Introduction

Survival analysis is a general term to describe techniques for analyzing data in which the outcome of interest is the time from a defined beginning point until the occurrence of a specified event.

Examples

- In a cancer treatment trial, the outcome of interest is the survival time of patients from the start of treatment until death.
- In a study of married couples, the outcome of interest is the time from the wedding until the birth of the first child.
- In a study of the carcinogenicity of a chemical, rats are exposed to the chemical and the outcome of interest is the time until a tumor develops.

Censoring

Survival time data are subject to *censoring*. Censoring occurs when the event of interest is not observed for some of the subjects in the study. Censoring occurs if

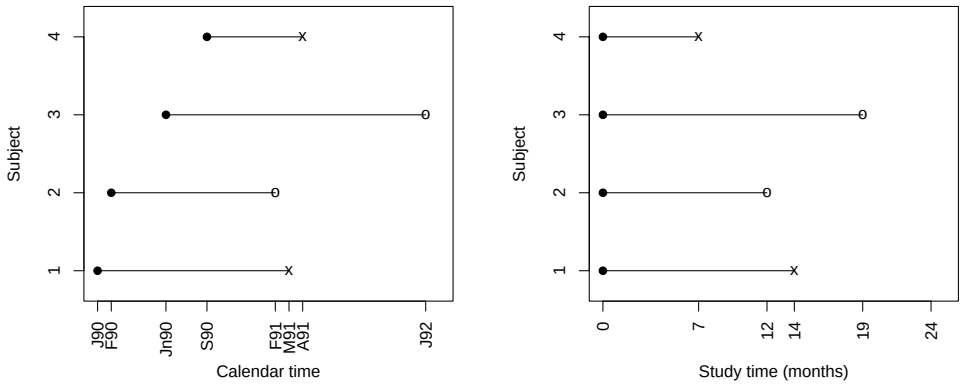
- the subject has not yet had the event when the study is terminated.
- the subject is lost to follow-up or withdrawn from the study.
- the subject dies from causes not relevant to the study.

In general, we assume that censoring is *non-informative*. That is to say, censoring should not convey information about the patient's outcome (event versus non-event).

Survival data

ID	Entry	End	Time (mos)	Event
1	01/01/90	03/01/91	14	Death
2	02/01/90	02/01/91	12	Lost to FU
3	06/01/90	12/31/91	19	Study ended
4	09/01/90	04/01/91	7	Death

Survival data depiction



Data issues

- Distribution of survival times tends to be positively skewed
  - Some observations have much longer survival times than others
  - Non-normal distribution
- Censoring
  - Survival times only partially observed
  - Comparison of mean survival time between groups not appropriate

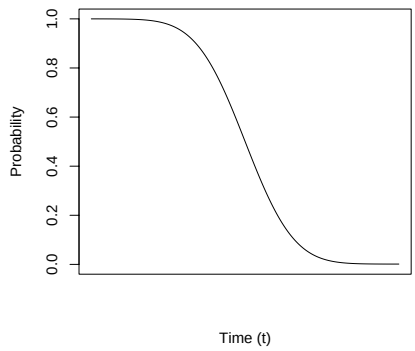
Terminology and notation

- T is the time to the specified event, also commonly referred to as *failure time*. T is a random variable and its observed value for a given subject is denoted as t.
- The *survival function*, $S(t)$, expresses the probability of surviving at least t time units. For example, a "five year survival rate" in cancer is simply the probability of surviving at least five years. The definition of the survival function is

$$S(t) = \text{Prob}(T > t).$$

Properties of $S(t) = \text{Prob}(T > t)$

- Non-increasing function of t
- $S(0) = 1$. In words, at the beginning of observation, no subject has had the event of interest.
- $S(\infty) = 0$. In words, if subjects were observed forever, everyone would eventually experience the event.



Estimation of $S(t)$

The most common estimator of the survival function is the *Kaplan-Meier estimator*, also known as the *Product-limit estimator*. It is a non-parametric estimator of survival, which means that it requires no distributional assumptions about the survival times.

We first introduce the following useful terminology and notation.

Kaplan-Meier estimator of $S(t)$

- Let k index the ordered (from smallest to largest) event times in the data. The event (failure) times are represented as t_k.
- The *risk set* at event time t_k refers to the collection of subjects at risk of failure just before time t_k.
- n_k is the size of the risk set associated with event time t_k.
- d_k is the number of events at event time t_k.

Then the Kaplan-Meier estimator of $S(t)$ is

$$\hat{S}_{KM}(t) = \prod_{\{k:t_k \leq t\}} \left(1 - \frac{d_k}{n_k}\right).$$

KM estimation - example

Consider the following *ordered* event times for seven subjects. The variable CENSOR is equal to 1 if an event is observed and 0 if the event time is censored.

ID	1	2	3	4	5	6	7
t	3	5	7	8	10	11	13
CENSOR	1	1	1	0	1	0	0

There are four event times ($k = 4$) summarized below.

k	t_k	n_k	d_k
1	$t_1 = 3$	7	1
2	$t_2 = 5$	6	1
3	$t_3 = 7$	5	1
4	$t_4 = 10$	3	1

KM estimation - example (cont.)

$$\hat{S}_{KM}(t) = \prod_{\{k:t_k \leq t\}} \left(1 - \frac{d_k}{n_k}\right),$$

Time, t	$\{k : t_k \leq t\}$	$\hat{S}_{KM}(t)$
$[0, 3)$	none	1
$[3, 5)$	$k = 1$	$\left(1 - \frac{1}{7}\right) = 6/7 \doteq 0.86$
$[5, 7)$	$k = 1, 2$	$\left(1 - \frac{1}{7}\right)\left(1 - \frac{1}{6}\right) = 5/7 \doteq 0.71$
$[7, 10)$	$k = 1, 2, 3$	$\left(1 - \frac{1}{7}\right)\left(1 - \frac{1}{6}\right)\left(1 - \frac{1}{5}\right) = 4/7 \doteq 0.57$
$[10, 13)$	$k = 1, 2, 3, 4$	$\left(1 - \frac{1}{7}\right)\left(1 - \frac{1}{6}\right)\left(1 - \frac{1}{5}\right)\left(1 - \frac{1}{3}\right) = 8/21$ $\doteq 0.38$

KM estimation in SAS

```
data one;
    input t cind;
    cards;
3 1
5 1
7 1
8 0
10 1
11 0
13 0
;
run;
proc lifetest data = one method=km;
    time t*cind(0);
run;
```


KM estimation in SAS (cont.)

Product-Limit Survival Estimates

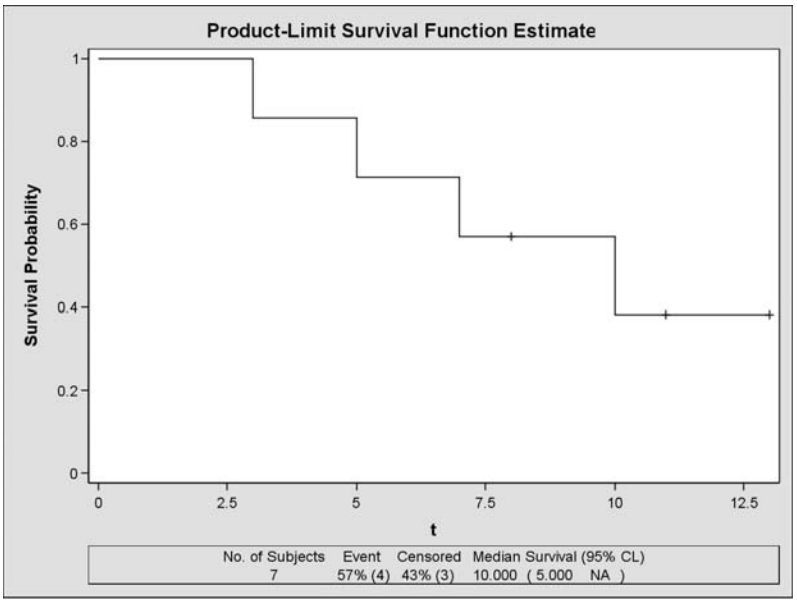
t	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	7
3.0000	0.8571	0.1429	0.1323	1	6
5.0000	0.7143	0.2857	0.1707	2	5
7.0000	0.5714	0.4286	0.1870	3	4
8.0000*	.	.	.	3	3
10.0000	0.3810	0.6190	0.1993	4	2
11.0000*	.	.	.	4	1
13.0000*	.	.	.	4	0

NOTE: The marked survival times are censored observations.

Graphing survivor function in SAS

```
ods html;
ods graphics on;

proc lifetest data = one method=km;
    time t*cind(0);
    survival plots = (survival);
run;

ods graphics off;
ods html close;
```


Graphing survivor function in SAS (cont.)



Note that $\hat{S}_{KM}(t)$ changes at an event time and remains constant between event times.

HIV example

A large HMO wishes to evaluate the survival time of its HIV+ members using a follow-up study. Subjects were enrolled in the study from January 1, 1989 to December 31, 1991. The study ended on December 31, 1995. After a confirmed HIV diagnosis, members were followed until death due to aids or AIDS-related complications, until the end of the study, or until the subject was lost to follow-up. The primary outcome of interest is survival time after a confirmed diagnosis of HIV. 100 subjects were enrolled into the study.

HIV example (cont.)

The data consist of the following variables:

ID Subject ID

TIME Survival time (months)

AGE Age (years) of subject at time of enrollment

DRUG Use of prior injecting drug use (1 = Yes, 0 = No)

CENSOR Censoring indicator (1 = Death observed, 0 = censored)

HIV example in SAS

```
ods rtf file='I:\Survival\hivsurv.rtf';
ods graphics on;

proc lifetest data = one;
    time time*censor(0);
    survival plots=(survival);
run;

ods graphics off;
ods rtf close;
```

(Segue to .rtf output)

SEs of survival estimates

The most common estimator of the SE of KM estimated survival times is the *Greenwood estimator*. It has the following form:

$$\widehat{SE}(\hat{S}(t)) = \hat{S}(t)\sqrt{\sum_{\{k:t_k \leq t\}} \frac{d_k}{n_k(n_k - d_k)}}.$$

The 95% CIs for the estimated survival times based on this formula is

$$\hat{S}(t) \ \pm \ 1.96 \cdot \ \widehat{SE}(\hat{S}(t)).$$

One drawback to using this method of CI construction is that it can lead to lower limits that are less than 0 or upper limits greater than 1. When this occurs, the CI is truncated at the boundary.

SEs and CIs in SAS

The Greenwood estimated standard errors of the KM survival times are produced by default when you run PROC LIFETEST. See the .rtf output file. Since the standard linear-type 95% CI for $S(t)$ (shown on Slide 20) can lead to upper/lower endpoints that are impossible, a number of transformations of the survival function have been proposed so that the resulting interval is contained between 0 and 1, the most common of which is the *log-log* transformation.

The 95% CI for $S(t)$ based on the log-log transformation is

$$\left[\hat{S}(t)\right]^{\exp(1.96\hat{\tau}(t))} \leq S(t) \leq \left[\hat{S}(t)\right]^{\exp(-1.96\hat{\tau}(t))}$$

where

$$\hat{\tau}^2(t) = \frac{\widehat{SE}^2(\hat{S}(t))}{\left[\hat{S}(t)\log(\hat{S}(t))\right]^2}$$

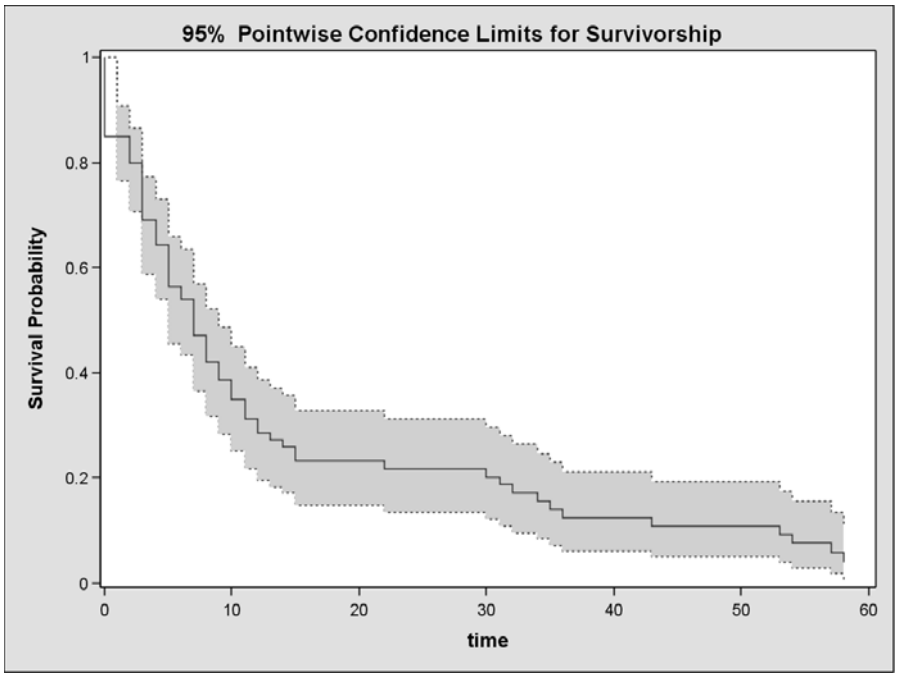
is the estimated variance of $\log(-\log(\hat{S}(t)))$.

SEs and CIs in SAS (cont.)

The following code produces a graph of the survival function and the corresponding log-log 95% CIs (the default), as well as an output data set containing the endpoints of the intervals.

```
ods html;
ods graphics on;

proc lifetest data = one;
    time time*censor(0);
    survival out = survcis plots=(survival, cl);
run;

ods graphics off;
ods html close;
```

SEs and CIs in SAS (cont.)



SEs and CIs in SAS (cont.)

```
proc print data = survcis;
run;
```

Obs	time	_CENSOR_	SURVIVAL	CONFTYPE	SDF_LCL	SDF_UCL
1	0	.	1.00000		1.00000	1.00000
2	1	0	0.85000	LOGLOG	0.76359	0.90672
3	1	1	0.85000		.	.
4	1	1	0.85000		.	.
5	2	0	0.79880	LOGLOG	0.70566	0.86523
6	2	1	0.79880		.	.
7	2	1	0.79880		.	.
8	2	1	0.79880		.	.
9	2	1	0.79880		.	.
10	2	1	0.79880		.	.
11	3	0	0.68937	LOGLOG	0.58622	0.77176

Reported percentiles

In addition to the KM estimated survival times and their SEs, SAS also reports estimates of the most common percentiles of the survival times, namely the 25th, 50th (median) and 75th percentiles. For the HIV data, we have the following output.

```
             Quartile Estimates
             Point     95% Confidence Interval
Percent    Estimate      [Lower      Upper)
     75     15.0000     10.0000     34.0000
     50      7.0000      5.0000      9.0000
     25      3.0000      2.0000      4.0000

    Mean    Standard Error
 14.5912            1.9598

NOTE: The mean survival time and its standard error were
      underestimated because the largest observation was censored
      and the estimation was restricted to the largest event time.
```

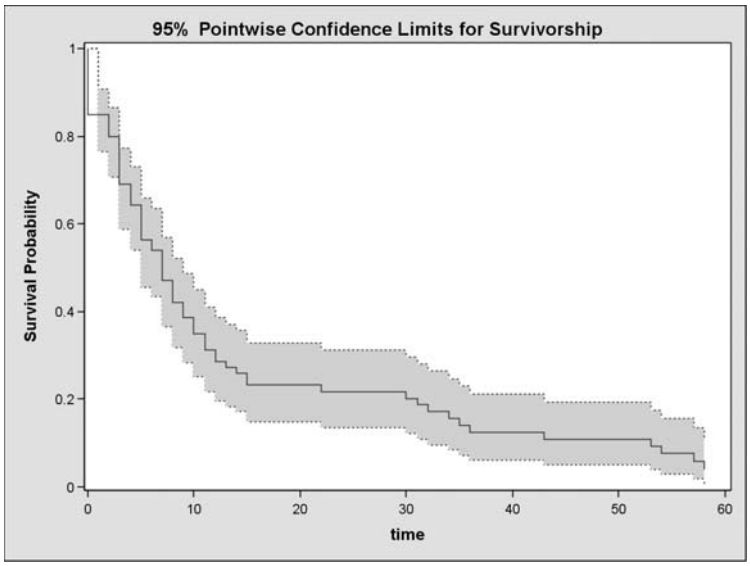

Interpreting percentiles

You need to be a little careful interpreting percentiles of the survival times. For example, let t_{25} be the 25th percentile of the survival times. This means that 25% of the observed survival times are equal to or smaller than t_{25}. In other words, 25% of the population has failed by time t_{25}. This means that 75% of the population survived beyond t_{25}. In general, if t_p is the pth percentile of the survival times, then

$$S(t_p) = \text{Prob}(T > t_p) = 1 - (p/100).$$

Compare the reported percentiles with the KM estimate of survival.

Percent	Quartile Estimates Point Estimate	95% Confidence Interval [Lower	Upper)
75	15.0000	10.0000	34.0000
50	7.0000	5.0000	9.0000
25	3.0000	2.0000	4.0000

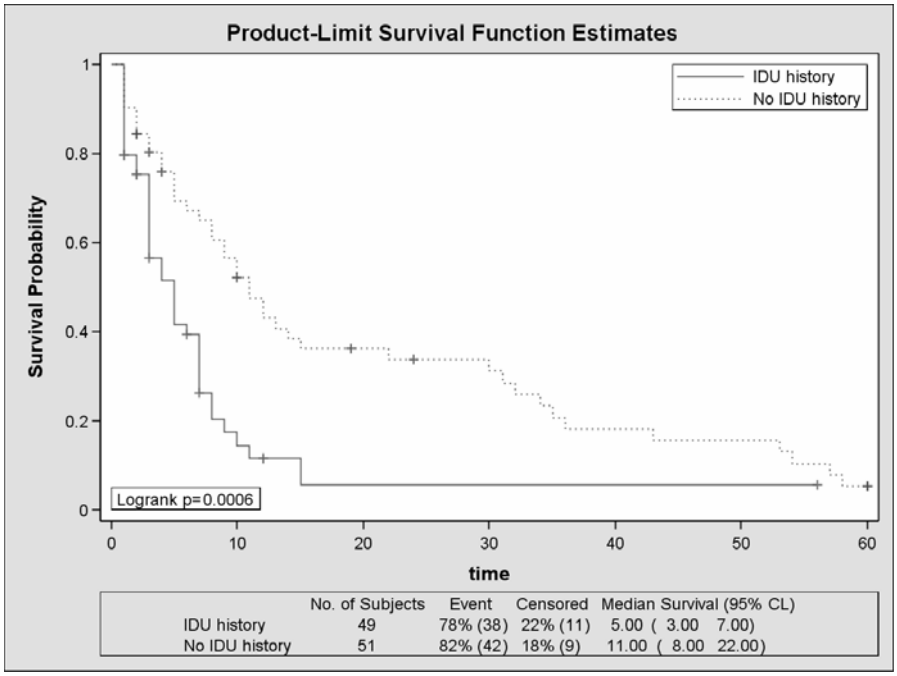


Comparing survival functions

Suppose we want to compare survival experience between subjects with and without a history of injecting drug use. This is accomplished easily in PROC LIFETEST using the STRATA statement.

```
ods html;
ods graphics on;

proc lifetest data = one;
    time time*censor(0);
    strata drug;
    survival plots=(survival);
run;

ods html close;
ods graphics off;
```

Comparing survival functions (cont.)



The Log-Rank test

Are the observed differences in survival significant? One way to answer this question is using the *Log-Rank test*.

Let $S_1(t)$ be the survival function for Group 1.
Let $S_2(t)$ be the survival function for Group 2.

The log-rank test tests the following null versus alternative hypothesis:

$$H_0 : S_1(t) = S_2(t) \text{ for all } t$$
$$H_A : S_1(t) \neq S_2(t) \text{ for at least one } t$$

The Log-Rank test (cont.)

The idea behind the log-rank test is to construct a 2×2 contingency table of group membership versus survival for each event time, t. The data from the sequence of tables are accumulated using the Mantel-Haenszel test statistic.

The Log-Rank test (cont.)

Let t_j, $j = 1, \ldots, J$, be the ordered failure times in the pooled sample. (Here, the total number of observed failure times is J.) At each time t_j construct the following table:

	Group		
Status	Group 1	Group 2	
Failed at t_j	d_{1j}	d_{2j}	D_j
Survive past t_j	s_{1j}	s_{2j}	S_j
	n_{1j}	n_{2j}	N_j

The Log-Rank test (cont.)

- The observed number of failures in Group 1 is $O_j = d_{1j}$.
- The expected number of failures in Group 1 (under the null hypothesis) is $E_j = n_{1j}D_j/N_j$.
- The variance, v_j, of d_{1j} is $(n_{1j}n_{2j}D_jS_j)/(N_j^2(N_j - 1)$.

Then the log-rank test is

$$Q = \frac{\left[\sum_{j=1}^{J}(O_j - E_j)\right]^2}{\sum_{j=1}^{J} v_j}.$$

Under the null hypothesis, $Q \sim \chi^2_1$. (Note: The degrees of freedom of the test are "number of groups - 1".)

The Log-Rank test in SAS

When you use the STRATA statement in PROC LIFETEST (see Slide 28), the log-rank test is performed on the groups identified by the variable in the STRATA statement.

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	11.8556	1	0.0006
Wilcoxon	10.9104	1	0.0010
-2Log(LR)	20.9264	1	<.0001

Conclusions

The log-rank test is highly significant ($p = 0.0006$). Therefore, we reject the null hypothesis and conclude that the distributions of survival times for HIV+ patients with and without a history of injecting drug use are significantly different.

Some words of caution

The log-rank test is the most powerful test for the specific alternative

$$H_A : S_1(t) = [S_2(t)]^c, \ c \neq 1.$$

It is not very powerful for other alternatives for which $S_1(t)$ is different from $S_2(t)$. This means that failing to detect a significant difference between the survival functions for two groups can be attributed to any of the following:

1. H_0 is true
2. Lack of power because of inadequate sample size
3. Lack of power due to departure from the assumption of the alternative for which the log-rank test is most powerful.

Checking for proportional hazards

$S_1(t) = [S_2(t)]^c,\ c \neq 1$ is known as the *proportional hazards assumption* (more on this later). To assess the validity of this assumption, we use the following fact.

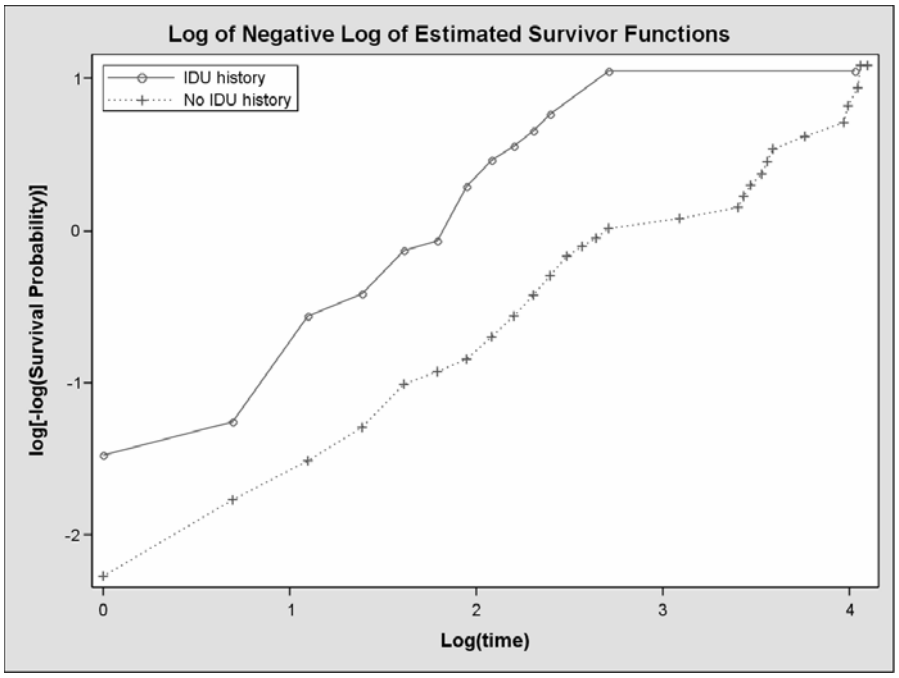
$$\begin{aligned} \log S_1(t) &= c \log S_2(t) \\ \iff -\log S_1(t) &= c(-\log S_2(t)) \\ \iff \log(-\log S_1(t)) &= \log c + \log(-\log S_2(t)) \end{aligned}$$

Checking for proportional hazards (cont.)

So, if we plot $\log[-\log[S_1(t)]]$ on the same graph with $\log[-\log[S_2(t)]]$ we should see two curves that are separated by a constant distance, $\log c$. We construct this plot directly in SAS.

```
ods html;
ods graphics on;

proc lifetest data = one;
    time time*censor(0);
    strata drug;
    survival plots=(survival, lls);
run;

ods html close;
ods graphics off;
```

Checking for proportional hazards (cont.)



If proportional hazards function not met

If the assumption of proportional hazards is not met, there are alternative tests that weight the contributions to the numerator and denominator of Q. We include the references for completeness.

1. Gehan E.A. A generalized Wilcoxon test for comparing arbitrarily signly censored samples. *Biometrika*, **52**, 203-223, 1965.
2. Tarone R.E. and Ware J. On distribution free tests for equality of survival distributions. *Biometrika*, **64**, 156-160, 1977.
3. Prentice R.L. Linear rank tests with right-censored data. *Biometrika*, **65**, 167-179, 1978.

Testing more than two groups

The log-rank test can be generalized to testing equality of the survivor functions for more than two groups, where the alternative hypothesis is that at least two of the survival functions are different. The form of the test statistic is similar, and its distribution under the null hypothesis of equality of the survivor functions is chi-square with d degrees of freedom, where $d =$ number of groups - 1.

Testing more than two groups (cont.)

Consider the HIV data, and suppose we are interested in testing the hypothesis

$$H_0: \quad S_1(t) = S_2(t) = S_3(t) = S_4(t)$$

$$H_A: \quad \text{At least two of the survivor functions are different}$$

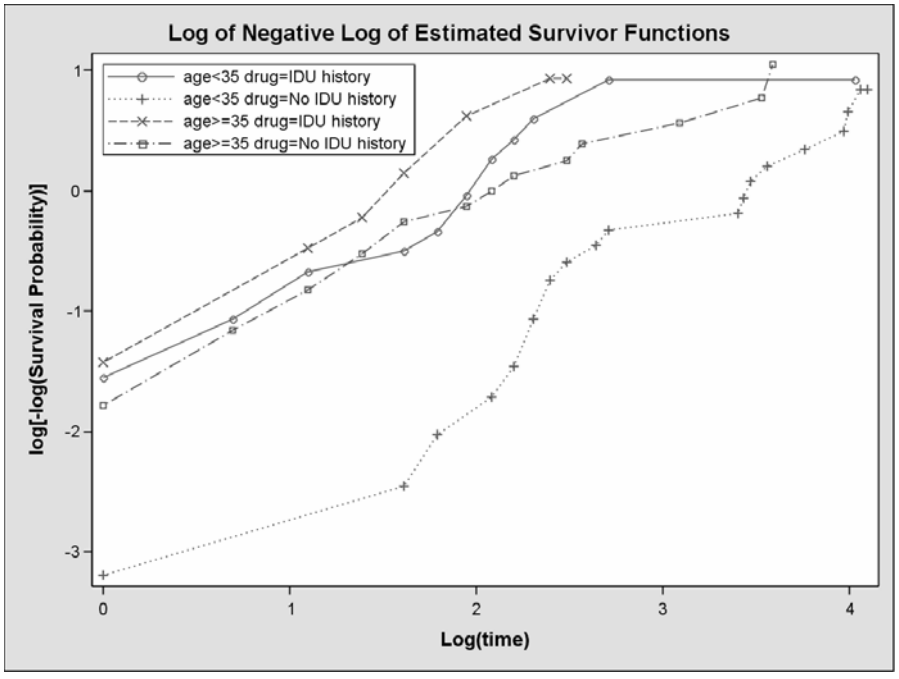
where

- Group 1 = AGE < 35 with no history of IDU
- Group 2 = AGE < 35 with history of IDU
- Group 3 = AGE ≥ 35 with no history of IDU
- Group 4 = AGE ≥ 35 with history of IDU

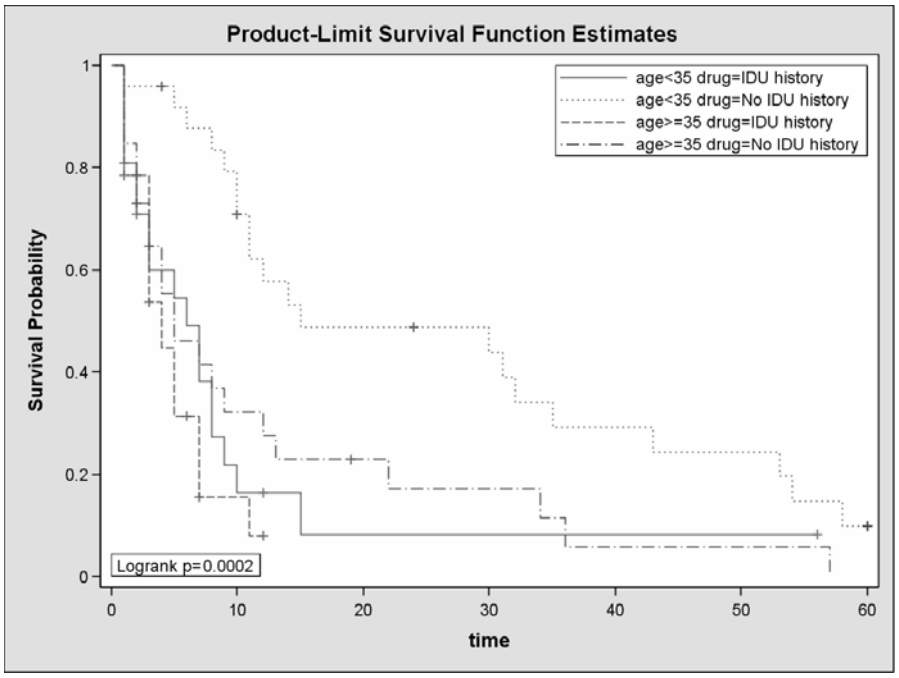
Testing more than two groups in SAS

```
ods html;
ods graphics on;

proc lifetest data = one notable;
    time time*censor(0);
    strata age(35) drug;
    survival plots=(survival, lls);
run;

ods html close;
ods graphics off;
```


Testing more than two groups in SAS (cont.)



Testing more than two groups in SAS (cont.)



Testing more than two groups in SAS (cont.)

Stratum	drug	age	Total	Failed	Censored	Percent Censored
1	IDU history	<35	21	17	4	19.05
2	No IDU history	<35	25	20	5	20.00
3	IDU history	>=35	28	21	7	25.00
4	No IDU history	>=35	26	22	4	15.38
Total			100	80	20	20.00

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	20.2473	3	0.0002
Wilcoxon	19.9514	3	0.0002
-2Log(LR)	33.2148	3	<.0001