WEEK 1

A) The Kaplan-Meier Plot

0) Premises

The survival analysis was born in the 17^{th} century.

- In the first course of this series it will be explained how Statistics plays a crucial role in understanding almost anything involving quantitative data.
- In the second course they will explain hot to assess the relation between a set of patient characteristics and a normally distributed outcome variable;
- In the third course how to analyse a binary outcome variable such as the presence or absence of diabetes and tricky issues such as how to select your predictors for the model when you have a lot of possible ones to choose from.

1) Why not just survival analysis? There are some particular features of the public health datasets. Public health considers also the population angle. Let's suppose we have found a given predictor that shows a strong relation with the outcome, this could be for example liver disease as a strong predictor of mortality. Being it a specific feature that only a few ppl have public health will focus on more common features. It is fundamental to decrease the mortality rate in public health

2) What is survival analysis?

It's is not only used to analyze survival data. This is also why sometimes is called "time to event analysis" because the event of analysis isn't always death. It is a family method, and we will see the 2 most common and also how its is different from regression analysis.

I) Events that can be analysed: patient outcome, what happened to patient between the beginning and end of the experiment. It could be death but also non-fatal events. In the latter case we need to be careful especially with the ppl who died before having any of those outcomes.

II) In logistic regression we are interested in whether the outcome happened, while in survival analysis we are interested also in how long it took to the patient to have that outcome -> time to event.

For ex. We want to know the effects of the lifestyle on death. We follow the ppl chosen, until they die monitoring their lifestyles. Another difference btw logistic and survival analysis. If we follow ppl for say, 10 years in the lifestyle ex. it is very common for some ppl to be "lost to follow-up" before the end of the study. They may move away or drop out, so we don't know if they are still alive. And for non fatal outcomes is a problem, but not always for survival analysis.

III) Summary: survival analysis can be used to explore the relation btw patient factors of interest and the time to any binary event. Logistic regression is only interested in the facts of that binary event. This event can be death but it can also be disease recurrence, etc.

B) The KM plot and Log-rank test

1) Kaplan-Meier plot

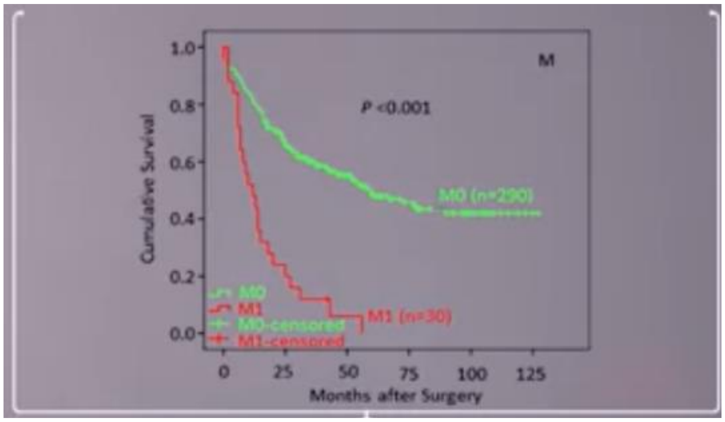
If we developed a serious disease that could shorten our life we would like to know how long we've got, but also the chances of surviving the next year or the next 5 years. The km plot allows us to answer these questions. It also let's us compare survival for different patient groups; it estimates the probability of surviving to any given time point.

This probability is technically known as the SURVIVAL FUNCTION. When we plot it for a range of times we get the km plot.

Example: the TNM staging system is useful in predicting survival following a cancer diagnosis.

- T = size of the tumor
- N = describes the nearby lymph nodes that are involved;
- M = describes distant metastasis spread to other parts of the body

In his example the km plot shows the survival probability over time since surgery by the m component, so whether the cancer has spread. At the time of surgery, on the left the cumulative survival probability is 100% . As ppl die the probability falls giving the line a jagged appearance.



- The green m0 line is for ppl without spread to other body parts;
- The red m1 line is for ppl for whom cancer has spread.

The prognosis for the latter group is worse than for the former, the line separates after the very beginning.

There is a test to compare the survival curves called LOGRANK test and a p<0.001 is highly significant. We can also see the presence of the term censored: when a patient is censored it means that he/she has dropped out of the study and we do not know whether they are still alive (second core concept after the survival function). Every time a patient is censored a vertical line is placed over the line so it looks like a cross. The logrank test compares the survival function for 2 little groups of patients and gives a summary p-value.

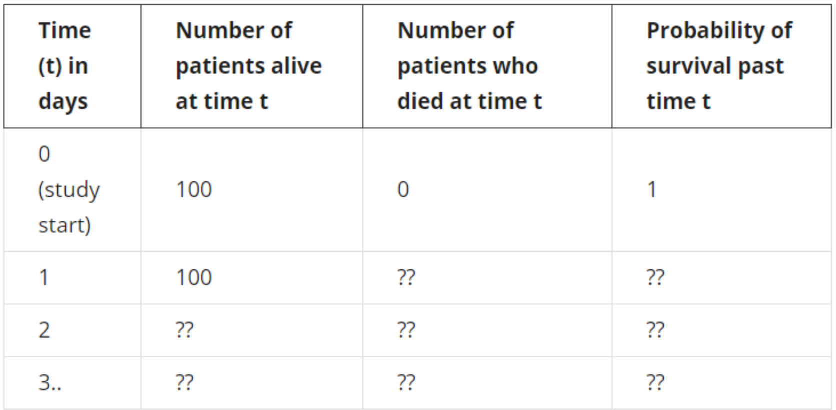
2) Life tables

I) Life tables are used to measure the probability of death at a given age and the life expectancy at varying ages. There are 2 types of life tables:

- o Cohort or generational life tables = they take an actual set of ppl born at the same time (usually the same day of the same yy) and follow them for their whole lives. The mortality experience of such cohort teaches us a lot;
- o Current or period life tables = take a hypothetical cohor of ppl born at the same time and uses the assumption that they are subject to the age-specific mortality rates of a region or country. The rates are often calculated using census data as the base population and actual age-specific death rates during the census year.

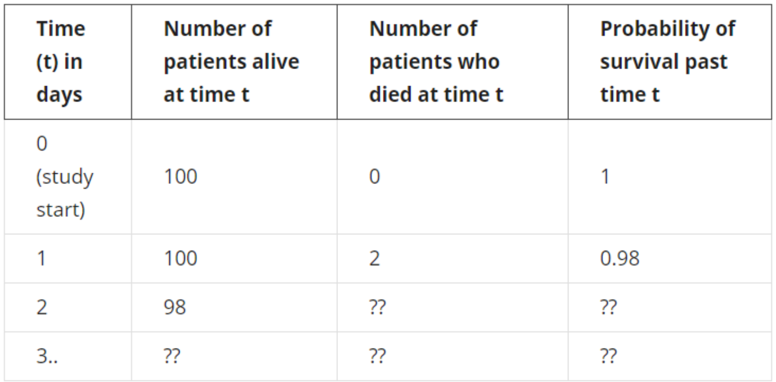
II) How are life tables constructed?

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 and when they get it. The latter is often measured in days. At time zero a table of the numbers of ppl with and w/o the outcome at each time point will look like this (suppose 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..	??	??	??

The probability of surviving at least to time t=0 is 100% and this probability is known as the SURVIVAL FUNCTION. Let's now say that 2 ppl die the day after they enrolled, the new life table will look 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..	??	??	??

And so on and so forth. This assumes that everybody enters the study at the same time, t=0 and that no one leaves it expect by death, ignoring the cases of *drop out* and *"lost to follow up".* The technical term for this is that these ppl are CENSORED.

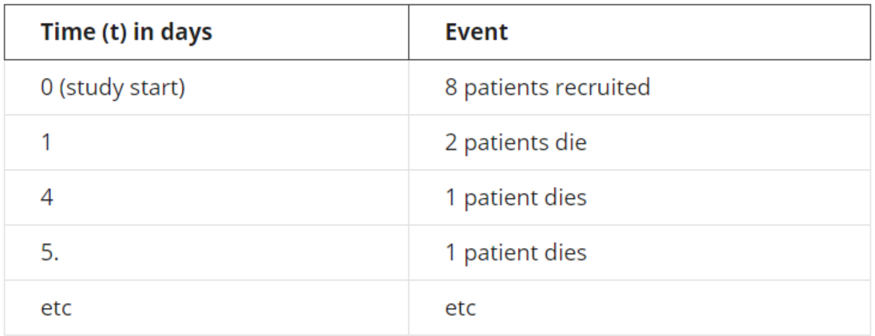
III) How to calculate a K-M table and plot by hand

SURVIVAL CURVE = plot of the survival function versus time.

The Kaplan-Meier method can be used to estimate the survival curve from the observed survival times w/o the assumption of an underlying probability distribution. Other kinds of survival analysis do require some kind of underlying distribution for the survival times (later), but the reason why KM is so popular is that id doesn't make any of such assumptions (any statistical assumption would need to be tested for validity).

Example adapted from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1065034

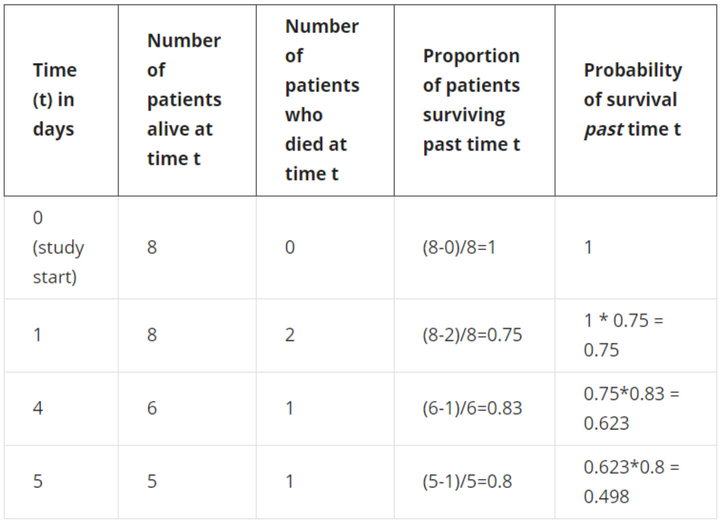
Let's suppose we are monitoring patients after a particular treatment; after 5 days of follow up e have the following info:



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

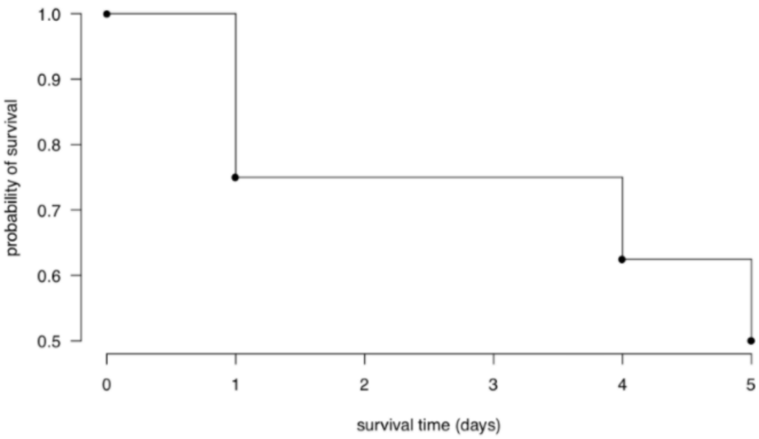
We can determine how many patients were alive at any given day and how many died and when. How to compute the probability of survival past time t? We start by computing the proportion of patients that survive day t, what proportion make it to the next day alive? On day 0, the day the study begins, there are no deaths. Hence the proportion of surviving is 1. On the following day, 2 out of 8 patients don't make it -> 75% survive the day.

Now that we know the proportions we need to compute the probabilities. With 0 deaths on day 0 the probability of surviving is 1. For the next probability the K-M tables come 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

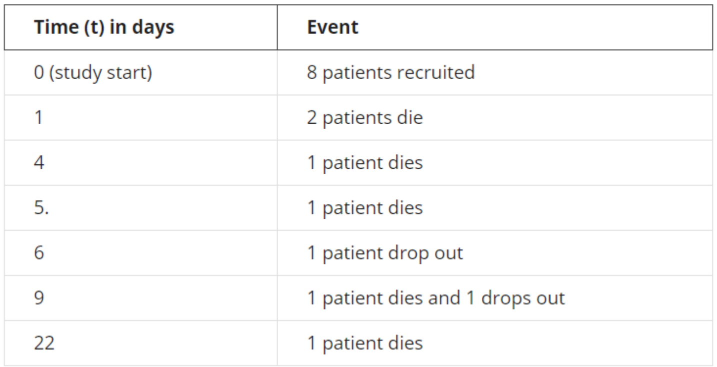


Time (t) in days	**Number of patients alive at time t**	**Number of patients who died at time t**	**Proportion of patients surviving past time t**	**Probability of survival *past* time t**
0 (study start)	8	0	(8-0)/8=1	1
1	8	2	(8-2)/8=0.75	1 * 0.75 = 0.75
4	6	1	(6-1)/6=0.83	0.75*0.83 = 0.623
5	5	1	(5-1)/5=0.8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

If we plot now the time column against the probability column we end up with the survival curve, and by connecting the dots using steps we draw our 1st survival curve:



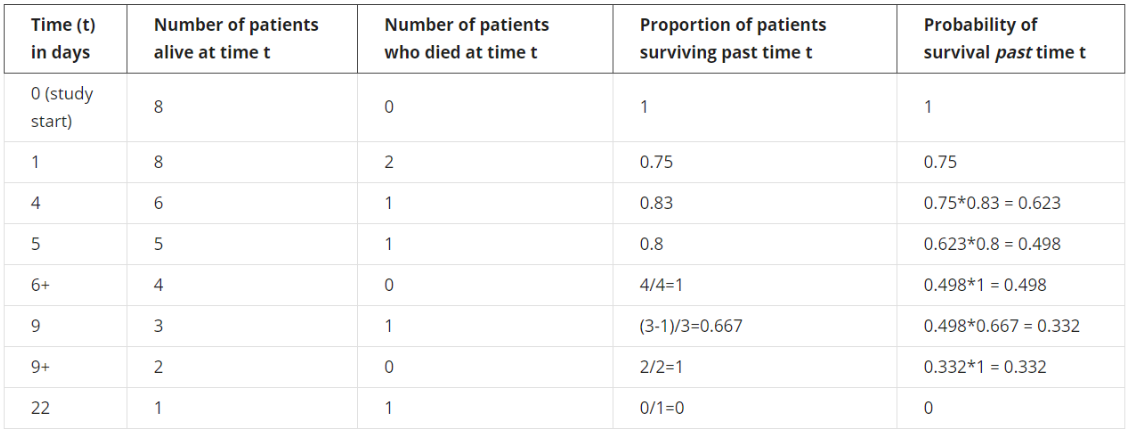
We now extend the time of observation to 2 weeks including some drop outs. The patients who dropped out are censored and should be treated differently, in particular they are classified neither as ‘survived’ nor as ‘dead’ on any given day. We simply deduct them from the number of patients alive. When 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 patients by taking them off the count of patients alive at time t.



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

At time 6 and 9 we need to subract one persone from the risk set, the number of patients at risk of death. At times when no one dies, the proportion surviving that time point is 1 so the cumulative probability of survival past time t in the last column is unchanged.

At the last time point, t=22, there is only one person left in the risk set, namely only 1 person who we are still following, giving a final probability of survival beyond t=22 of zero.



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

3) What is Heart Failure and How to run a KM plot in R

I) The dataset

We are going to use a set of simulated data based on real hospital administrative data in England for a group of patients admitted for heart failure. Every public hospital in the country must submit records for every admission; private hospitals also submit records for any NHS patients that they treat. The other UK countries and Ireland have similar databases. These can be linked to the national death registry in order to capture deaths that occur after discharge. The simulated extract contains a random sample of emergency (unplanned) admissions for heart failure.

II) What is heart failure and why it is a major health policy problem.

The dataset we will use and how to analyze it. Heart failure is a condition where heart is unable to pump blood around the body properly, usually because the heart has become too weak or stiff. The many causes, particularly coronary heart disease, high blood pressure, arrhythmias but also less common one such as arrhythmias.

Main symptoms are shortness of breath and swollen ankles, it is common in the elderly who are increasing in number in many countries. Emergency readmission rates are high, typically around 1 in 4 in 30 days of discharge. Many countries have adopted electronic hospital admission databases that record info admitted to the hospital, diagnosis, age, gender, date of admission and discharge, death, etc typically used for insurance and billing but also for research.

III) How can we use this dataset to investigate the predictors of mortality in patients with heart failure? With start with the km model. We will use the package “survival” also for the COX regression. *survfit* is one of the objects created by R

IV) The log-rank test compares the survival time by gender. It’s the most popular method of comparing the survival of patient groups that takes the whole follow-up period into account. Its big advantage is that it you don’t need to know anything about the shape of the survival curve or the distribution of survival times. It’s based on a comparison of the observed numbers of deaths and the numbers of deaths expected if in fact there were no difference in the probability of death between the groups (genders in this case) and uses a chi-squared test. The resulting p-value that you should have got is high, at 0.8. There’s therefore no good evidence of a difference between the genders in their survival times.

From (link: http://www.sthda.com/english/wiki/survival-analysis-basics#log-rank-test-comparing-survival-curves-survdiff)

The *log-rank test* is the most widely used method of comparing two or more survival curves. The null hypothesis is that there is no difference in survival between the two groups. The log rank test is a non-parametric test, which makes no assumptions about the survival distributions. Essentially, the log rank test compares the observed number of events in each group to what would be expected if the null hypothesis were true (i.e., if the survival curves were identical). The log rank statistic is approximately distributed as a chi-square test statistic.
The function *survdiff*() [in *survival* package] can be used to compute *log-rank test* comparing two or more survival curves.
The log rank test for difference in survival gives a p-value of p = 0.0013, indicating that the sex groups differ significantly in survival. (less than 0.01 se non erro).

Compare the survival times for patients 65 and over with those under 65: This time, we got a low p value, one that's way below the conventional 5% threshold, so you'd conclude that survival times do differ by whether you've turned 65. But which group live longest after their hospital admission? You'd expect the younger group to live longest of course, but you don't know that until you look at the above table. The 115 younger patients (those with age_65plus = 0) had 18 observed deaths, but you would expect 67 under the null hypothesis of no difference in survival times by age group. In contrast, the older group had more deaths than expected under the null, which confirms your instinct that younger patients live significantly ($p<0.001$, in fact very near zero) longer after hospital admission than older ones do.