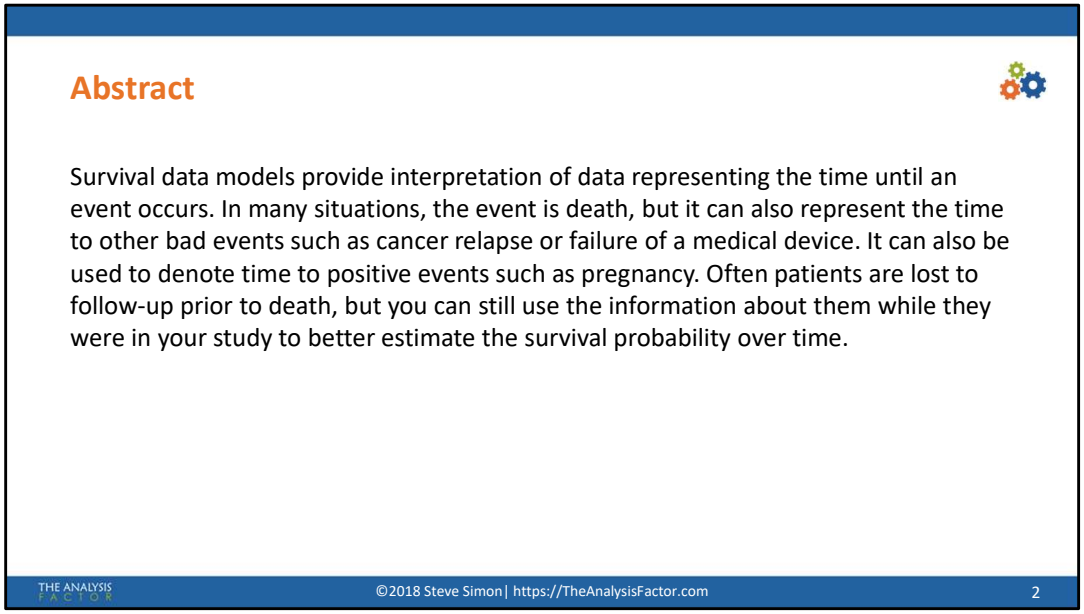
Survival Analysis: Models for Time to Event Data

Module 1: An Introduction to Kaplan-Meier Curves

Steve Simon

THE ANALYSIS FACTOR



In these diagrams, square edged boxes represent observed variables, and rounded or oval boxes represent latent variables, sometimes called factors:

Abstract (continued)

This is done using the Kaplan-Meier curve, an approach developed by Edward Kaplan and Paul Meier in 1958. In this talk, you will see a simple example using fruit fly data and learn how to interpret the Kaplan-Meier curve to estimate survival probabilities and survival percentiles.

Most of this talk is based on a web page I wrote in 2008:
http://www.pmean.com/08/SimpleKm.html

In these diagrams, square edged boxes represent observed variables, and rounded or oval boxes represent latent variables, sometimes called factors:

Fruit Fly Data (Round 1)

Where does this data come from?

The following data represents survival time for a group of fruit flies and is a subset of a larger data set found at the Data and Story Library (DASL). The data set has been slightly modified to simplify some of these explanations.

There are 25 flies in the sample, with the first fly dying on day 37 and the last fly dying on day 96. If you wanted to estimate the survival probability for this data, you would draw a curve that decreases by 4% (1/25) every time a fly dies.

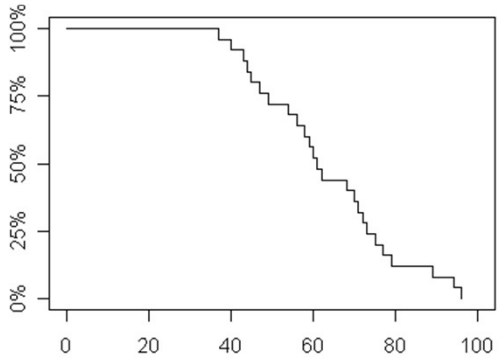
Fruit Fly Data (Round 1)

At each date, the survival probability drops by 1/25.

37 96%	58 60%	73 24%
40 92%	59 56%	75 20%
43 88%	60 52%	77 16%
44 84%	61 48%	79 12%
45 80%	62 44%	89 8%
47 76%	68 40%	94 4%
49 72%	70 36%	96 0%.
54 68%	71 32%	.
56 64%	72 28%	

Fruit Fly Data (Round 1)

A graphical depiction of the survival probability



Fruit Fly Data (Round 2)

Let's alter the experiment

Now let's alter the experiment. Suppose that totally by accident, a technician leaves the screen cover open on day 70 and all the flies escape. You're probably worried that the whole experiment has been ruined. But don't be so pessimistic. You still have complete information on survival of the fruit flies up to their 70th day of life. Here's how you would present the data and estimate the survival probabilities.

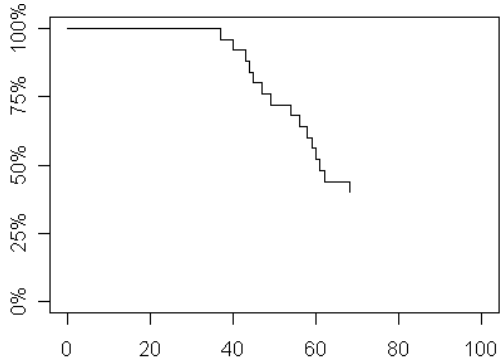
Fruit Fly Data (Round 2)

You can still estimate some survival probabilities

37 96%	58 60%	70+ ?
40 92%	59 56%	70+ ?
43 88%	60 52%	70+ ?
44 84%	61 48%	70+ ?
45 80%	62 44%	70+ ?
47 76%	68 40%	70+ ?
49 72%	70+ ?	70+ ?
54 68%	70+ ?	.
56 64%	70+ ?	

Fruit Fly Data (Round 2)

Here is a graph of the survival probabilities



Fruit Fly Data (Round 2)

What you can still estimate

We clearly have enough data to make several important statements about survival probability. For example, the median survival time is 61 days because roughly half of the flies had died before this day.

By the way, you might be tempted to ignore the ten flies who escaped. But that would seriously bias your results. The median survival time, for example, of the 15 flies who did not escape, for example, is only 54 days which is much smaller than the actual median.

Fruit Fly Data (Round 3)

Another change to the data

Let's look at a third experiment, where the screen cover is left open and all but four of the remaining flies escape. It turns out that those four remaining flies who didn't bug out will allow us to still get reasonable estimates of survival probabilities beyond 70 days. Here is the data and the survival probabilities.

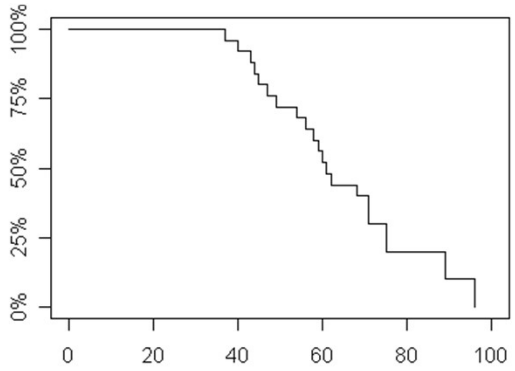
Fruit Fly Data (Round 3)

Here are the estimated survival probabilities

37 96%	58 60%	70+ ?
40 92%	59 56%	75 20%
43 88%	60 52%	70+ ?
44 84%	61 48%	70+ ?
45 80%	62 44%	89 10%
47 76%	68 40%	70+ ?
49 72%	70+ ?	96 0%
54 68%	71 30%	
56 64%	70+ ?	

Fruit Fly Data (Round 3)

Here is a graph of the survival probabilities



Fruit Fly Data (Round 3)

What you do with the six escaped flies is to allocate their survival probabilities equally among the four flies who didn't bug out. This places a great responsibility among each of those four remaining flies since each one is now responsible for 10% of the remaining survival probability, their original 4% plus 6% more which represents a fourth of the 24% survival probability that was lost with the six escaping flies.

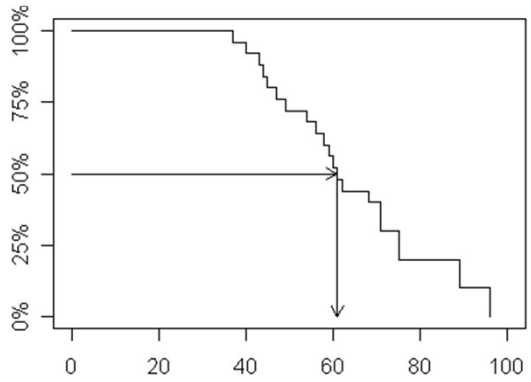
Fruit Fly Data (Round 3)

Informative censoring

If the censoring mechanism were somehow related to survival prognosis, then you would have the possibility of serious bias in your estimates. Suppose for example, that only the toughest of flies (those with the most days left in their short lives) would have been able to escape. The flies destined to kick the bucket on days 70, 71, 72, and 73, were already on their deathbeds and unable to fly at all, much less make a difficult escape. Then these censored values would not be randomly interspersed among the remaining survival times, but would constitute some of the larger values. But since these larger values would remain unobserved, you would underestimate survival probabilities beyond the 70th day.

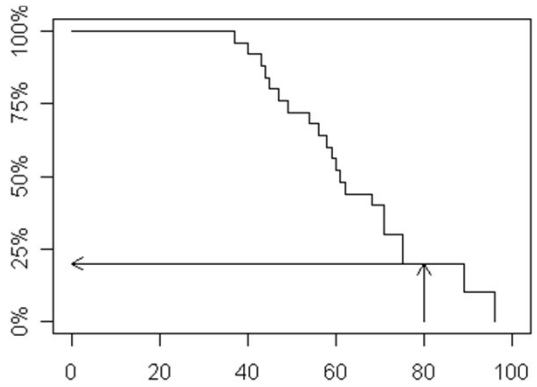
Fruit Fly Data (Round 3)

Interpretation: 50th percentile = 61



Fruit Fly Data (Round 3)

Interpretation: 80 week survival probability = 20%



Hand Calculation of Kaplan-Meier curve

Table 2.1 of Hosmer, Lemeshow, and May

Data: 6, 44, 21+, 14, 62

Time	Censor
6	1
14	1
21	0
44	1
62	1

Hand Calculation of Kaplan-Meier curve

Calculate number at risk (ni) and deaths (di) at time=i.

Data: 6, 44, 21+, 14, 62

Time	Censor	ni	di
6	1	5	1
14	1	4	1
21	0	3	0
44	1	2	1
62	1	1	1

Hand Calculation of Kaplan-Meier curve

Calculate the conditional probability of survival.

Data: 6, 44, 21+, 14, 62

Time	Censor	ni	di	(ni-di)/ni
6	1	5	1	4/5
14	1	4	1	3/4
21	0	3	0	3/3
44	1	2	1	1/2
62	1	1	1	0/1

Hand Calculation of Kaplan-Meier curve

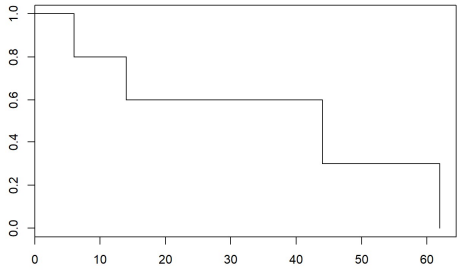
Compute the cumulative product.

Data: 6, 44, 21+, 14, 62

Time	Censor	ni	di	(ni-di)/ni	Cumulative Product
6	1	5	1	4/5	4/5 = 0.8
14	1	4	1	3/4	4/5*3/4= 0.6
21	0	3	0	3/3	4/5*3/4*3/3=0.6
44	1	2	1	1/2	4/5*3/4*3/3*1/2=0.3
62	1	1	1	0/1	4/5*3/4*3/3*1/2*0/1=0.0

Hand Calculation of Kaplan-Meier curve

Here's a Kaplan-Meier graph, similar to Figure 2.2



Kaplan-Meier Curve for a Larger Data Set

WHAS100 data, first six rows

id	admitdate	foldate	los	lenfol	fstat	age	gender	bmi
1	03/13/1995	03/19/1995	4	6	1	65	0	31.38134
2	01/14/1995	01/23/1996	5	374	1	88	1	22.65790
3	02/17/1995	10/04/2001	5	2421	1	77	0	27.87892
4	04/07/1995	07/14/1995	9	98	1	81	1	21.47878
5	02/09/1995	05/29/1998	4	1205	1	78	0	30.70601
6	01/16/1995	09/11/2000	7	2065	1	82	1	26.45294

Kaplan-Meier Curve for a Larger Data Set

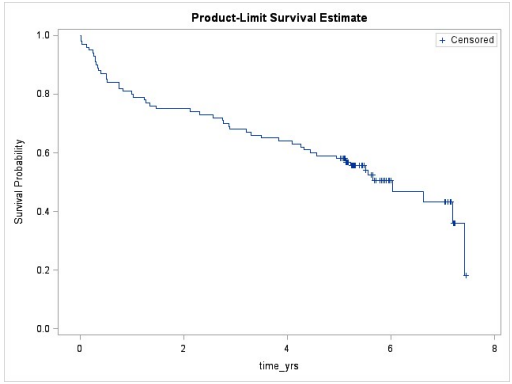
Data dictionary for WHAS100

This is a tab delimited file with 100 rows and 9 columns of data.

id, a sequential code from 1 to 100
admitdate, Admission Date, formatted as mm/dd/yyyy
foldate, Follow Up Date, formatted as mm/dd/yyyy
los, Length of Hospital Stay in Days
lenfol, Follow Up Time in Days
fstat, Vital Satus, 1 = Dead, 0 = Alive
age, Age at Admission in years
gender, 0 = Male, 1 = Female
bmi, Body Mass Index, kg/m^2

Kaplan-Meier Curve for a Larger Data Set

The overall Kaplan-Meier curve (SAS)



Kaplan-Meier Curve for a Larger Data Set

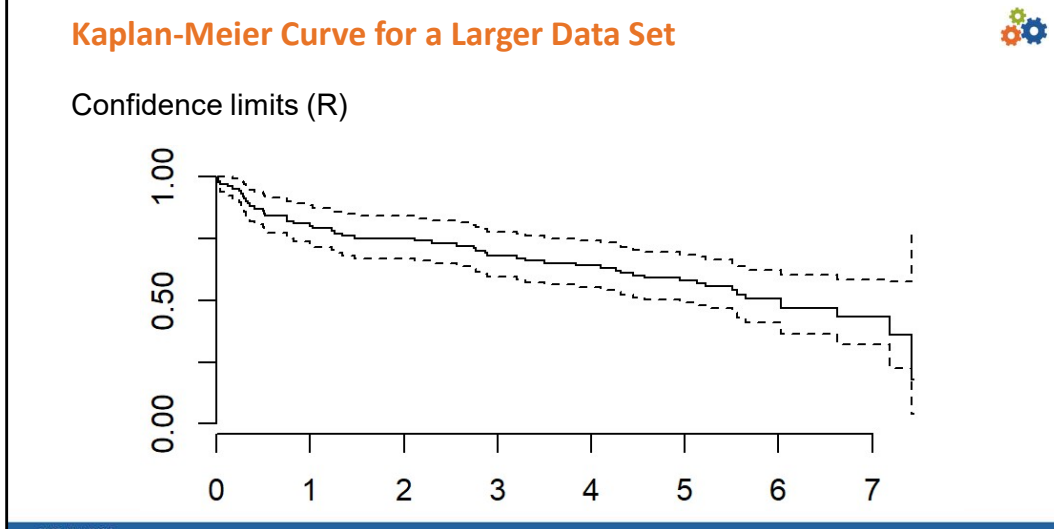
Formula for confidence limits

Since the Kaplan-Meier curve is a product of conditional probabilities, you can, with relative ease, compute the variance on a log scale and then transform back to the original scale.

$$Var(S(t_i)) = S(t_i)^2 \sum_{j \le i} \frac{d_j}{n_j(n_j - d_j)}$$

The full derivation requires knowledge of change of variable methods that you might have learned in your mathematical statistics class. Details are on pages 28-29 of Hosmer, Lemeshow, and May. There are other formulas for calculating confidence limits, but the limits based on the variance shown above works well in practice.

Kaplan-Meier Curve for a Larger Data Set



Kaplan-Meier Curve for a Larger Data Set

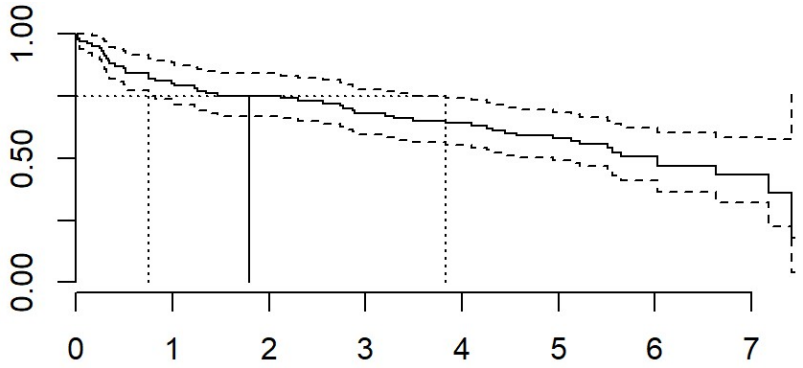
Quartile confidence limits

You can get confidence limits for the median survival time, the quartiles or any other survival percentile by extrapolating horizontally along the confidence limits of the Kaplan-Meier curve.

For some percentiles, the horizontal line may not ever cross the upper confidence limit. In that case, you can set the upper confidence limit to plus infinity.

Kaplan-Meier Curve for a Larger Data Set

How you can visualize quartile confidence limits.



Kaplan-Meier Curve for a Larger Data Set

Quartile confidence limits (SAS)

SAS produces quartile confidence limits and estimated mean by default. The mean estimate is biased if the last observation is censored.

Summary Statistics for Time Variable time_yrs

Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	7.41958	LOGLOG	7.18412	.
50	6.02601	LOGLOG	4.44627	7.41958
25	1.79603	LOGLOG	0.75017	3.29911

Mean	Standard Error
4.78754	0.29242

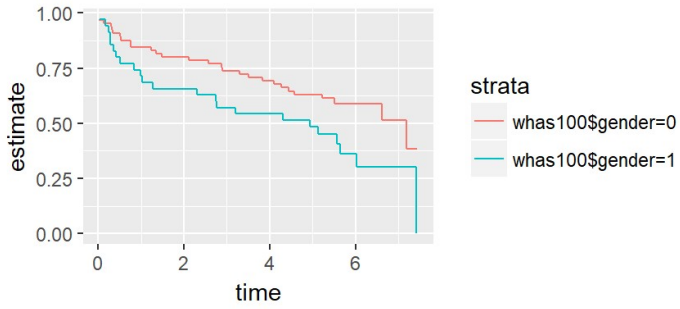
Kaplan-Meier Curve for a Larger Data Set

Comparing two or more Kaplan-Meier curves

If you want to compare the survival curves for two subgroups, you should first draw the two subgroup Kaplan-Meier curves on the same graph.

Kaplan-Meier Curve for a Larger Data Set

Comparing two or more Kaplan-Meier curves



The Log-Rank Test

Formulation

The formulation of the log-rank test, as described in Hosmer, Lemeshow, and May is a bit opaque.

d_{ij} is the number of events in the i^{th} group at time t_j (i=1,2)
n_{ij} is the number of patients at risk in the i^{th} group at time t_j
$d_j = d_{1j} + d_{2j}$; $n_j = n_{1j} + n_{2j}$

The Log-Rank Test

Formulation found in Hosmer, Lemeshow, and May

$$e_{1i} = \frac{n_{1i} d_i}{n_i}$$

$$v_{1i} = \frac{n_{1i} n_{0i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$$

$$Q = \frac{(\sum_i (d_{1i} - e_i))^2}{\sum_i V_i}$$

The Log-Rank Test

A simpler formulation

This looks a bit mystifying, but if you define

$$p_i = \frac{d_i}{n_i}$$

then e_{1i} and v_{1i}

$$e_{1i} = n_{1i} p_i$$

$$v_{1i} = n_{1i} p_i (1 - p_i) \frac{n_i - n_{1i}}{n_i - 1}$$

are just the mean of a binomial distribution and the variance of a binomial distribution with a finite population correction factor. Equivalently, the latter is the variance of a hypergeometric distribution.

The Log-Rank Test

Hand calculation on a small data set

Calculate the number of deaths and the number at risk at each time point.

Male 6, 44+, 98, 114
Female 14, 44, 89+, 98, 104

Time	d_{0i}	n_{0i}	d_{1i}	n_{1i}	n_i	d_i
6	1	4	0	5	1	9
14	0	3	1	5	1	8
44	0	3	1	4	1	7
98	1	2	1	2	2	4
104	0	1	1	1	1	2
114	1	1	0	0	1	1

The Log-Rank Test

Hand calculation on a small data set

Compute the expected value and variance at each time point.

Male 6, 44+, 98, 114
Female 14, 44, 89+, 98, 104

Time	d_{0i}	n_{0i}	d_{1i}	n_{1i}	n_i	d_i	$e_{1i} = \frac{n_{1i}d_i}{n_i}$	$v_{1i} = \frac{n_{1i}n_{0i}d_i(n_i-d_i)}{n_i^2(n_i-1)}$
6	1	4	0	5	1	9	5*1/9=0.556	4*5*1*(9-1)/(9^2*8)=0.2469
14	0	3	1	5	1	8	5*1/8=0.625	3*5*1*(8-1)/(8^2*7)=0.2344
44	0	3	1	4	1	7	4*1/7=0.571	3*4*1*(7-1)/(7^2*6)=0.2449
98	1	2	1	2	2	4	2*2/4=1.000	2*2*2*(4-2)/(4^2*3)=0.3333
104	0	1	1	1	1	2	1*1/2=0.500	1*1*1*(2-1)/(2^2*1)=0.2500
114	1	1	0	0	1	1	0*1/1=0.000	1*0*1*(1-0)/(1^2*0)=0.0000

The Log-Rank Test

Log rank test (SAS)

Ignore the covariance statistics that SAS produces. They are of limited relevance for more complex settings, but are totally useless for a two group test.

Rank Statistics		
gender	Log-Rank	Wilcoxon
0	-6.6200	-459.00
1	6.6200	459.00

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	3.9714	1	0.0463
Wilcoxon	3.4624	1	0.0628
-2Log(LR)	4.4183	1	0.0356

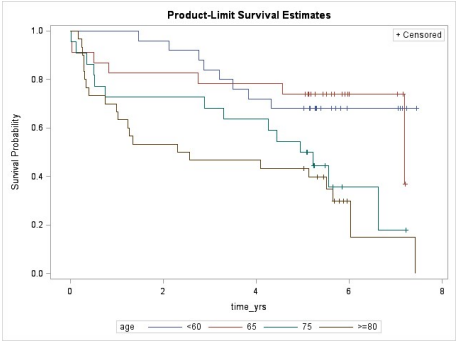
The Log-Rank Test

How to handle continuous outcomes

The log rank test cannot easily handle continuous predictor variables. For these variables, you should really consider a more sophisticated model like a Cox proportional hazards model (coming up in the next lecture). But you can get a rough preliminary idea of what is going on with a continuous predictor by categorizing it using one or more cut-points. Here's an example using age.

The Log-Rank Test

How to handle continuous outcomes (SAS)



The Log-Rank Test

How to handle continuous outcomes (SAS)

Rank Statistics		
age	Log-Rank	Wilcoxon
<60	-7.5195	-490.00
65	-5.9178	-385.00
75	3.7738	201.00
>=80	9.6635	674.00

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	15.5721	3	0.0014
Wilcoxon	12.2981	3	0.0064
-2Log(LR)	17.2401	3	0.0006

The Log-Rank Test

Test for trend

The log rank test for more than two groups treats the groups in a nominal fashion—order is not important. For this particular data set, and many others, you might prefer a test for trend. This is available in most statistical packages, but we will not show the details here.

The Log-Rank Test

Limitations

The log rank test:

- works well when you're comparing a treatment group to a control group
- you can also use it when you have three or more groups

But the log rank test does not extend beyond this:

- you cannot include a continuous predictor,
- you cannot analyze data with multiple predictors, and
- you cannot do risk adjustment