Life tables: the basics

Life tables are used to measure the probability of death at a given age and the life expectancy at varying ages. Actuarial science and of course life insurance companies need to know this in detail, but we in public health do too. There are two different kinds of life table:

- Cohort or generational life tables
- Current or period life tables

Cohort life tables take an actual set of people born at the same time, usually in the same year or even on the same day of the same year, and follow them up for their whole lives. Several countries, including Norway, Denmark and the US, have these "birth cohorts" such as the Millennium Cohort Study in the UK that follows up people born in 2000. The mortality experience of such a cohort teaches us a lot and is great for history, but it's unlikely to be completely relevant to people born at other time points.

Period life tables take a hypothetical cohort of people born at the same time and uses the assumption that they are subject to the age-specific mortality rates of a region or country. These rates are often calculated using census data as the base population and actual age-specific death rates during the census year (and typically also one year either side).

How are life tables constructed? In a common type of epidemiological study called a cohort study, a set or cohort of patients are enrolled at time zero and then followed up to see who gets the outcome of interest, such as death, and when they get it. The latter will often be measured in days since the study start, but not necessarily. In theory you could measure it in milliseconds, but that's pretty silly unless you're looking at something like biochemical reactions. At time zero, a table of the numbers of people with and without the outcome at each time point will look like this. Let's suppose that we start off with 100 patients.

Time (t) in days	Number of patients alive at time t	Number of patients who died at time t	Probability of survival past time t
0 (study start)	100	0	1
1	100	??	??
2	??	??	??
3..	??	??	??

Everybody makes it past time zero, so the probability of surviving at least to time t=0 is 1, or 100%. This probability is technically known as the survival function, one of two core concepts in survival analysis. Let's now say that two people die the day after they are enrolled. The life table then looks like this:

Time (t) in days	Number of patients alive at time t	Number of patients who died at time t	Probability of survival past time t
0 (study start)	100	0	1
1	100	2	0.98
2	98	??	??
3..	??	??	??

The calculations continue in that way. However, this assumes that everybody enters the study at the same time, t=0, and no one leaves it except by death. It ignores the more realistic case when people drop out or are "lost to follow-up". The technical term for this is that these people are censored. Censoring is a really important concept in survival analysis. There are different forms, but the type due to people dropping out – or when people are still alive at the study end – is the most common. The Kaplan-Meier table and associated plot is the simplest (but not the only) way of estimating the survival time when you have drop-outs.

How to calculate a Kaplan-Meier table and plot by hand

The plot of the survival function versus time is called the survival curve. The Kaplan-Meier method can be used to estimate this curve from the observed survival times without the assumption of an underlying probability distribution. Some other kinds of survival analysis do require some kind of underlying distribution for the survival times, which we'll discuss later in the course, but one reason why the KM method is so popular is that it doesn't make any such assumptions. As you've seen in previous courses, whenever you make an assumption in statistics, you have to test whether it's valid.

To better understand the Kaplan-Meier method we'll now use it to draw a survival curve. Suppose we are monitoring patients after a particular treatment. After 5 days of follow-up we have the following information (example adapted from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1065034

Time (t) in days	Event
0 (study start)	8 patients recruited
1	2 patients die
4	1 patient dies
5.	1 patient dies
etc	etc

We're interested in the event 'death'. We can easily determine how many patients were alive at any given day and how many died and when.

Time (t) in days	Number of patients alive at time t	Number of patients who died at time t	Proportion of patients surviving past time t	Probability of survival *past* time t
0 (study start)	8	0		
1	8	2		
4	6	1		
5	5	1		

But how do we compute the probability of survival past time t? Start by computing the proportion of patients that survive day t, i.e. of those alive at the beginning of day t, what proportion make it to the next day alive? On day 0, the day the study begins, there are no deaths. Everybody survives. Hence the proportion surviving is 1. On the following day, 2 out of 8 patients don't make it; in other words, 75% survive the day.

Time (t) in days	Number of patients alive at time t	Number of patients who died at time t	Proportion of patients surviving past time t	Probability of survival *past* time t
0 (study start)	8	0	(8-0)/8=1	
1	8	2	(8-2)/8=0.75	
4	6	1	(6-1)/6=0.83	
5	5	1	(5-1)/5=0.8	

Now we know the proportions, but what are the probabilities? With no deaths on day 0, the probability of surviving is 1. Computing the next probability is a bit trickier. The basic idea underlying Kaplan-Meier tables comes into play here: the probability of surviving past day t is simply the probability of surviving past day t-1 times the proportion of patients that survive on day t. Let's see it together:

Time (t) in days	Number of patients alive at time t	Number of patients who died at time t	Proportion of patients surviving past time t	Probability of survival *past* time t
0 (study start)	8	0	(8-0)/8=1	1
1	8	2	(8-2)/8=0.75	1 * 0.75 = 0.75
4	6	1	(6-1)/6=0.83	0.75*0.83 = 0.623
5	5	1	(5-1)/5=0.8	0.623*0.8 = 0.498

If we now plot the time column against the probability column, we end up with a survival curve. We plot the time on the x-axis, running from 0 on the left to the highest day count, i.e. 5 in this example, on the right. The probability of survival goes on the y-axis, with 0 on the bottom and 1 as the maximum (Figure 1).

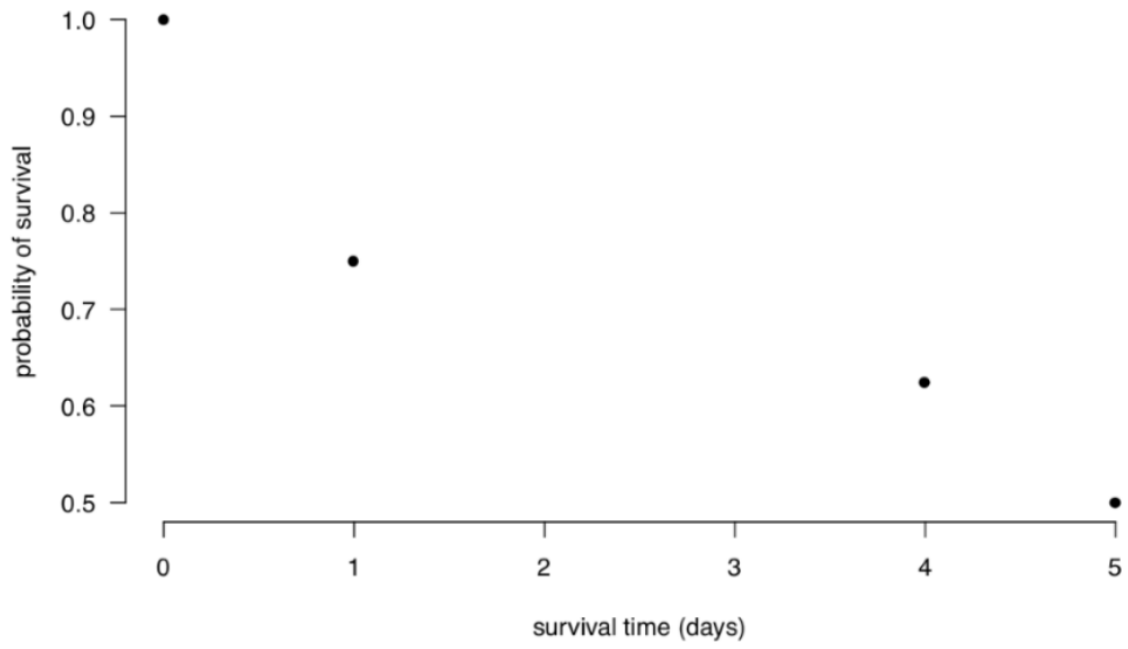

Figure 1. Data points for survival curve for Kaplan-Meier example

If we now connect the dots using steps, first horizontal then vertical, we have drawn our first survival curve (Figure 2).

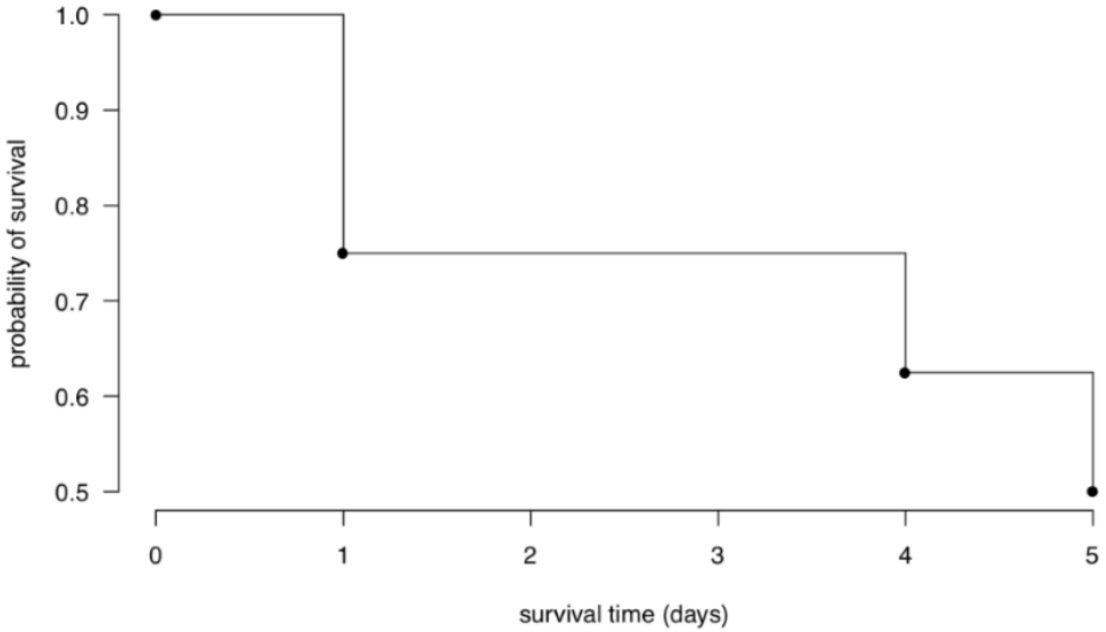

Figure 2. Survival curve for Kaplan-Meier example

You might wonder why the "steps" involve a horizontal line followed by a vertical line and not the other way around. This is because the probability is assumed to be the same until the next death occurs. For example, there's a death at day 1 but then no more deaths until day 4.

Let's follow up our patients for another two weeks. This is what we have:

Time (t) in days	Event
0 (study start)	8 patients recruited
1	2 patients die
4	1 patient dies
5.	1 patient dies
6	1 patient drop out
9	1 patient dies and 1 drops out
22	1 patient dies

As you can see, we now have some drop outs, i.e. patients whose outcome we don't know exactly. These patients are censored and should be treated differently from patients that die. When a patient is censored at time t, we know the patient was alive at time t, but we don't know whether the patient has died or survived. For this reason, censored patients are classified neither as 'survived' nor as 'died' on any given day. We simply deduct them from the number of patients alive. When there are censored patients at the same time as patients that die, we deal first with patients that die. Then we add a new line, mark it with a little '+' right after the time count and denote the censored patient(s) by taking them off the count of patients alive at time t.

Time (t) in days	Number of patients alive at time t	Number of patients who died at time t	Proportion of patients surviving past time t	Probability of survival *past* time t
0 (study start)	8	0	1	1
1	8	2	0.75	0.75
4	6	1	0.83	0.75*0.83 = 0.623
5	5	1	0.8	0.623*0.8 = 0.498
6+	4	0	4/4=1	0.498*1 = 0.498
9	3	1	(3-1)/3=0.667	0.498*0.667 = 0.332
9+	2	0	2/2=1	0.332*1 = 0.332
22	1	1	0/1=0	0

At time t=6 and t=9, we need to subtract one person from the risk set, the number of patients at risk of death. For t=6, for instance, 4 people enter that time period but one drops out, leaving 3 to go forward to t=7, which means that at the time of the next death, t=9, the proportion surviving is (3-1)/3.

At times when no one dies, the proportion surviving that time point is 1, or 100%, so the (cumulative) probability of survival past time t in the last column is unchanged.

At the last time point, t=22, there's only one person left in the risk set, i.e. only one person who we're still following up, and they then die, giving a final probability of survival beyond t=22 of zero.