# Advanced Topics in Survival Analysis 1

## Transcript

For full resource, see: <https://www.ncrm.ac.uk/resources/online/all/?id=20860>

Hello. I'm Dr Oliver Perra and this is the first part of my resources on Advanced Topics in Survival Analysis.

 In this first part, I will first provide a summary of the key hazard and survival functions for discrete time events. It's something I covered in other resources I have prepared for the National Centre for Research Methods.

 I will then talk about the hazard and survival functions when time is measured continuously and I will particularly talk about challenges and the importance of moving from instance to intervals.

 And then I will introduce the Kaplan-Meier method talking about the estimates, the cumulative hazard function and the importance of using Kernel-Smoothed hazard functions.

 So, in this summary, I will remind you that survival analysis is used when we are interested in studying the occurrence of events as well as the factors that may influence this occurrence and the timing of this occurrence.

 So here I give some examples of events that might be of interest in survival analysis. These events may happen only once, for example, death of a patient or the onset of puberty, but we can also apply these methods to reoccurring events, for example, we may be interested in the occurrence of depression symptoms over years whereby some individuals may experience depression symptoms more than once.

 Another example I may use, it's in finance. If we were interested in studying if and when shareholders will sell the shares in the stock market, that is something that we can also study with survival analysis.

 And to run survival analysis, we need to clearly define an event of interest. That is, we need to define a change in participant status that we can observe. For example, individuals move from pre-pubertal to puberty or they move from not experiencing depression symptoms to experiencing depression symptoms.

 The other key precursor to survival analysis is identifying the start of time. When we are interested in advanced occurrence, we need to identify a point where all the participants in a study are in only one status, so they haven't experienced the event yet.

 In the puberty example, we need to identify an age when all the participants had not yet started puberty. In the stock market example, the beginning of time may be the opening of the stock market. That is a point where all the shareholders own their shares before operation starts. Again, for an introduction to survival analysis, you can refer to the resources I had created for the National Centre for Research Methods.

 When we are running survival analysis, it’s then important to specify a metric for measuring time. So, what are the units that the researchers will use to record the passing of time and the timing of the events of interest?

 With current technology, it may be possible to have very precise measures of event occurrence. For example, when we use eye tracking devices we could have measurements in milliseconds. However, some type of events can only be recorded within discrete intervals. For example, career progression within academia takes place within years and it doesn't take place within months or days.

 As I will illustrate in the following of this presentation, even when we can record events taking place using a very precise time measurement, most time we need to constrain the measurement within time intervals.

 A key statistic in survival analysis is the hazard function and here I will provide a premise. I will only focus on repeatable events in the rest of these presentations, and this is because they are easier to understand and the examples will be easier to follow. So, I will only focus, or mostly focus, on events that can only happen once, for example, death.

 So, when we consider non-repeatable events, the hazard function represents the risk that a participant in the study will experience the event within a time interval, provided that the individual has not experienced the event before this interval. And to indicate intervals, I will use the lowercase letter j whereas the lower case letter i will indicate an individual in the study.

 So, the hazard function, in other words, it indicates the risk of the event occurring, taking place within a specific interval j. And the table here is a live table from a study in which male secondary students reported if they had the first sexual intercourse. The students reported every sexual intercourse within each school year, so the time interval here are school grade years and here, for example, I highlighted the statistics for time for second year Grade 9. The hazard function in this case is calculated simply by dividing the number of participants who experienced the event within that interval by the number of participants that were at risk of experiencing the event at the start of the interval. In other words, the hazard function is the number of participants who experienced the event for the first time within that interval divided by the number that had not yet experienced the event at the beginning of the interval. And in this example, if we take Grade 9, the hazard function is 18/158.

 I will also highlight that when we consider events measured in discrete time, the hazard function is a conditional probability. It represents the probability of event occurrence conditionally on participants not having experienced the event before.

 When we measure event occurrence using a continuous measure of time, we encounter a problem. The issue is that there is an infinite number of instance where the event can take place and here I try to represent it with the fading colours on a line and the red dot represents an event. This means that the probability of an event taking place at any given instant will approach zero as the units of time become smaller.

 Consider, for example, that we are trying to measure the first occurrence of craving for tobacco during a day among a group of people who are trying to quit. So, the first occurrence of cravings can be measured in hours and minutes. But even then, if we had better instruments, we could potentially measure the first occurrence of cravings in seconds or, in principle, we can even measure them in milliseconds and so on which means that we can potentially break down time into infinitesimally small units making in this way the probability of an event occurring within a given instant virtually zero.

 And so how can we resolve this problem? The way to resolve this problem is by accumulating instants of time into intervals. In other words, time is divided into a series of intervals. For example, if we measure event occurrence in minutes, minutes are just intervals made up of 60 seconds and in this example I reported here graphically with the timeline divided into small boxes, small frames, we might consider the occurrence of an event within five time intervals.

 So, this may seem trivial, but it has an important implication because the hazard function then is calculated as the probability of the event taking place within a specific interval, provided that it did not occur before as the interval with approaches zero and that's an important caveat. So, as the interval approaches zero, we measure the occurrence of event within that interval if conditionally on the participant not having experienced the event before that interval.

 But since we are considering probability of events within intervals, this probability will change with intervals of different width. So, if we consider the probability of experiencing cravings for the first time within minutes, the probability will change when we consider intervals of one minute or intervals of five minutes.

 So, to ensure we have a common metric then, we must divide the conditional probability of the event taking place within an interval by the interval’s width. And the issue that we need to consider, the key issue here, is that by doing so, the hazard function does not represent a probability anymore, but rather a rate, represent the rate of the probability taking place within an interval.

 And in this example, if we consider intervals of five minutes, so the hazard function will represent the conditional probability of the event taking place each five years.

 And this calculation of the hazard function whereby when we consider discrete time, it's a conditional probability. When we consider continuous time, it’s a rate. This also has repercussions on the interpretation of hazard function when we consider different types of events. And here I try to represent that. So, when we are considering events that are repeatable, the interpretation of the hazard function is linked with an individual and in fact if the events are repeatable, the hazard function represents how often an individual can expect the event to take place per unit of time.

 For example, if we are studying cravings for cigarettes and we considered that cravings can happen more than once in a day, the hazard function here will represent the expected number of cravings of an individual per five minutes, per units of five minutes. But if the event is non-repeatable, the interpretation of the hazard rate is linked with a hypothetical population of individuals. For example, if we are studying how long before investors sell the shares of a company during a specific day in the stock market and we consider units of time in periods of five minutes, a hazard rate of 0.04 means that within a five minute interval we expect 4% of investors that still hold the company's shares to sell them.

 I will come back to the hazard function with a practical example, but for now I also wanted to remind you about another important statistic in survival analysis, the survival function. The survival function is the probability that a randomly selected participant will survive past an interval j. That is, it's basically the probability that the participant will not experience the event within a time interval j. And the key distinction between the hazard and the survival function is that the hazard function represents risk that is uniquely associated with an interval whereas the survival function cumulates information about the investor risk that is the non-occurrence of events over time intervals. But since the hazard is an indicator of occurrence of an event and the survival function is an indicator of event non-occurrence they are obviously linked. And in the example of the study of first sexual intercourse, we can see that the probability of non-occurrence in a discrete time period like a year is calculated as the inverse of the hazard function in that interval multiplied by the survival function of the previous period so that we take into account the reduction in the risk set in previous intervals. Well, basically that means that we can estimate the survival function by calculating the product of the inverse hazard function across intervals that preceded the interval of interest.

 So, that example showed how we can calculate the survival function when we are considering discrete time, but when we consider continuous time, we use the same logic. Once we break continuous time into small intervals rather than instants, we can then use the same logic to calculate the survival function. So, here the survival function at time interval tj, so tj represents specific time interval j, the survival function at time tj is the probability of surviving past this interval tj and we can obtain this probability by multiplying the successive probabilities of surviving through each interval from the interval one to interval tj. And these probabilities are just the complement of the conditional probability of experiencing the event during the interval, here indicated as p̂, so that we can express basically the survival function with the formula here that indicates that basically we can estimate the survival function by multiplying the inverse of the conditional probability of experiencing the event across all the intervals that preceded the interval of interest.

 Right, having said all that, I now provide a practical example. I've been using the data set called “lung” from the survival library in R, so this is the data set that results from a study of lung cancer patients and the event of interest was death of patients measured in days from the start of the study. There were other variables that represented, for example, the sex of the patients, the age of the patients at the start of the study and so on.

 So, using this example, the target event is death of the patients and the occurrence of the event. As I said, it’s calculated there in days. However, in order to summarise the data, we might use a different time interval. In this example I could have used months and report the hazard and survival function of patients every month, but here I used trimesters, so periods of three months. Obviously, when you are reporting live tables of events that are measured in continuous time you need to find a way to represent the statistics, you need to find an interval that can represent the statistics in a meaningful way, but also ideally we want to find intervals that are large enough to produce stable estimates and estimates that are not too noisy. But also, we want to identify intervals that are fine enough that can reveal patterns of interest. So, it's a difficult balance and you may need to try different ways of categorising time into different intervals.

 Here note that I also collapsed the last three trimesters into a single interval, so it's possible to summarise data for intervals that have different length. In this case I did that because the risk set in the last three trimesters was just five participants and only one did experience the event within these three trimesters.

 So, here the key statistics I report are the p̂ which represents the estimated conditional probability of event occurrence within the interval, the ĥ is the estimated hazard function and the ŝ the estimated survival function. Note that the p̂ and the ĥ are identical in all the rows save for the last one row because the unit of time here is trimester. So, hazard function is calculated as the conditional probability divided by the interval width, as I said before, but since the unit of time is trimester, the interval width is one in all the rows save for the last one where the interval width is three trimesters.

 And here also I reported that actuarial hazard and survival estimate which are calculated in a similar way as the hazard and survival function estimates with the difference that those estimates, the actuarial estimates, introduce some correction that reflects the assumption that all events and censoring are distributed equally throughout the time intervals of interest. So, this is analogous to assuming that events and censoring take place at random during the intervals of interest.

 So, these actuarial estimates partly adjust for the categorisation of time in similar tables. You can find more information in the references I provided, but in general those different estimates provide similar estimates so there isn't, as you can see here, a large difference usually between those estimates.

 And here we can see that the risk of death seems to be relatively stable to then increase over trimester eight and nine. So, around the end of the second year. However, bear in mind that the risk set is a lot smaller in the later intervals. So, the estimates also may tend to be more erratic. But those estimates are calculated in ways I described and with the resources that I provide with the material that I provide with these resources. You can also check, there are scripts I've used to calculate these estimates so you can see how I have done those calculations, how I have estimated those statistics.

 I will now introduce the Kaplan-Meier method to analyse continuous time event occurrence. This is also known as the product limit method, but one of the advantages it also has maximum likelihood properties and that's another one of the reasons also why it has become quite popular. The key intuition in the Kaplan-Meier method is extending the discrete time method, but rather than rounding event times to create preset intervals, for example, periods of five minutes, the Kaplan-Meier method uses the raw event times to create intervals whereby each interval contains one absurd event occurrence.

 So, in other words, and I try to represent it here graphically in this line that is breaking down by the event of occurrence, the red dots, as you can see here the red dots basically determine where an interval stops and a new one is starting. So, in other words, the initial interval will be between time zero and time one when the first event takes place. The second interval of interest will be between time one and time two when the second event takes place and so on.

 And conventionally, the first interval ends just before the occurrence of the first event and as soon as the event takes place, a new interval starts which ends as soon as the other event, the next event, takes place and so on.

 So, the last interval then ends as soon as the last event takes place or it ends at infinity if the last recorded time corresponds to censoring, to a participant that is censored.

 So now given that we have broken down time into intervals defined by event occurrence, we can compute the conditional probability of event occurrence in the same way I illustrated before, and then multiply the complements of these probabilities to obtain the Kaplan-Meier estimates of the survival function. So, we just use the same methods I illustrated, but applied to time that has been divided into interval defined by event occurrence itself.

 And here I provide some of the formulae. So, the conditional probability of event occurrence is defined as the number of participants who experienced the event within an interval divided by the number of participants who had not yet experienced the event at the start of that interval. The survival function is defined as the cumulative product of the inverse conditional probability of previous intervals up to the interval of interest. And the Kaplan-Meier estimate of the hazard function will thus be obtained by dividing the conditional probability of event occurrence in an interval by the width of that interval. And remember, in this case, as I tried to represent graphically here, the width of the intervals will vary depending on event occurrence.

 So, here are also provided examples of how to calculate the conditional probability of event occurrence in interval one and in interval two and then the general formula considering a generic interval tj and similarly for the survival function and genetic interval tj and so on.

 And here I provide a worked example again considering the lung cancer data set. And again, you can check the data and the operations I have run by looking at the material attached with these results and, as I mentioned before, before I also provide the R scripts I have used to create those to estimate those statistics and create graphs that I'm going to show you.

 So, if we look at the interval between the first and the second event occurrence, so the second interval here after the initial time when no event had taken place, the second interval is between days five and 11. Note that on day 11, three individuals died, so three individuals experienced the event for the first time. So, we have a tie, basically, an instance where more than one participant experienced the event at that precise time.

 Now the conditional probability of event occurrence within the second interval between five and 11 days is one, so we only have one individual that experienced the event, divided by the risk set which was the full sample of 228 participants that were in the study since there was not yet any censoring. The conditional probability of death within 11 days and 11 and 12 days is three since three participants died on day 11 divided by the remaining risk set that is 227 and so on.

 So, using these conditional probabilities, we can then calculate the survival function for each interval that is the probability of a random individual in the data set that a random individual in the data set will survive past the interval. So, for example, the probability of surviving past 11 days is 98%.

 And finally, considering the hazard function, we just divide the conditional probability of event occurrence for each interval by the width of the interval and in this way we obtain the hazard function for each interval which represents the death rate per day within that interval.

 And as you probably realise already, the hazard functions calculated in this way tends inevitably to be very erratic because they depend on the intervals with. And for this reason, they are not usually reported by most software. Most software don't provide these hazard functions because they are quite erratic and difficult to interpret.

 What is usually reported instead is the cumulative hazard function. Here the cumulative hazard function is indicated by the capital H rather than a lower case h, and then the indicator for the time interval of interest. So, capital H followed by tj indicates their cumulative hazard function for a generic time interval tj.

 The cumulative hazard function totals the virtually infinite values of the hazard function that exists between the start of time, so t with zero, and the interval of interest tj. But since it cumulates the estimated risk of event occurrence, it does not provide information about unique risk at a specific time or a specific interval rather as the hazard function does. So, it's a useful function, but it doesn't represent a unique hazard within a time interval.

 And also, because the cumulative hazard function totals hazard, it has a value that has little meaning on its own. The hazard represents probability when we are talking about discrete time. So, it has a metric that is more intuitive whereas the cumulative hazard function doesn't have a metric that is very useful. However, it is useful in providing insights, so the cumulative hazard function is reported and used because it does provide some insight.

 And there are two ways to calculate the cumulative hazard function that I will now illustrate.

 The first and quite commonly used way to calculate the cumulative hazard function is the Nelson-Aalen method and this method basically sums up the product of the hazard function and the interval width across all intervals from the first interval to the one of interest, here indicated as interval j. The negative log survivor function instead uses the link between the hazard and the survival function and calculates the negative log of the Kaplan-Meier survival function to obtain an estimate of commutative hazard. Here I also want to specify that when I use log and when I talk about logs I am referring to natural logarithms. So, logarithm with base ten are instead indicated as Log 10, so when I use log, I am referring to logarithms with a natural base. I use the same, basically, I'm using the same notation that is used in software like STARTER or R.

 So, having said that, the two methods of calculating the cumulative hazard function are usually quite similar, but they tend to diverge as the risk set decreases and the Nelson-Aalen method is more susceptive to censoring, so some people prefer to report the negative log of the survival function, but the Nelson-Aalen is also very popular.

 And here just to explain it better and in a more concrete way, I provide an example of how to calculate the cumulative hazard function using that Nelson-Aalen method. So, you can see the right side of the slide for the example. You may notice here that since the hazard function is obtained by dividing the interval conditional probability by the interval wave, multiplying the conditional probability by the interval wave just returns the conditional probability so that here the cumulative hazard function is equivalent to the sum of the conditional probabilities before the interval of interest.

 Regardless of the methods used to calculate the cumulative hazard function, cumulative hazard functions have a common problem. That is, they draw attention to the upper right tails, and here I provide an example taken from the lung cancer data set. But the upper right tails usually are estimates that tend to be very unstable. And here I try to highlight this problem by drawing also the cumulative hazard function with confidence intervals. The line here represents the cumulative hazard function for the lung cancer data and the vertical ticks on the line represents censored cases. But the area that represented here indicates the 95% confidence interval of the cumulative hazard function, and here you can see that the confidence intervals indicate greater uncertainty in the estimates on the right side of the graph as more people die or leave the study and are therefore censored, the cumulative hazard function tends to be based on a smaller reset risk set so the estimates become more unstable and more erratic. It is therefore advisable to focus on earlier and more stable sample estimates than the ones that are estimated towards the end of the study.

 But in a nutshell, the cumulative hazard function is particularly useful in identifying periods where the rate of change in risk changes so that we can identify periods where the average slope, so to speak, of the hazard function is changing.

 So, one last thing before I close this presentation is that I mentioned that the hazard function estimates are too unstable and in fact they are not usually reported by software. But they are the only statistics that can provide estimates of the unique risk within specific time intervals.

 So, they are providing some information that is otherwise not that easy to gather from other statistics. So, to still use the hazard function estimates and get some insights into the phenomena of interest, we can rely on Kernel-Smoothed estimates. So, these are basically estimates of the average of a function, in this case the hazard function, that are estimated by aggregating together all the point estimates that are within the temporal vicinity of a focal time point. So, in other words, they provide an approximate value based on estimates that are near the focal point of interest.

 But as you can imagine, the shape of this average estimates, the smoothed estimates, depend on what we define as the vicinity of the focal point of interest, and this is called the bandwidth, which is basically the interval that will provide the point estimates that are being aggregated in the smoothing.

 So, we can choose different bandwidths for smoothed estimates. Here in the graph, I provide an example where I had used three different bandwidth to report the hazard function in the lung cancer data. So, the first bandwidth is 30 days. So, for example, if we take a focal point of 35 days, the estimates are aggregating all point estimates of risk of death between five and 70 days in order to create the smoothed estimate.

 The other two bandwidths here are 90 days and 180 days. So, as you can see, the narrower the bandwidth then the more erratic are the estimates. You see, for example, the estimates for 30 days are more erratic, whereas the estimates are smoother as the bandwidth becomes larger.

 But while we may like the smoother lines, we also need to consider that widening the bandwidth introduces more bias. If we take the estimated hazard of 35 days, in the first 30 day bandwidth this value represents the average population hazard between five days and 70 days. But when the bandwidth is 90 days, the estimated hazard at 35 days represents the average population value between zero and 120 days. So, while the estimates are smoother and less erratic when we enlarge the bandwidth, we risk losing the meaning of the individual values so we may have better looking lines, but we may lose key information.

 And one final point before I close this presentation is that the different statistics I've mentioned so far provide different types of information so it is important to consider them together to get some insight into the phenomenon of interest. The data set that I've used to create those graphs here is still the lung cancer data set and usually we start with the Kaplan-Meier survival function, which is easier to interpret because it has a more intuitive metric. Then we can usually add the cumulative hazard functions and then the smoothed hazard functions that can then provide information about periods of higher and lower risk.

 So, really the point here is that it's important to explore and consider different types of statistics when we are trying to get some insight into the phenomena of interest when we are exploring survival data.

 So, to summarise what I have said so far, analysing event occurrence when time is measured continuously causes some challenges that can be tackled by defining small but informative time intervals for data analysis. So, this effectively means that the time metric in analysis can differ actually from the metric of data collection. We may record some events like infants’ gaze movements in milliseconds thanks to the advent of eye tracking technology, but often the time units in the analysis are seconds.

 I then introduced the Kaplan-Meier method that allows to provide maximum likelihood estimates of the probability of a randomly selected participant surviving past the interval of interest and the method also allows to calculate the hazard function which represents a rate that is the probability of event on onset per unit of time. But since the estimates tend to be too erratic because of the variability in time intervals, it is common practice to inspect the smoothed hazard functions with particular care in the choice of bandwidth used in the smoothing process.

 And finally, the cumulative hazard function provides insights into the rate of change in hazard which can then be useful in providing better understanding and better intuition.

 All these indicators are very important together so researchers should inspect them to develop interpretations concerning the phenomenon on the study.

 So, thank you for your attention and please check the webpage of the National Centre for Research Methods for other resources and also don't forget that with this presentation there is other material attached including exercises that you can run using our software and some references.

 Thank you very much.

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK