




Marcello Pagano

[JOTTER 12 SURVIVAL ANALYSIS]

Survival curves, product limit method, censoring, log rank test

Survival Analysis



Regression where the end-point,
or dependent variable, is positive.

e.g. survival

time to an event

e.g. response

failure

pregnancy

infection

strike end

latency period

We continue our study of regression. This week we look at what is generically called survival analysis—essentially the study of positive outcome variables—and how to incorporate explanatory variables into the argument. This is an important topic that can be applied in a number of situations, not the least is the study of clinical trials, but we only have enough time to introduce you to the topic.

Quick review: In our study of regression, we first looked at simple linear regression. That meant that the means of the outcome variable, for fixed values of the explanatory variable, all lay on a straight line.

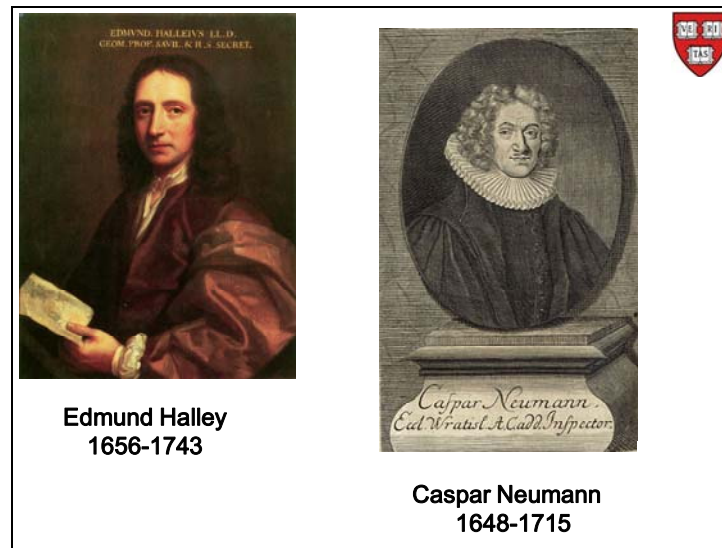
Sometimes, if the plotted line of the means is not straight, we can *transform*, or change the units of the measurements to obtain a straight line. We then extended these methods to multiple-regression by incorporating more than one explanatory variable.

When faced with a dichotomous outcome variable, its mean is a probability, and thus is constrained to take values between zero and one. As a result, a straight line cannot provide a good model for dichotomous regression, and so we turned to logistic regression (other transformations are available, of course).

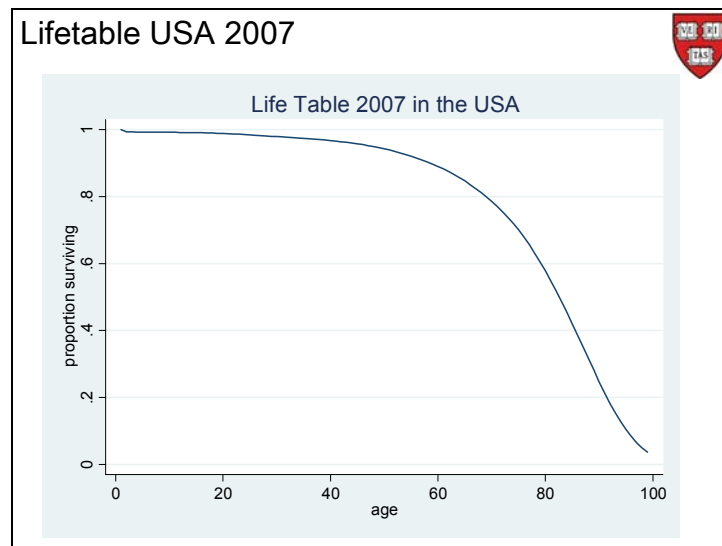
Now we look at the situation when the outcome variable is non-negative. For example, we might be interested in the time to an event, such as the time it takes to respond to cancer treatment or how long it takes for the tumor to shrink; or the time to failure, such as death; or in a fertility study, how long does it take for a woman to get pregnant; or when studying HIV transmission, how long before a partner gets infected with the virus.

All of these are non-negative variables, so once again we probably would not want to use linear regression in general, since a non-horizontal straight line is both positive and negative. Plus, more importantly, with survival data, for example, linear methods are not optimal.

When these outcome variables are positive, the first idea might be to simply transform them by taking their logarithms, and thus removing the positivity constraint. This approach is fine, although it needs to be modified some to include the possibility of an outcome of zero, but there are ways of handling that circumstance. But such an approach does not handle incomplete or censored data, too well. We explain the meaning of this below, after we introduce the new method we recommend. First, though we revisit how we handled survival data at the population level, a few weeks ago, because it is quite closely related to the proposed, new method.



Let us return to Edmund Halley and recall how his method for creating the life table out of Neumann's data.



Here is the graph we get from applying Halley's method to the data for 2007 in the USA. Along the horizontal we have age and on the vertical we have proportion surviving to a particular age. This way we see the survival experience for this constructed cohort of individuals.

Life Table for 2007 USA		Probability of dying between ages x to $x + 1$
Age		q_x
0-1		0.006761
1-2		0.000460
2-3		0.000286
3-4		0.000218
4-5		0.000176
5-6		0.000164
6-7		0.000151
7-8		0.000140
8-9		0.000124
9-10		0.000105
10-11		0.000091
11-12		0.000094
12-13		0.000132
13-14		0.000209
14-15		0.000314
15-16		0.000426
16-17		0.000529
17-18		0.000627
18-19		0.000715
19-20		0.000796

To remind ourselves, in order to construct this survival curve, or life table, we start with the life span broken into age intervals, and then obtain the conditional probabilities that someone entering an interval will die within that interval—the hazard function. With this set of probabilities we can trace the mortality, or life experience of a cohort subjected to these hazards, as it progresses through time. And that is the life table—it is a way to convert the hazard function into an associated survival function.

Life Table for 2007 USA		Probability of dying between ages x to $x + 1$	Number surviving to age x	Number dying between ages x to $x + 1$
Age		q_x	l_x	d_x
0-1		0.006761	100,000	676
1-2		0.000460		
2-3		0.000286		
3-4		0.000218		
4-5		0.000176		
5-6		0.000164		
6-7		0.000151		
7-8		0.000140		
8-9		0.000124		
9-10		0.000105		
10-11		0.000091		
11-12		0.000094		
12-13		0.000132		
13-14		0.000209		
14-15		0.000314		
15-16		0.000426		
16-17		0.000529		
17-18		0.000627		
18-19		0.000715		
19-20		0.000796		

In particular, to start the curve off, suppose 100,000 babies are born (enter the first interval). Of these, a proportion of 0.006761 die. That means 676 die.

Life Table for 2007 USA			
Age	Probability of dying between ages x to $x + 1$	Number surviving to age x	Number dying between ages x to $x + 1$
	q_x	l_x	d_x
0-1	0.006761	100,000	676
1-2	0.000460	99,324	46
2-3	0.000286		
3-4	0.000218		
4-5	0.000176		
5-6	0.000164		
6-7	0.000151		
7-8	0.000140		
8-9	0.000124		
9-10	0.000105		
10-11	0.000091		
11-12	0.000094		
12-13	0.000132		
13-14	0.000209		
14-15	0.000314		
15-16	0.000426		
16-17	0.000529		
17-18	0.000627		
18-19	0.000715		
19-20	0.000796		

That leaves 99,324 to reach their first birthday. Now, of those a proportion of 0.00460 will die. That means 46 ($=99,324 \times 0.00460$) of these will die before reaching their second birthday.

Life Table for 2007 USA			
Age	Probability of dying between ages x to $x + 1$	Number surviving to age x	Number dying between ages x to $x + 1$
	q_x	l_x	d_x
0-1	0.006761	100,000	676
1-2	0.000460	99,324	46
2-3	0.000286	99,278	28
3-4	0.000218		
4-5	0.000176		
5-6	0.000164		
6-7	0.000151		
7-8	0.000140		
8-9	0.000124		
9-10	0.000105		
10-11	0.000091		
11-12	0.000094		
12-13	0.000132		
13-14	0.000209		
14-15	0.000314		
15-16	0.000426		
16-17	0.000529		
17-18	0.000627		
18-19	0.000715		
19-20	0.000796		

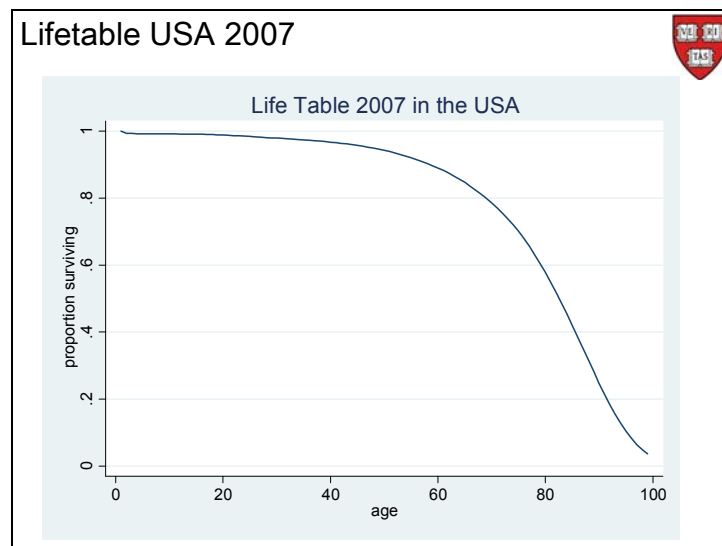
That means 99,278 ($=99,324 - 46$) reach their second birthday. Of those, the proportion who die before they reach their third birthday is 0.000286. So of these, 28 ($= 99,278 \times 0.000286$) die before they reach their third birthday. So 99,278 - 28 reach their third birthday.

And continuing this logic, we can tumble down the table to generate the survival experience of all 100,000 in this constructed cohort.

Life Table for 2007 USA			
Age	Probability of dying between ages x to $x + 1$	Number surviving to age x	Number dying between ages x to $x + 1$
	q_x	l_x	d_x
0-1	0.006761	100,000	676
1-2	0.000460	99,324	46
2-3	0.000286	99,278	28
3-4	0.000218	99,250	22
4-5	0.000176	99,228	17
5-6	0.000164	99,211	16
6-7	0.000151	99,194	15
7-8	0.000140	99,179	14
8-9	0.000124	99,166	12
9-10	0.000105	99,153	10
10-11	0.000091	99,143	9
11-12	0.000094	99,134	9
12-13	0.000132	99,125	13
13-14	0.000209	99,112	21
14-15	0.000314	99,091	31
15-16	0.000426	99,060	42
16-17	0.000529	99,018	52
17-18	0.000627	98,965	62
18-19	0.000715	98,903	71
19-20	0.000796	98,832	79

Now divide the third column (l_x) by 100,000 to turn the numbers into proportions and get the survival curve shown below.

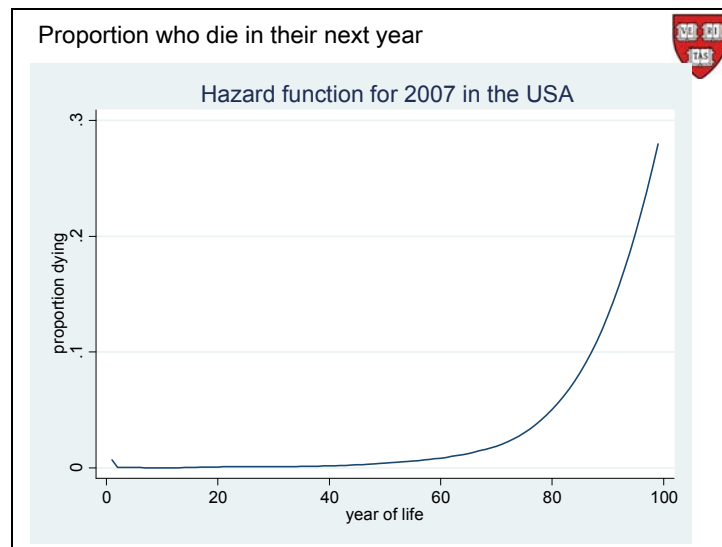
In passing, it should be noted that we could have taken another (more positive?) approach by looking at $(1-q_x)$ to find the probability of surviving the interval, and thence a particular l_x entry is simply the product of the $(1-q_x)$ and the l_x from the *previous* row. For example, $(1 - 0.006761)100,000 = 99,324$; $(1 - 0.000460)99,324 = 99,278$; etcetera. Why? (Answer, below.)



This is just a reminder of how we calculated the survival curve.

Life Table for 2007 USA			
	Probability of dying between ages x to $x + 1$	Number surviving to age x	Number dying between ages x to $x + 1$
Age	q_x	l_x	d_x
0-1	0.006761	100,000	676
1-2	0.000460	99,324	46
2-3	0.000286	99,278	28
3-4	0.000218	99,250	22
4-5	0.000176	99,228	17
5-6	0.000164	99,211	16
6-7	0.000151	99,194	15
7-8	0.000140	99,179	14
8-9	0.000124	99,166	12
9-10	0.000105	99,153	10
10-11	0.000091	99,143	9
11-12	0.000094	99,134	9
12-13	0.000132	99,125	13
13-14	0.000209	99,112	21
14-15	0.000314	99,091	31
15-16	0.000426	99,060	42
16-17	0.000529	99,018	52
17-18	0.000627	98,965	62
18-19	0.000715	98,903	71
19-20	0.000796	98,832	79

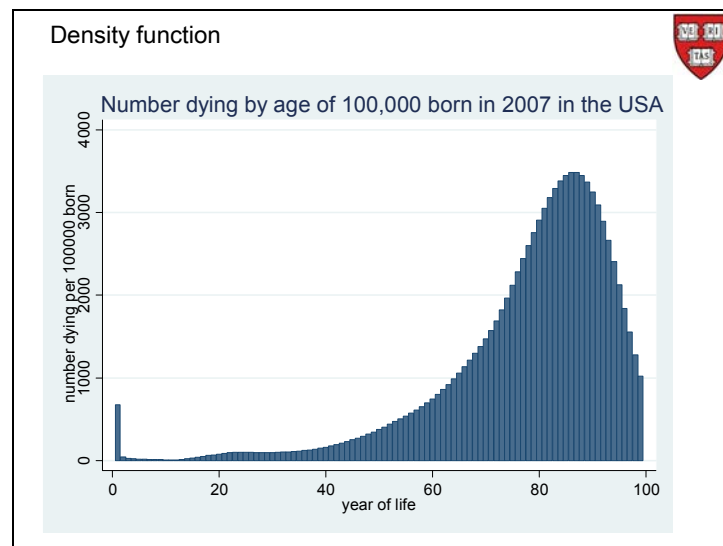
It is also interesting to look at the column labeled q_x . This gives us the values of the *hazard function* as x varies. That is the conditional probability of dying within an interval, given that one has survived to the beginning of the interval. So, for example, if we consider two-year-olds we see that the probability is 0.000286 that one of them does not reach their third birthday.



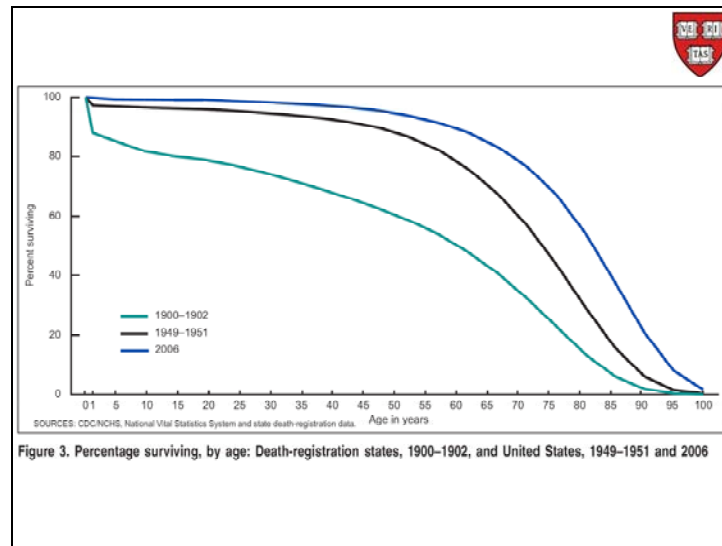
Plotting the hazard function shows that there is a blip in the first year of life, but subsequently, until age 40 or 50, the chance of dying within the next year of life is negligible. After that it increases exponentially. (Check this linearity out for yourself by plotting this function using a log scale on the vertical axis.)

Life Table for 2007 USA			
	Probability of dying between ages x to $x + 1$	Number surviving to age x	Number dying between ages x to $x + 1$
Age	q_x	l_x	d_x
0-1	0.006761	100,000	676
1-2	0.000460	99,324	46
2-3	0.000286	99,278	28
3-4	0.000218	99,250	22
4-5	0.000176	99,228	17
5-6	0.000164	99,211	16
6-7	0.000151	99,194	15
7-8	0.000140	99,179	14
8-9	0.000124	99,166	12
9-10	0.000105	99,153	10
10-11	0.000091	99,143	9
11-12	0.000094	99,134	9
12-13	0.000132	99,125	13
13-14	0.000209	99,112	21
14-15	0.000314	99,091	31
15-16	0.000426	99,060	42
16-17	0.000529	99,018	52
17-18	0.000627	98,965	62
18-19	0.000715	98,903	71
19-20	0.000796	98,832	79

The last column, the one labeled d_x , can also be plotted to show when each of the 100,000 in the cohort, die. It tells us how densely the individuals are packed into each age category and thus is called the density function (when normalized; i.e. divide by 100,000.)

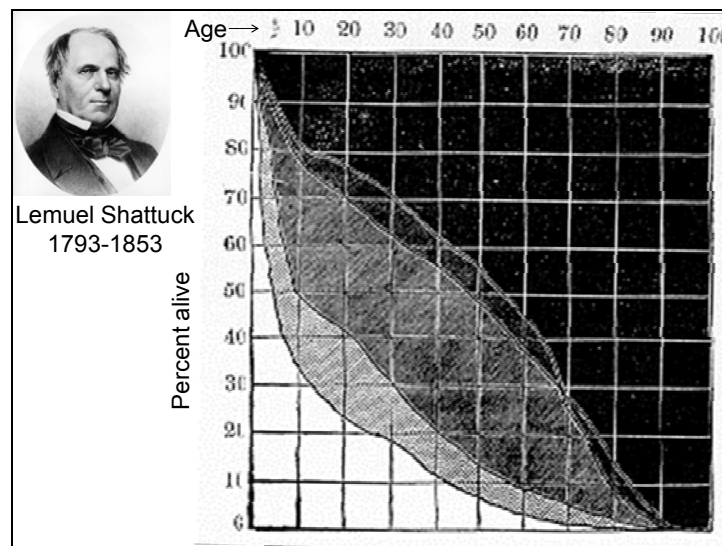


Technically the density function should have total area equal to one—easier to see if we divide the vertical scale by 100,000. From this we see that a relatively big number of babies die in their first year of life, but then very few die until the late teens. Then mortality picks up, with most of us dying in our late eighties, early nineties, and then decreases since there are not too many of us who live to our late nineties.



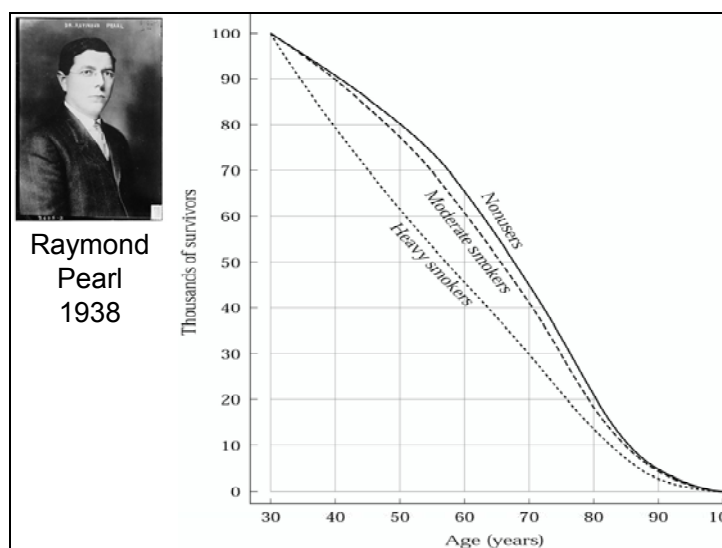
It is interesting to compare these life tables over the last century. At the beginning of the last century we see a very high mortality in the first year of life, which we do not see anymore. This improvement in mortality in the first years of life has had a huge impact on our life expectancy—remember that life expectancy is the area under the curve—because lifting the curve at the left resulted in lifting the curve throughout the life span.

The length of life has not changed much in the last century. What has changed is the proportions living longer—the curves do not extend much beyond where they have extended in the past, just the level of the curve on the right has increased.



These are the life tables Lemuel Shattuck published¹ in the document that was instrumental in the forming of the first State Health Department in the US: Massachusetts.

¹ Report to the Committee of the City Council appointed to obtain the census of Boston for the year 1845: embracing collateral facts and statistical researches, illustrating the history and condition of the population, and their means of progress and prosperity, Lemuel Shattuck, Boston (Mass.) John H. Eastburn, City Printer (1846)



We also looked at the depiction of the study of 7,000 individuals by Raymond Pearl comparing the three groups—nonsmokers, moderate smokers, and heavy smokers². These curves can be read to evaluate the effects of smoking.

Table 3. Cumulative percent of never-married males and females 15–19 years of age who have ever had sexual intercourse before reaching selected ages, by age, race, and Hispanic origin: United States, 1988, 1995, and 2002

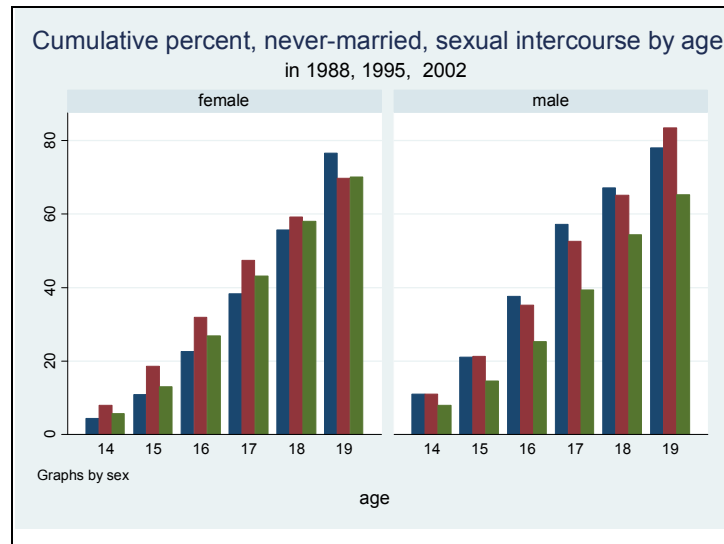
[Table 3-Supplement](#)

Characteristic	Female			Male		
	1988	1995	2002	1988	1995	2002
All never-married ¹	51.1	49.3	45.5	60.4	55.2	45.7
Age						
14 years	4.4	8.0	5.7	11.0	11.0	7.9
15 years	10.9	18.6	13.0	21.1	21.3	14.6
16 years	22.6	31.9	26.8	37.6	35.2	25.3
17 years	38.3	47.4	43.1	57.2	52.6	39.4
18 years	55.7	59.2	58.0	67.1	65.1	54.3
19 years	76.5	69.7	70.1	78.0	83.4	65.2

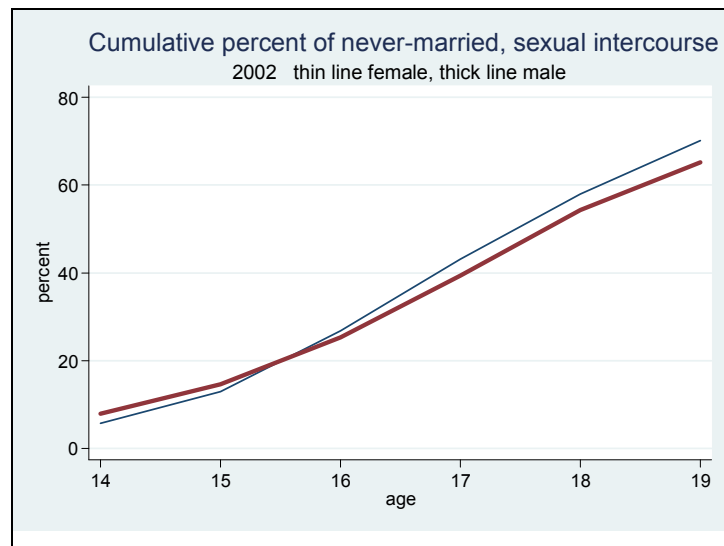
Survival methods can also be used to examine the time to an event, such as the first time to sexual intercourse. These are the results of three such surveys from the Centers for Disease Control and Prevention, for studying teenage behavior³.

² Pearl, R., 1938 Tobacco smoking and longevity. *Science* **87**: 216–217.

³ Abma JC, Martinez, GM, Mosher, WD., Dawson, BS. Teenagers in the United States: Sexual activity, contraceptive use, and childbearing, 2002. National Center for Health Statistics. *Vital Health Stat* 23(24). 2004.



We can plot these data as bar graphs. We can see that, as it should, the bars increase monotonically from left to right for each sex and for each survey.



Choosing a particular year, 2002 here, one can plot the cumulative curves. These cumulative curves are complementary to the survival curves in that the later go from 100% monotonically down to zero, whereas these curves monotonically increase from zero (not shown because we only have information in the 14-19 year age groups) to 100%.

With survival curves, if instead of plotting the percent who survived we plot the percent who die, we would then get a cumulative curve that monotonically increases with time; just like the one above. One minus the survival curve gives us a cumulative curve, namely a curve representing the percent who die.

Product Limit Method

Longitudinal

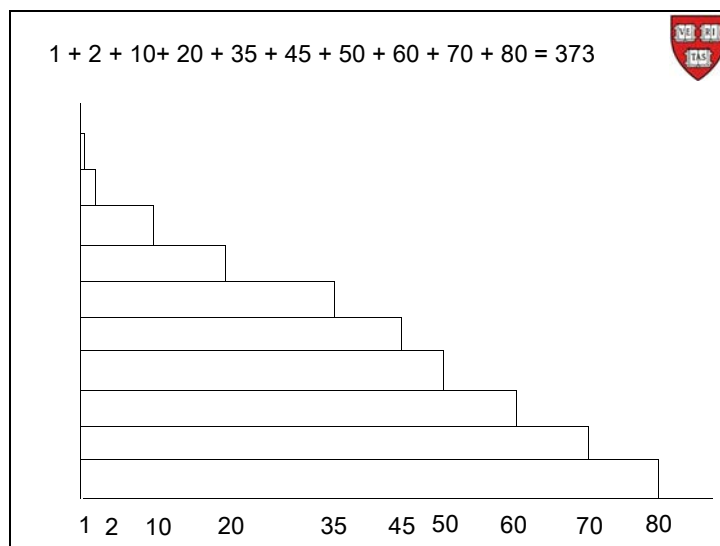


Note that all of these are cross-sectional studies attempting to describe a longitudinal process.

Another approach is to follow the cohort longitudinally.

In all of these examples that we have seen up to now, except for one, we tried to quantify a longitudinal experience—people living through their life spans—by using cross-sectional data—what happens to people over a year of life, by and large. It is a cohort we have constructed, to try to emulate what we might observe were we to follow them for a hundred years, or more.

What if we did have a longitudinal cohort, such as we had with the sticks—this is the exception we noted, above? How would we then estimate the survival curve? This is the longitudinal approach we now follow.



When we looked at the 10 sticks, we actually followed them until they all died, and this is the resulting survival curve. It is the properties of such a curve that we now study.

Longitudinal



Note that all of these are cross-sectional studies attempting to describe a longitudinal process.

Another approach is to follow the cohort longitudinally.

Consider a sampling approach, i.e. not the whole population.

Exactly what we look at is the situation where we want to make inference about a population survival curve, when having information on only a random sample from that population—think of the ten sticks as a random sample of ten individuals from the population. So we look at making inference about another curve in the population, namely the survival curve.

AIDS Data



Interval from AIDS to death Hemoph. < 41

Patient	Survival (months)
1	2
2	3
3	6
4	6
5	7
6	10
7	15
8	15
9	16
10	27
11	30
12	32

A case in point: Here is a sample of 12 individuals from an old study of hæmophiliacs under the age of 41, and we see their survival beyond their diagnosis with AIDS. And the question that was being asked in this aspect of this study was, whether there is an age effect on survival subsequent to being diagnosed with AIDS.

Life table approach:	
$[t, t+1)$	# dying at t
0-1	0
1-2	0
2-3	1
3-4	1
4-5	0
5-6	0
6-7	2
7-8	1
8-9	0
9-10	0
10-11	1

We can take a life table approach and split the time interval into months. Here are the first eleven such intervals. And we can note how many died in each interval—the left endpoint of the interval is considered in the interval, whereas the right endpoint is considered to be in the subsequent interval.

Life table approach:		
$[t, t+1)$	# survive to $t+$	# dying at t
0-1	12	0
1-2	12	0
2-3	11	1
3-4	10	1
4-5	10	0
5-6	10	0
6-7	8	2
7-8	7	1
8-9	7	0
9-10	7	0
10-11	6	1

Now create another column that keeps a tally of how many survive to the beginning of the interval. To continue with our prescription for creating a life table, we need to know what the probabilities are of failing within an interval.

Life table approach:

$[t, t+1)$	Prob dying in t to $t+1$	# survive to $t+$	# dying at t
0-1	0/12	12	0
1-2	0/12	12	0
2-3	1/12	11	1
3-4	1/11	10	1
4-5	0/10	10	0
5-6	0/10	10	0
6-7	2/10	8	2
7-8	1/8	7	1
8-9	0/7	7	0
9-10	0/7	7	0
10-11	1/7	6	1

Create a new column that calculates the probability of dying within the interval; the number who die in the interval divided by the number entering the interval. This gives us all we need to create a life table.

Life table approach:

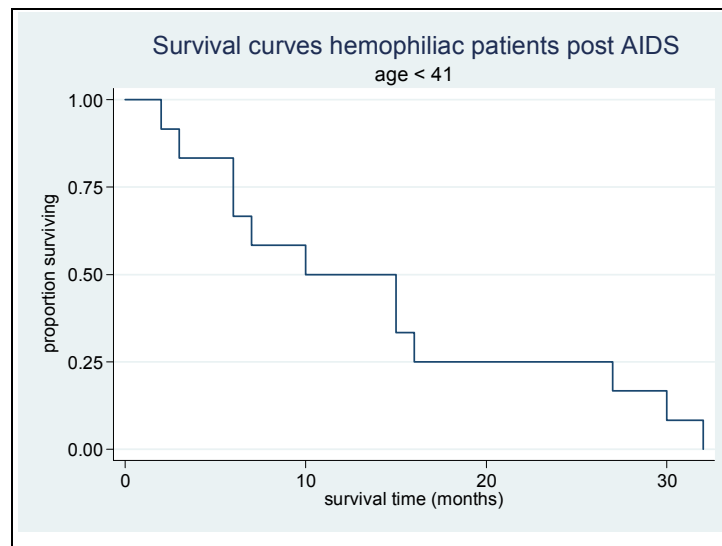
$[t, t+1)$	Prob dying in t to $t+1$	# survive to $t+$	# dying at t	$S(t)$
0-1	0/12	12	0	1
1-2	0/12	12	0	1
2-3	1/12	11	1	$11/12=0.92$
3-4	1/11	10	1	$10/12=0.83$
4-5	0/10	10	0	$10/12=0.83$
5-6	0/10	10	0	$10/12=0.83$
6-7	2/10	8	2	$8/12=0.67$
7-8	1/8	7	1	$7/12=0.58$
8-9	0/7	7	0	$7/12=0.58$
9-10	0/7	7	0	$7/12=0.58$
10-11	1/7	6	1	$6/12=0.50$

And voila, we've got the survival curve. So the survival curve is just this, the number surviving of the 12 who came in. To standardize, we divide by 12, to give us the last column.

That would be the life table approach to this summary. We could plot this survival curve, but before doing that, let us do some cleaning up. We see that nothing happens to the curve except when an event—a death in this example—happens. For example, the curve starts at 1 and continues there until at two months it goes down to 0.92. Then at three months it goes down to 0.83 and stays there until it reaches six months. Thus we could replace the 0-1 and 1-2 intervals with a 0-2 interval, and the 3-4, 4-5, and 5-6 intervals with a 3-6 interval. Or even, replace all the intervals by just the time points at which an event happens.

Product Limit Method: only at deaths				
$[t, t+)$	Prob dying in t to $t+$	# survive to $t+$	# dying at t	$S(t)$
0	0/12	12	0	1
2	1/12	11	1	$11/12=0.92$
3	1/11	10	1	$10/12=0.83$
6	2/10	8	2	$8/12=0.67$
7	1/8	7	1	$7/12=0.58$
10	1/7	6	1	$6/12=0.50$
15	2/6	4	2	$4/12=0.33$
16	1/4	3	1	$3/12=0.25$
27	1/3	2	1	$2/12=0.17$
30	1/2	1	1	$1/12=0.08$
32	1	0	1	0

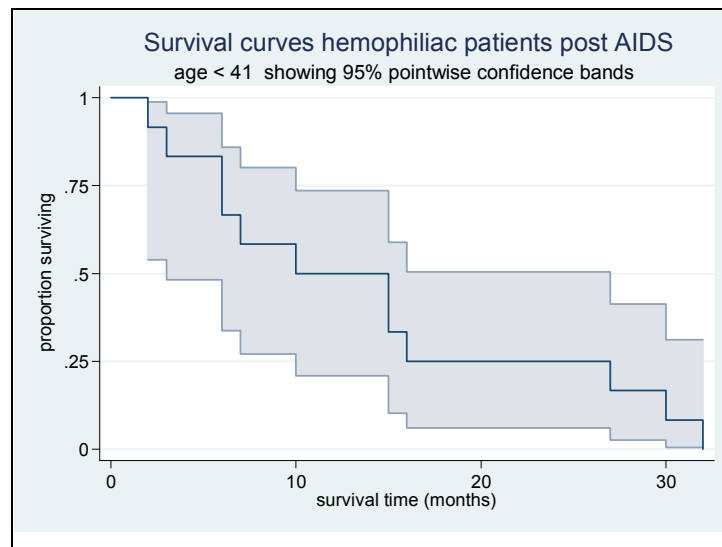
This is the product limit method of estimating the survival curve. It represents a minor change over the life table method.



Here is what the survival curve looks like, and it reminds us of the stick example we considered earlier.

Product Limit Method: only at deaths				
$[t, t+)$	Prob dying in t to $t+$	# survive to $t+$	# dying at t	$S(t)$
0	0/12	12	0	1
2	1/12	11	1	$11/12=0.92$
3	1/11	10	1	$10/12=0.83$
6	2/10	8	2	$8/12=0.67$
7	1/8	7	1	$7/12=0.58$
10	1/7	6	1	$6/12=0.50$
15	2/6	4	2	$4/12=0.33$
16	1/4	3	1	$3/12=0.25$
27	1/3	2	1	$2/12=0.17$
30	1/2	1	1	$1/12=0.08$
32	1	0	1	0


Looking at the rightmost column, the survival curve, we see that it is calculated as a ratio; the number alive of the number who started. This reminds us that these 12 patients were a sample of patients. In particular we have the estimate of the survival at that point and we can construct a confidence interval at any point in time⁴.



We can collect all these confidence intervals at every time point and display them as the shaded region above.

⁴ The theory for these confidence intervals are beyond the scope of this course. For the interested reader we refer you to, Greenwood, M. 1926. The natural duration of cancer. *Reports on Public Health and Medical Subjects* 33: 1–26.

So when you have a sample, you can treat this curve just like any statistic we have seen before, *and in particular* there is uncertainty associated with it, which we can quantify.




```
. sts list if age==1
      failure _d: death
analysis time _t: surv
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf. Int.]	
2	12	1	0	0.9167	0.0798	0.5390	0.9878
3	11	1	0	0.8333	0.1076	0.4817	0.9555
6	10	2	0	0.6667	0.1361	0.3370	0.8597
7	8	1	0	0.5833	0.1423	0.2701	0.8009
10	7	1	0	0.5000	0.1443	0.2085	0.7361
15	6	2	0	0.3333	0.1361	0.1027	0.5884
16	4	1	0	0.2500	0.1250	0.0601	0.5048
27	3	1	0	0.1667	0.1076	0.0265	0.4130
30	2	1	0	0.0833	0.0798	0.0051	0.3111
32	1	1	0	0.0000	.	.	.

Here is the *Stata* command which will *display* the numbers used *in* the above graph.

Censored Observations



Incomplete Longitudinal — Censoring

Suppose some of the patients are still not “dead” at the time the analysis is performed —

censored observations

One of the many complications associated with longitudinal studies, is *something* that we *call* censoring:

Consider a typical clinical trial. *It has* a start date when the trial *opens* to accepting patients; let us say January 1, 2006. Not all the patients arrive that day, but they accumulate over time, *and we typically assume that they arrive at random*. Suppose *the* endpoint of interest is survival. What that typically means is that *survival of the patients is measured beyond* a start point, *such as the point in time when the patient joins the trial*. In the prior graph it was the survival time after being diagnosed with AIDS. There is an implicit assumption of uniformity here; for example, that two months survival after being diagnosed in

January is to be considered the same as two months survival after being diagnosed in June. We make this assumption, and that allows us to draw a single curve incorporating all patients as starting from the same point.

Now suppose the study has been going on for a year and you would like to monitor the trial. For example, you may wish to conclude that the treatments are not too toxic, or, if two treatments are being compared, that one treatment is not far better than the other, since it would not be ethical to continue to put patients on the lesser treatment. So you decide to do an interim analysis of the trial. We should note that there are a number of famous trials that have been stopped earlier than planned because of interim results. Care must be taken when stopping trials early, and the statistical estimation methods subsequently used must reflect this early termination, otherwise they are statistically unsound.⁵

Returning to the analysis of an ongoing trial, the measurements may reflect that some of the patients have reached the endpoint, death, say, while others have not reached the endpoint and are still alive. How to calculate the survival curve? These “partial” observations—patients you have observed for some time but have not yet reached the endpoint—do provide partial information about the time to the endpoint, but not as much information as those who have died. For example, for the patient who is still alive three months after entering the study we only know that survival will be more than three months without knowing the exact survival time.

We could discard these partial data points, but by discarding the information we have on these patients we run the risk of introducing bias into our study. For example, if all the patients on the one treatment in a randomized trial are dead at the time we do our analysis, whereas all the patients on the other treatment are alive, then surely that could be very informative. So discarding such data is not a general solution.

These data are called censored observations. Why the censoring occurred may be informative—for example, if we see that a patient is responding poorly to a medication we would take that patient off the trial, and even though that would be a censored observation, it is also informative since we may classify it as a treatment failure—and that leads to a whole area of research that we do not pursue further, here. So we assume henceforth that the censoring occurs at random, or is non-informative—we cannot read anything into the censoring. For example, the family moved out of town and that is why we lost the patient—often labeled, lost to follow up.

Let us look at an example with uninformative censoring.

⁵ Bassler D, Briel M, Montori VM, Lane M, et al, Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA. 2010 Mar 24;303(12):1180-7.

Product Limit Method: only at deaths				
$[t, t+)$	Prob dying in t to $t+$	# survive to $t+$	# dying at t	$S(t)$
0	0/12	12	0	1
2	1/12	11	1	11/12=0.92
3+	1/11	10	1	10/12=0.83
6	2/10	8	2	8/12=0.67
7	1/8	7	1	7/12=0.58
10+	1/7	6	1	6/12=0.50
15	2/6	4	2	4/12=0.33
16	1/4	3	1	3/12=0.25
27	1/3	2	1	2/12=0.17
30	1/2	1	1	1/12=0.08
32	1	0	1	0

Returning to our hæmophiliac example, let us change these data and say, just for argument's sake, that the two patients, the one who died at 3 months and the other who died at 10 months, were both censored (indicated by a + sign); i.e. they were still alive at those time points, but we know nothing about them beyond those time points.

What this means is that Columns two through five should now be modified.

Product Limit Method: only at deaths				
$[t, t+)$	Prob dying in t to $t+$	# survive to $t+$	# dying at t	$S(t)$
0	0/12	12	0	1
2	1/12	11	1	11/12=0.92
3+	0/11	11	0	11/12=0.92
6	2/10	8	2	
7	1/8	7	1	
10+	0/7	7	0	
15	2/6	4	2	
16	1/4	3	1	
27	1/3	2	1	
30	1/2	1	1	
32	1	0	1	0

The first change to make is that zero persons died at both 3 months and 10 months. So the correct entries in Column two are 0/11 and 0/7, respectively. Also the entries in Column three should be 11 and 7, because 11 survived 3 months—now patient at time 3 did not die, so 11 reached time 3 and 11 were still alive immediately subsequent to time 3—and similarly for time 10 months—7 reached and survived 10 months. That means there should not be any change to the survival curve at either 3 months or 10 months. The changes to column four are clear. How to change column 5, requires some thought.

Product Limit Method: only at deaths					
[t,t+)	Prob dying in t to t+	# survive to t+	Prob surv in t to t+	# dying at t	S(t)
0	0/12	12	12/12	0	1
2	1/12	11	11/12	1	11/12=0.92
3+	1/11	10	10/11	1	10/12=0.83
6	2/10	8	8/10	2	8/12=0.67
7	1/8	7	7/8	1	7/12=0.58
10+	1/7	6	6/7	1	6/12=0.50
15	2/6	4	4/6	2	4/12=0.33
16	1/4	3	3/4	1	3/12=0.25
27	1/3	2	2/3	1	2/12=0.17
30	1/2	1	1/2	1	1/12=0.08
32	1	0	0	1	0

Let us return to the complete data tableau and introduce a fourth column, the more optimistic view of column two; namely, the probability of surviving the interval t to $t+$ (whereas column two is the probability of dying in the interval t to $t+$). So at every row, columns two and four need to sum to one.

The advantage of this new column (column 4) is that it can be used to calculate the probability of survival to any point. Pick a row. Let us say we choose the “6 month” row, and ask, what is the probability of surviving beyond 6 months?

One way to answer this is to observe that we must first survive to just before 6 months and then survive through the sixth month. From the above table, the probability of surviving to just before 6 months is $10/12=0.83$. Then the probability of surviving from 6 to just beyond 6 is $8/10$. So the probability of surviving to just beyond 6 months is $10/12 \times 8/10 = 8/12 = 0.67$ —the multiplicative rule of probability: $P(A \cap B) = P(A)P(B|A)$.

We can carry out this logic for every row and thus generate the whole table.

$[t, t+)$	Prob dying in t to $t+$	Prob surv in t to $t+$	$S(t)$
0	0/12	12/12	1
2	1/12	11/12	11/12=0.92
3+	1/11	10/11	10/12=0.83
6	2/10	8/10	8/12=0.67
7	1/8	7/8	7/12=0.58
10+	1/7	6/7	6/12=0.50
15	2/6	4/6	4/12=0.33
16	1/4	3/4	3/12=0.25
27	1/3	2/3	2/12=0.17
30	1/2	1/2	1/12=0.08
32	1	0	0

This slide schematically shows how to generate the whole survival curve. This is the recipe we use below to incorporate the censoring.

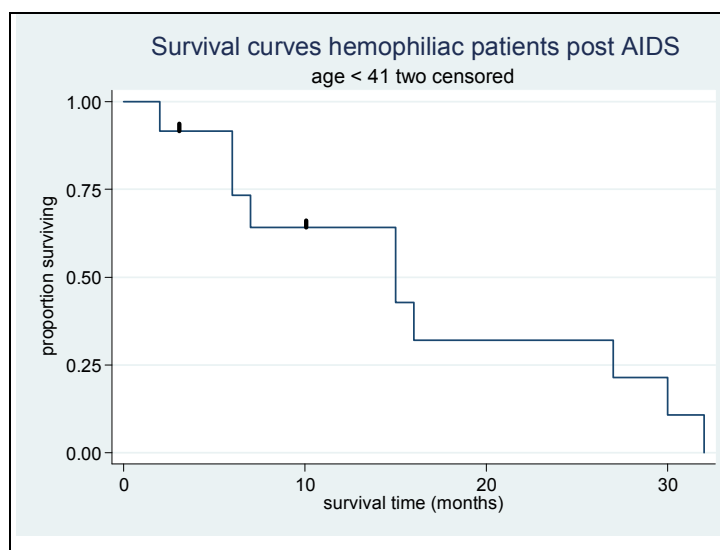
$[t, t+)$	Prob dying in t to $t+$	# survive to $t+$	Prob surv in t to $t+$	$S(t)$
0	0/12	12	12/12	1
2	1/12	11	11/12	0.92
3+	0/11	11	11/11	11/11x0.92 = 0.92
6	2/10	8	8/10	8/10x0.92 = 0.73
7	1/8	7	7/8	7/8x0.73 = 0.64
10+	0/7	7	7/7	7/7x0.64 = 0.64
15	2/6	4	4/6	4/6x0.64 = 0.43
16	1/4	3	3/4	3/4x0.43 = 0.32
27	1/3	2	2/3	2/3x0.32 = 0.21
30	1/2	1	1/2	1/2x0.21 = 0.11
32	1	0	0	0

Let us return to the censored observations and apply this logic to the new situation.

First let us look at the observation at month three. Now nobody dies, so column 2 is 0/11 and column 3 is 11. Thus column 4 becomes 11/11. Similarly at month ten, column 2 is 0/7 and column 3 is 7 and column 4 is 7/7.

Now with the new column 4 we can generate a new column 5, just like we did before in the complete data (no censoring) case.

This is called the product limit method.

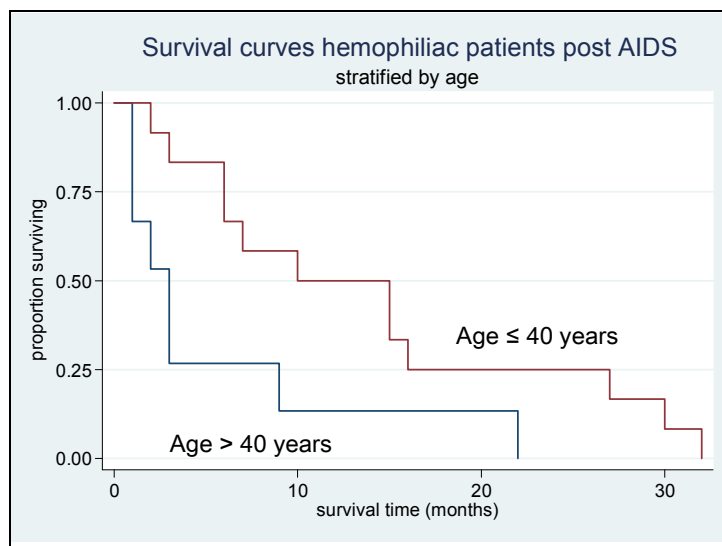


Here is the resultant curve—sometimes called the Kaplan-Meier curve. Two marks, one at 3 months and one at 10 months, show where the censored observations happened. So at the censored observations the curve remains flat and there are no steps.

Intuitively, one can see from the curve what the method does. Returning to the sticks or the curve when everyone died, then each step size was the same, $1/10$ for the sticks and $1/12$ for these 12 hæmophiliacs. That is exactly what happens in this curve, which includes censored observations, too in the beginning—at 2 months 1 person dies and the curve goes down $1/12$. Now when we reach the person who is censored at 3 months, the curve does not go down, but we must still utilize the $1/12$ of the whole that this person represents. What this method does is distributes that $1/12$ equally to the 10 people to the right of 3 months. Now each person to the right is going to be “worth” $1/12$ plus $1/12 \times 1/10 = 11/10 \times 1/12 = 0.092$. So at 6 months when 2 die, the curve should go down by 0.184, which indeed it does ($0.92 - 0.73 = 0.19$), up to roundoff. Similarly at 7 months, when another patient dies, the curve goes down by 0.09.

This time at 10 months when another patient is censored, that patient carries weight of 0.09 that needs to be distributed to the remaining 5 patients to the right of 10 months, so each of these five will carry weight of $0.09 + 1/5 \times 0.09 = 0.108$. That is the size of the steps the curve will take for every death that occurs.

This redistribution of the weights to the right of censored observations is like saying: as long as this person is alive the person enters into the denominator. Once the person leaves the study, I only know that the person is still alive and will die subsequent to the point in time when they left the study. How survival behaves beyond this point we can appeal to those still in the study, so let us say that this person who has left the study is equally likely to behave like any one of the remaining patients, and thus this method of equidistribution of the weight.



We now have seen how to estimate the survival curve and treat it like a statistic and quantify the uncertainty by calculating the confidence interval at points along the curve. We can also start to think about making comparative statements about two, or more, survival curves. For example, we can estimate the survival curves for the two samples of hæmophiliac patients once they have been stratified by age—say age 40—and ask whether these two curves are statistically different.

In order to answer that question, there are a number of statistical tests one can perform. A popular one is the logrank test. It calculates a “distance” between these two curves and, as usual, then determines how likely it is that the distance as large, or larger, can be attributed to chance.

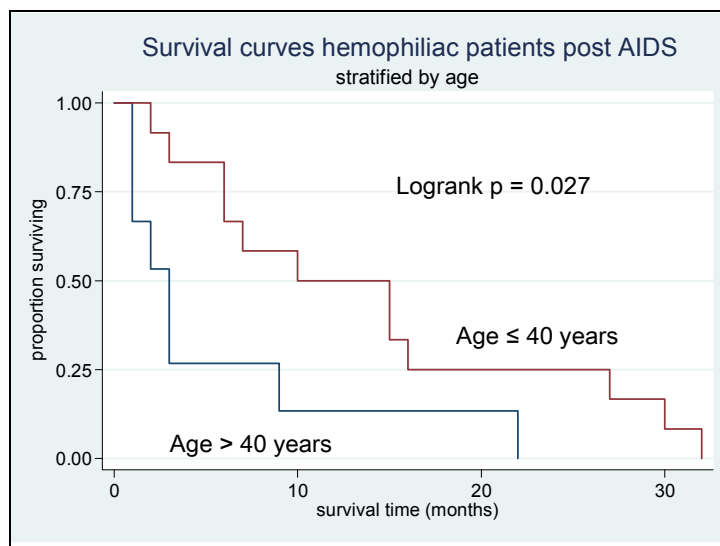
Logrank Test

At each *death* point construct a 2x2 table:

	Dead	Alive	Total
Treat 1			
Treat 2			
Total			

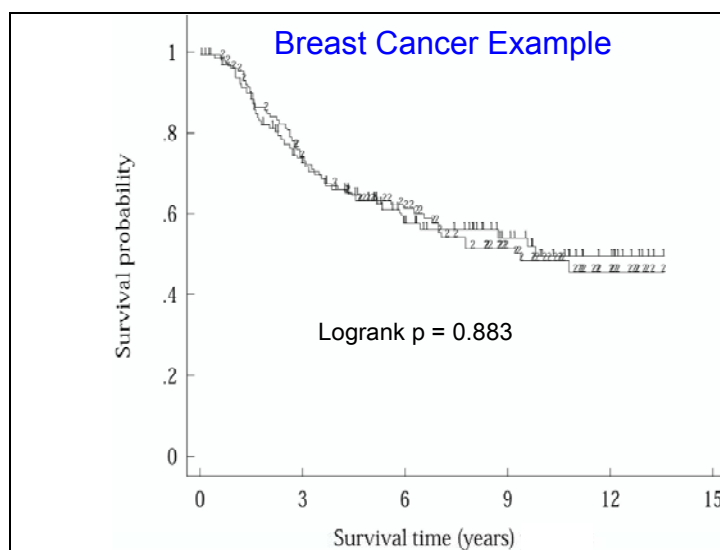
Then treat as g independent 2x2 tables in Mantel-Haenszel

The logrank test looks at every time point when either, or both, of the curves makes a change (where a death occurred), and creates a 2x2 table, as above. If there are a total of g distinct death times between the two groups, then we have g such tables. The logrank test then treats these as g independent 2x2 tables and appeals to the Mantel-Haenszel method for handling 2x2 independent tables.



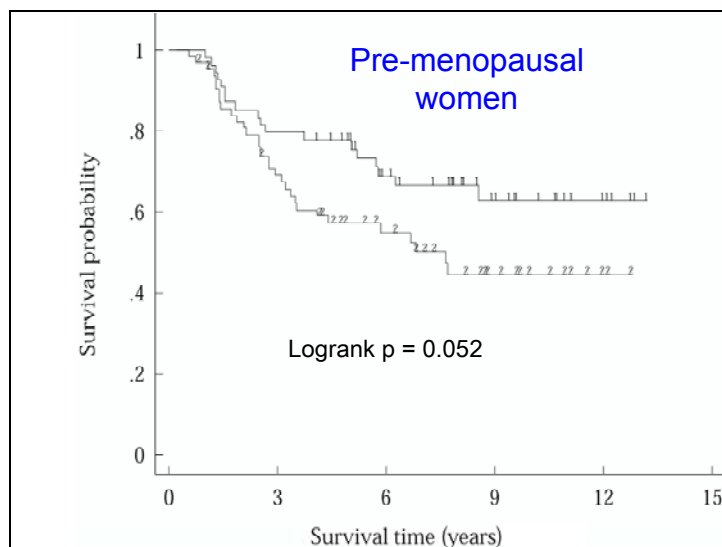
The result of the logrank test in this example yields a p-value of 0.027. So, on the basis of these two samples, we reject the null hypothesis that there is no survival difference between these two age groups at the 5% level.

We can also think about bringing in other covariates to explain the differences, and indeed, we can entertain the thought of extending regression consideration to the survival setting. That might lead us to the Cox model, but that is beyond the scope of this course. We look at one last example to show that even in a survival situation, we must still remain alert to the same problems we had before.

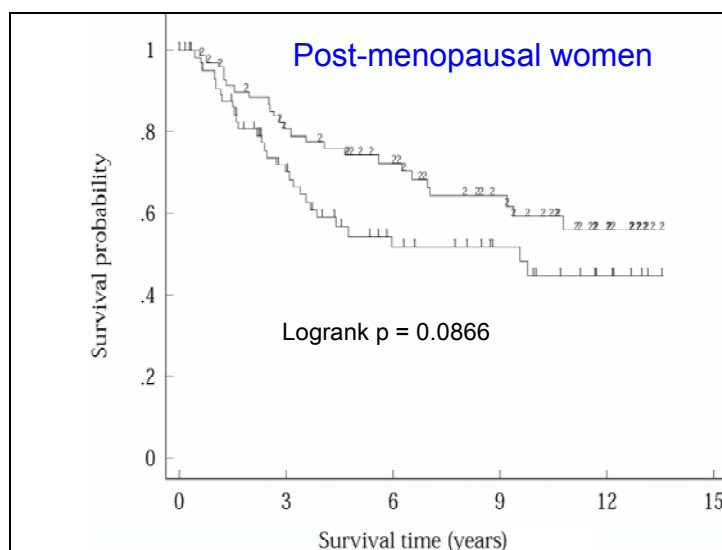


Here is a breast cancer study to compare survival experiences for two treatments. There is a lot of censoring evident, and to distinguish the two treatments, a “1” is used to indicate the censoring times for the first treatment, and a “2” indicates the censored values for patients on the second treatment.

From this graph it looks like there is no evidence of treatment differences as far as survival is concerned. The log rank p-value is 0.883 agreeing with our visual assessment that the difference between the two curves could easily be explained by sampling variability.



Instead of looking at the whole sample, let us just look at premenopausal women. Now we see a separation of the two curves, with the "1"s on top. The log rank p-value is much less than before although, at 0.052, is not significant at the 5% level.



When we look at post-menopausal women we once again see a separation of the survival curves. The separation is not sufficiently large to be significant at the 5% level by the log rank test, since the p-value is 0.0866.

It is interesting to note that both curves show separation when we look at the subgroups determined by menopausal status, neither significant, but much larger separations than when we look at the women as a single group; ignoring classification by menopausal status. Also, note that the positioning of the curves has flipped; whereas treatment “1” seems to be superior for the pre-menopausal women, treatment “2” seems to be superior for post-menopausal women. The reason for showing you this example was for you to see that no matter how complicated the outcome, or statistic, we are measuring, the Yule effect (Simpson’s Paradox) can rear its ugly head if you ignore a covariate.