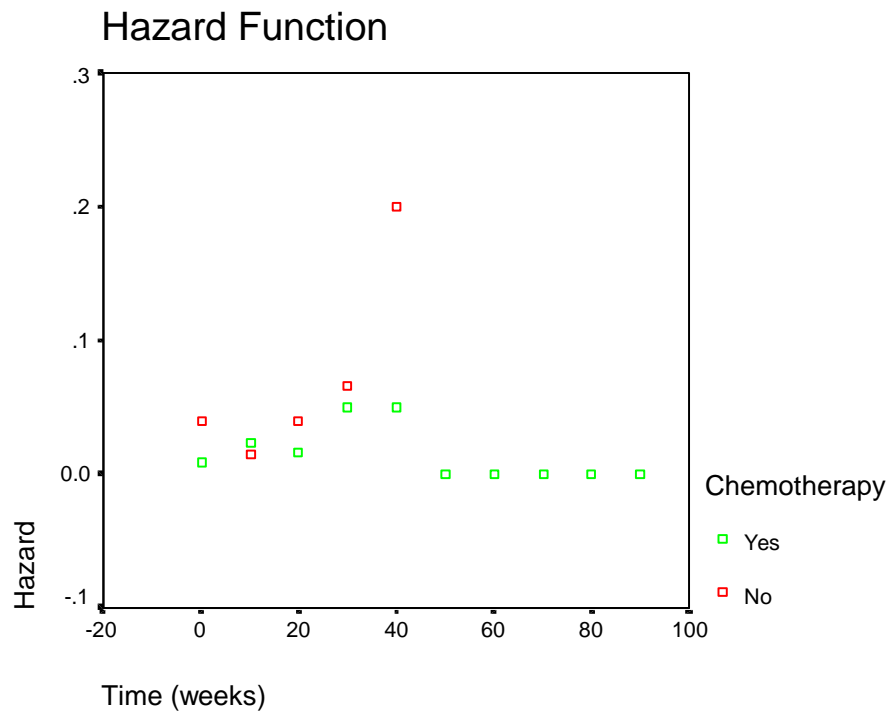
Comparison of survival experience using the Wilcoxon (Gehan) statistic

Survival Variable	TIME	Time (weeks)				
grouped by	CHEMO	Chemotherapy				
Overall comparison statistic	2.741	D.F.	1	Prob.	.0978	
Group label	Total N	Uncen	Cen	Pct Cen	Mean Score	

0	No	12	11	1	8.33	-4.1667
1	Yes	11	7	4	36.36	4.5455

The optional plots requested included the survival function, the hazard function, and the probability density function as presented below.





**Interpreting Example 1:** The life table presents several pieces of information. Below is a list of the items and a definition of each.

- 1) Interval start time: Indicates the beginning of the new time interval. In this example, time was expressed in weeks and displayed in 10-week groups (/Interval= subcommand).
- 2) Number entering this interval: The number of cases for whom the terminal event had NOT occurred at the beginning of the interval.
- 3) Number withdrawn during the interval: The number of cases for whom the terminal event had NOT occurred, but for whom no further data was available (censored cases).
- 4) Number exposed to risk: The number entering the interval minus half of the number withdrawn in the interval. In the 10-week interval of the 'No chemo' condition, 8 cases were entered and one was withdrawn. Thus, the number exposed to risk was 7.5. (Further discussion of this may be found in Luke (1993).)
- 5) Number of terminal events: The number of cases for whom the terminal event occurred within the interval.
- 6) Proportion terminating: The number terminating divided by the number exposed to risk.
- 7) Proportion surviving: 1 minus the proportion terminating.
- 8) Cumulative proportion surviving: Used to estimate the survival function (graph). Usually referred to as the probability of surviving at least until the beginning of the specified interval without experiencing the terminal event. It is NOT the same as the proportion surviving due to consideration of censored cases. For the 'Chemo' condition, 72% survived the 10-20 week interval. Calculated as the product of the proportion surviving for the current and all prior intervals.
- 9) Probability density: The estimated probability of the terminal event occurring during the interval.
- 10) Hazard Rate: Used to estimate the hazard function (graph). Also known as the event risk. The proportion of those who survived to a given interval who are expected to reach the terminal event within the interval. (For the hazard equation, see Luke, 1993 p 221).
- 11) SE of Cumulative Surviving: Estimate of the variability of the cumulative proportion surviving.
- 12) SE of Probability Density: Estimate of the variability of the probability density function.
- 13) SE of Hazard Rate: Estimate of the variability of the hazard function.

The survival function indicates the median survival for those in the 'no chemo' condition is between 30 and 40 weeks, and for the 'chemo condition it is between 40 and 50 weeks. The hazard function indicates that those in the 'no chemo' condition were less likely to survive (the hazard was higher) in the first 10 weeks (0-10) and after 20 weeks than the 'chemo' condition. In addition, the Wilcoxon comparison indicates that there is a slight but non-significant difference between the two groups.

### **Example 2:**

This example uses data from an early version of the GAIN-I. In this example, we look at length of stay (retention) in treatment first overall and then by race. First, we computed a length of stay (LOS) variable for all clients regardless of whether they were still in

treatment or not. For those who were still in treatment, LOS was computed as the time from their intake until the end of the study. For clients who dropped out, LOS was computed as the time from intake until the day they were determined to be dropped out and were discharged. The following syntax was used: (Discharge date existed only as a date field (dis\_date), intake date existed as both a date field (idate) and as individual variables for year (intk\_yy), month (intk\_mm) and day (intk\_dd).)

```
compute slos=((dis_date-idate)/86400)+1.
if (missing(slos) and intk_yy gt 0)
  slos=(yrmoda(98,06,30)-yrmoda(intk_yy,intk_mm,intk_dd)+1).
```

The next step is to identify, calculate, or create a status variable for the terminal event. In this case, discharge status. A variable with 3 values was created: 1=Discharged, 2=Dropped Out, 3=Still in treatment. In each case, values of 3 were censored, and values of 1 or 2 were considered terminal events.

The first analysis used the following syntax:

```
SURVIVAL
TABLE=slos
/INTERVAL=THRU 1200 BY 30
/STATUS=dtype (1 2)
/PRINT=TABLE
/PLOTS ( SURVIVAL Hazard )=slos .
```

The following is the top section (through 600 days) of output showing the key output variables. As noted below the data, the median overall survival time is just over 547 days. The bolded item shows the interval in which the median survival time falls.

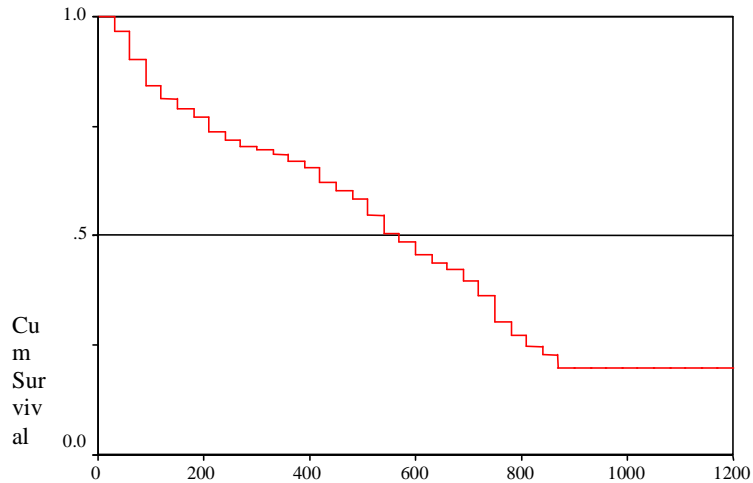
This subfile contains: 1015 observations

Life Table									
Survival Variable SLOS									
Intrvl Start Time	Number Entrng this Intrvl	Number Wdrawn During Intrvl	Number Exposd to Risk	Number of Termnl Events	Propn Termi- nating	Propn Sur- viving	Cumul Propn Surv at End	Proba- bility Densty	Hazard Rate
.0	1009.0	1.0	1008.5	33.0	.0327	.9673	.9673	.0011	.0011
30.0	975.0	1.0	974.5	67.0	.0688	.9312	.9008	.0022	.0024
60.0	907.0	1.0	906.5	58.0	.0640	.9360	.8431	.0019	.0022
90.0	848.0	1.0	847.5	32.0	.0378	.9622	.8113	.0011	.0013
120.0	815.0	19.0	805.5	20.0	.0248	.9752	.7912	.0007	.0008
150.0	776.0	31.0	760.5	20.0	.0263	.9737	.7704	.0007	.0009
180.0	725.0	42.0	704.0	29.0	.0412	.9588	.7386	.0011	.0014
210.0	654.0	41.0	633.5	17.0	.0268	.9732	.7188	.0007	.0009
240.0	596.0	33.0	579.5	11.0	.0190	.9810	.7052	.0005	.0006
270.0	552.0	25.0	539.5	7.0	.0130	.9870	.6960	.0003	.0004
300.0	520.0	30.0	505.0	8.0	.0158	.9842	.6850	.0004	.0005
330.0	482.0	17.0	473.5	11.0	.0232	.9768	.6691	.0005	.0008
360.0	454.0	15.0	446.5	9.0	.0202	.9798	.6556	.0004	.0007
390.0	430.0	9.0	425.5	22.0	.0517	.9483	.6217	.0011	.0018
420.0	399.0	15.0	391.5	11.0	.0281	.9719	.6042	.0006	.0009
450.0	373.0	12.0	367.0	12.0	.0327	.9673	.5845	.0007	.0011
480.0	349.0	20.0	339.0	21.0	.0619	.9381	.5483	.0012	.0021
510.0	308.0	20.0	298.0	24.0	.0805	.9195	.5041	.0015	.0028
540.0	264.0	7.0	260.5	9.0	.0345	.9655	<b>.4867</b>	.0006	.0012
570.0	248.0	24.0	236.0	14.0	.0593	.9407	.4578	.0010	.0020
600.0	210.0	16.0	202.0	8.0	.0396	.9604	.4397	.0006	.0013

The median survival time for these data is 547.06

The survival function is often depicted with a gridline showing the median—the point at which 50% of those in the study have reached the terminal event. In this case, that point is where 50% of the clients had been discharged or dropped out, close to 600 days. The line does not reach 0 because some long stay clients have not yet been discharged. Note that it is important to have accurate time data. In this case, if a client was missing a discharge date, it had to be assumed that he/she had not been discharged.

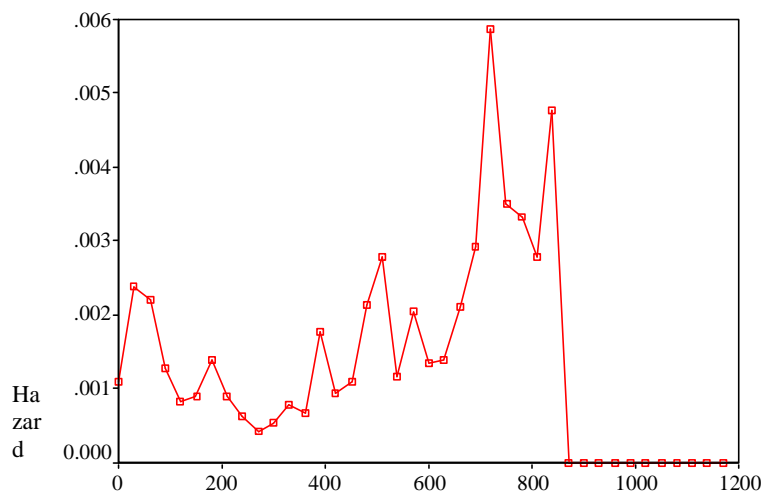
Figure 1. Survival Function



SLOS

The hazard function indicates that there is a somewhat high probability of being discharged early in treatment. This probability is fairly low until just over one year, when it begins to rise again. The probability of discharge is highest around two years.

Figure 2. Hazard Function



SLOS

In order to compare by race, the following syntax was used:

```
SURVIVAL
  TABLE=slos by race (4 5)
  /INTERVAL=THRU 1200 BY 30
  /STATUS=dtype (1 2)
  /compare slos by race
  /PRINT=TABLE
  /PLOTS ( SURVIVAL Hazard )=slos by race .
```

The output (not shown) lists two life tables, one for each value of race requested (in this case, 4-black and 5-white). The survival and hazard functions are shown below. In addition, the /COMPARE subcommand resulted in the following (appearing immediately after the life tables in the same text output.) This indicates that there is a significant difference between clients of different racial groups. When looking at the survival and hazard functions, it is clear that this may be due to a number of black clients who were not discharged after long stays.

Comparison of survival experience using the Wilcoxon (Gehan) statistic

Survival Variable SLOS		Race				
grouped by RACE		Race				
Overall comparison		statistic	7.448	D.F.	1	Prob. .0064
Group	label	Total N	Uncen	Cen	Pct Cen	Mean Score
4	Black	803	378	425	52.93	20.0847
5	White	195	98	97	49.74	-82.7077

Figure 3 Survival Function by Race

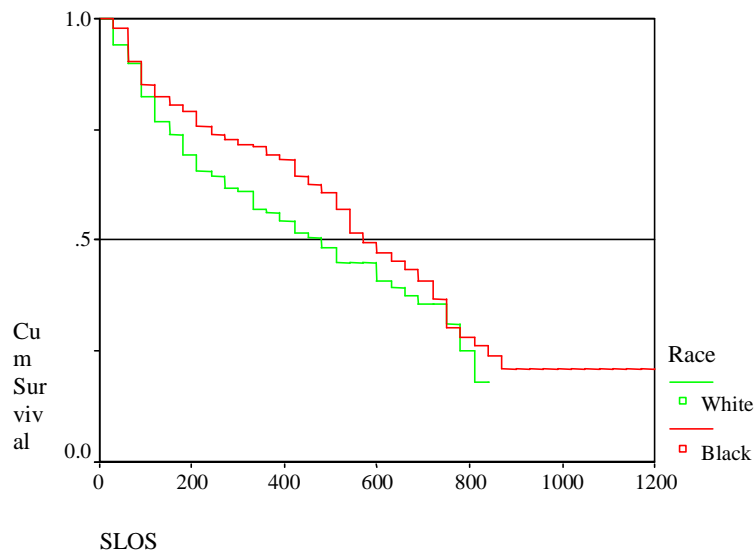
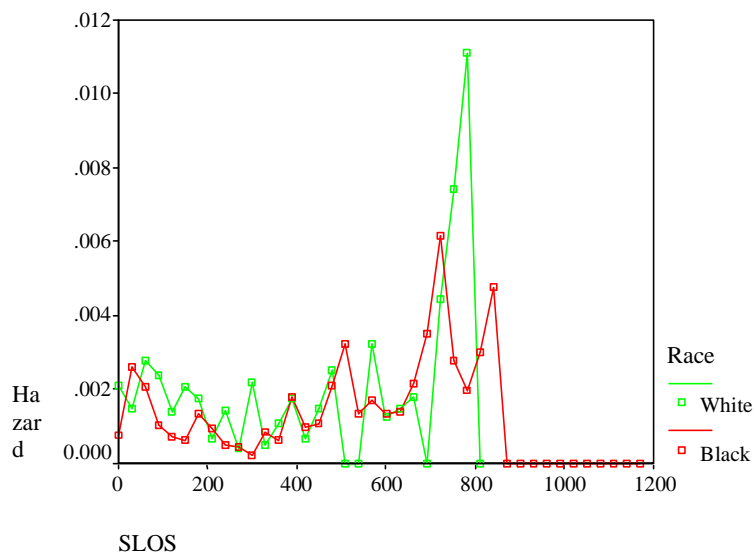


Figure 4. Hazard Function by Race



The median retention in treatment for black clients is 563.97, while for white clients it is 459.26. It would also be good to look at the survival functions by race comparing completing to dropping out. This is done by changing the /STATUS= line to be (1) for time to completing treatment and (2) for time to dropping out.

**Comments:** This paper contains some very simple examples of survival analysis. There are many other uses for survival analysis including modeling, competing risk analysis, non-proportional hazard models. For example, the survival curves for blacks and whites are not only at different levels, they are also shaped differently, suggesting that the effect of race varies by time. In this case, further analysis using a non-proportional hazard model is indicated. To consider competing risks, create separate survival functions for each possible risk. For instance, in Example 2, you might wish to create a second status variable equal to 1 if either discharge or dropout had occurred and compare LOS by the original discharge status. This would give you a separate curve for clients who were discharged and those who dropped out. Luke (1993) and Willett & Singer (1991) both go into much greater detail about using survival analysis in these ways and recommend further reading. In addition, there are several options for comparing groups, inputting and saving data discussed in the SPSS manuals. Finally, it is generally necessary to edit the survival and hazard plots to correct the minimum and maximum scale values for each axis, add a .50 gridline to the survival function, and add interpolation lines to the hazard function.

**Describing These Procedures.** The results in Example 2 would normally be described by including a graph of the survival and or hazard functions. The following text might be used.

*Figure 1 presents the graph of the estimated survival function. The horizontal axis represents time and the vertical axis represents the proportion of clients still in treatment.*

Survival analysis indicated that the overall median days of retention in treatment (where the proportion reaches .5 (50%)) was slightly more than 547 days. The curve indicates relatively long survival times and a fairly even rate of leaving treatment across time periods. The changes in the slope of the curve between 200 and 300 days and again between 300 and 400 days, indicates that the overall rate of leaving treatment changes over time.

Figure 2 presents the overall hazard function for leaving treatment. As can be seen, there is an increased risk of leaving treatment before the first 100 days. Further analysis shows that this is primarily due to clients who dropped out (graph not shown). The hazard rate declines and remains somewhat stable until around 400 days and another large peak at about 2 years (720 days).

Figures 3 and 4 present the survival and hazard function by race. As can be seen, white clients are retained for shorter periods (median LOS=459.3) than are black clients (564.0) (Wilcoxon(Gehan)=7.4; D.F.=1; Prob.= .0064). This is due to some very long-stay black clients who are still in treatment.

### **References:**

- Cox, D.R. (1972) Regression models and life tables. *Journal of the Royal Statistical Society*. 34 (Series B), 187-202.
- Dennis, M. Scott, C.K, Godley, M.D., & Funk, R. (1999). Comparisons of Adolescents and Adults by ASAM profile using GAIN data from the Drug Outcome Monitoring Study (DOMS). Bloomington, IL: Chestnut Health Systems (<http://www.chestnut.org/li/posters>).
- Kalbfleisch, J.D. & Prentice, R.L. (1980). *The statistical analysis of failure time data*. New York: John Wiley and Sons.
- Luke, D.A. (1993) Charting the progress of change: A primer on survival analysis. *American Journal of Community Psychology*. 21(2), 203-246.
- Statistical Program for the Social Sciences (SPSS 1997). *SPSS Advanced Statistics 7.5*. Chicago, IL: Author ([www.spss.com](http://www.spss.com)).
- Statistical Program for the Social Sciences (SPSS 1999). *SPSS Advanced Models 10.0*. Chicago, IL: Author ([www.spss.com](http://www.spss.com)).
- Willett, J.B & Singer, J.D. (1991). From Whether to When: New methods for studying student dropout and teacher attrition. *Review of Educational Research*. 61(4), 407-450.