# Introduction to Survival Analysis – part2

## Transcript

Video: <https://www.youtube.com/watch?v=QcTB8Qp3rqU>

Full resource: <https://www.ncrm.ac.uk/resources/online/all/?id=20850>

Hello and welcome. I'm Oliver Perra, and I'm going to provide a second presentation where I will introduce Survival Analysis. This is the second or three presentations and with the resources, we will find more exercises and material on Survival Analysis.

In the first presentation I have talked about the type of questions that Survival Analysis can answer, and I said that Survival Analysis can answer questions about wherever and when an event of interest takes place. I've also mentioned the prerequisites for Survival Analysis. The first one is to define with clarity what is the event and what is the event occurrence we're interested in. That implies defining different states that the event can assume and individuals are supposed to transition from one state to another, for example, from being alive to being dead.

The second prerequisite, it's identifying the beginning of time. So, the starting point of the study, where all the individuals in the study have not yet experienced the event but are eligible to do so.

And finally, it's important to specify the time metric of the study in what units of time we're going to use and those should be the smallest possible units that are relevant to the event of interest.

In this presentation, I will first talk about censoring and why is it a key challenge in Survival Analysis? I will then introduce a way to summarise survival statistics using live tables. This will allow me to introduce the hazard function and survivor function. I want to highlight that in this and the next presentation I will only talk about examples of discrete time events, events that are measured in coarse time units like years. I do this because in my experience it’s the best way to understand the concepts and principles of Survival Analysis. Once these concepts are developed, I find it is easier to understand the more sophisticated methods that are used to describe and analyse continuous time event occurrence, for example, the Kaplan-Meier method or Cox regression models.

Survival Analysis provides a solution to a key challenge when we analyse data on event occurrence. In the first presentation I provided this fiction example and I will use it again to illustrate censoring. So, suppose we are interested in investigating whether and when young people experience suicidal thoughts for the first time ever. Assume we ran the study from adolescence where 14 years of age and follow them up until they were 32 years. As it happens, Participant 5 dropped out from the study before reporting the target event, before reporting any suicidal thoughts, whereas Participant 4 finished the study but never reported any suicidal thoughts.

So, in those two cases, we don't know the timing of the event. And when we don't know the timing of the event for participants, we talk about censoring. But these examples illustrate the reasons why censoring happens. Either the participants never experience the event of interest during the study, as Participant 4 did, or some participants leave the study before, drop out from the study before we can recall the event, like Participant 5.

There may be studies where there is no censoring. For example, if we were investigating life expectancy of a group of individuals, provided that we study those individuals long enough until they eventually pass away and none of those individuals drop out from the study, then we will not have censored data. But censoring is usually present in studies of event occurrence.

And the problem with censoring is very similar to the problem of missing data in other types of analysis. In some instances, we can be confident that censoring is not linked to the target event occurrence. In these cases, we say the censor is non-informative. It is basically independent from the target event taking place. For example, Participant 4 here is censored by design since we finished the study at that point in time, the decision to end the studied at age 32 is independent from the event we are interested in. But take Participant 5, however. If the participant left the study independently from the target event, we would also say that in this case censoring was not informative. For example, the participant may have left the study because they emigrated. So again, in these cases we can be confident that censoring was non-informative and we can assume then that all participants that remained a study after censoring are representative of participants that would have remained in the study had censoring not occurred. But if we didn't know why Participant 5 is censored, we may worry that the reason that led to the participant leaving the study were linked to the target event. So, for example, if the participants experienced suicidal thoughts and those thoughts were so severe that the participant couldn't take place in the study any longer, then leaving the study is directly linked with event occurrence, but we didn't record the event of occurrence, we don't have information about the event occurrence. So, in this case we say that censoring is informative and that is a threat to the validity of the study.

In the remainder of this and the next presentation, I will use the data set to illustrate the methods I'm going to present. This data set comes from a study published in 1996 in child development. Male heterosexual adolescents were followed up from Grade Year 7 to Grade Year 12, so roughly from 12 to 13 years to 17, 18 years of age. The beginning of time then was Grade 7, whereas the target event was in adolescents reporting that they had started having sexual relationships within a certain year. There were also some covariates which I will describe in the next presentation.

I have modified this data set though to introduce some non-informative censoring. In fact, I have censored some data at random just for illustrating some of the problems with censoring. The data set, and there are scripts to run the examples that I am going to illustrate are available with the course materials, so you're welcome to have a look at them.

I wanted to draw your attention to the structure of the data here. You will notice there isn't a variable that reports whether the event took place or not, but we can inform this information by looking at the combination of time and sense of variable. For example, in case one, Participant 1, the participant is not censored, so we can infer that the event took place in Year 9.

But if we look at Participant 2, we can see that the case is censored, so the participant left the study without the event ever taking place. In the study, the researchers stopped following up participants after Grade 12, so the censoring is not informative, is not related to the target event taking place. The data were censored by design.

So, I'm going to introduce live tables because they provide the key summary statistics regarding event of occurrence over time. So, in the roles of these tables, we have the time intervals, for example, Grade 7 to 8, Grade 8 to 9 and so on. It's important to notice that the intervals in the rows include initial time of the interval. So, when we consider Year 7, that interval includes the start of Year 7, but the intervals exclude the concluding time. So, the interval for Year 7 excludes the start of Year 8.

The n risk column reports the number of participants that were in the study and had not yet experienced the target event at the beginning of that specific interval. So, for example, at the beginning of Interval 7, Year 7, the beginning of time, all participants had not yet had sexual intercourse by design. The n risk means that this is the number of participants at risk of experiencing the event for the first time. The n event reports the number of participants who experience the event during that interval. For example, 15 participants had reported initiating sexual intercourse for the first time between Grade 7 and Grade 8.

The n censor is the number censored, so those are the participants who left the study before the targeted event was ever recorded. Note that the number at risk in every interval is the number of participants in the study minus the sum of those that had experienced the event in the previous interval and those censored in the previous interval. So, for example, the risk set in Year 10 is 158 minus 18 and minus 6, so 134.

The column labelled hazard here reports the hazard function. The hazard function is defined as the conditional probability that an individual will experience the event during a specific time interval provided, this is the conditional bit, provided that the participant did not experience the event in previous time intervals.

So here I have reported a formal definition of the hazard function. So, it’s h(tij) where i indicates an individual in the study, a participant, and j indicates a time period. T with i is the time period j when the individual experiences the target event. The hazard function for individual i in interval j then is the probability that the event will occur in the current time period given that it must occur now or sometime in the future. So, in a way this definition specifies the converse of the local definition. Since the target event has not yet occurred, the formula expresses the probability of the event in the current time interval provided that it should occur in the current or future time intervals.

The point is that the hazard function tells us whether and when the target event is more likely to take place, so provides the key information for Survival Analysis. Since these probabilities are conditional on the participant not experiencing the event before the time interval of interest, the hazard function represents the probability of event occurrence among those participants that are still eligible to experience the event in that interval. That is excluding those that had already experienced the event or have left the study.

So, this consideration emphasises the importance of non-informative censoring. To be able to have informative probabilities, we must be able to assume that those 134 that entered the study in Year 10, for example, are representative of all the cases that would have entered the study in Year 10, had there not been any censoring.

And to illustrate how the hazard function is calculated, let's consider a time interval. In Year 9, for example, the risk set is 158 and 18 participants experience the event for the first time. So, the probability of the event taking place in Year 9 is 18/158, so 0.114. It's important to note that these proportions are maximum likelihood estimates of the discrete time hazard function. And also, those are the discrete limits of the Kaplan-Meier hazard estimates for continuous time which you will encounter when you engage with Survival Analysis beyond my introductory presentation. So, it's important to remember that those are the discrete limits of a function that is more general and more generally applicable to discrete time event occurrence.

And here also plotted hazard function as a line that links the estimates for the different time periods of interest. Since hazard is a probability, it will vary between zero and one. And in this example we can see that the probability of initiating sexual intercourse among male adolescents in the study was low in the first two years, but increased and was higher in Year 12 when the study finished.

Overall, the hazard function then estimates the unique risk of the event within a time interval. The survival function instead accumulates the estimate of risk over time periods and it estimates the probability that a randomly selected individual will survive past that specific time interval. So, the probability that the individual will not experience the target event within that time interval or before.

While at the start of the study all participants are survivors by definition, they have not yet experienced the event. By the end of the first year, for example, here the probability of surviving is approximately 92%. By the end of Year 8, it’s 88%. And since this is a cumulative estimate, the probability of surviving will never increase over time. But to understand how this function is calculated, let's consider the first two intervals where there is no censoring. Generally, the formula considers those who have not yet experienced the target event within are certain interval divided by the number of participants. So, at the end of Year 8, 158 participants have not yet experienced sexual intercourse. In fact, they are in the reset at the beginning of Year 9, so we divide 158/180 participants in the sample and this tells us that the probability of surviving of not initiating sexual intercourse within Year 8 is approximately 88%.

However, when we have censoring in the data, we do not know how many adolescents did not experience the target event by the end of those time intervals. But we can get around this problem considering that hazard function tells us about the probability of event occurrence within a time interval. So conversely, also tells us about the probability of the event not taking place before the survival function can be calculated based on the estimated hazard function.

And for example, if in the first interval the probability of experiencing the event is 0.083, we subtract this from one to obtain the probability of surviving during the first time interval. So, we obtain the survival function directly from the hazard function.

In Year 8, the hazard function is 0.042, so we can subtract this from one to learn the probability of the event not taking place within that interval. But we then need to consider that only 0.917 of the original sample where in the risk set at the start of Year 8. So, to estimate the cumulative probability of surviving past Year 8, we need to multiply the probability of event non-occurrence by the proportion of participants that are still at risk. So, we subtract one, we subtract 0.042 from one and multiply it by 0.917 and we obtain the survival function for that interval.

Here I also include the formal expression of the survival function and basically the same logic that I've used with time eight can apply when there is censoring. For example, the survival function in Year 12 is 1 minus 0.262, the hazard function multiplied by 0.501. So, the reset in the previous time interval.

So, the formal definition indicates that the survival function for interval j is the product of the survival function in the previous interval j minus one, and one minus the estimated hazard function for period j.

And this survival function is also useful in providing an estimate of central tendency that median lifetime. The median lifetime is the time value where the estimated survival function is 0.50 or, in other words, is the point where half of the sample has experienced the target event. In cases like this one where the time measure is discrete, the methods used to estimate the median lifetime is interpolation and you can find details in references and the supporting material for this resource.

So here the median lifetime is 11.01. So, by the start of Year 11, roughly 16 to 17 years of age, half of adolescents in this sample we have initiated sexual intercourse, while half of them have not yet done so.

And finally, both functions, plotting them together, is also useful in providing insights into the phenomenon of interest. As I discussed, the two functions are linked and this link can be seen by noting that in the first two grades of the study, the hazard function is low, and in the same year, the survival function remains relatively shallow. But after Year 8, the hazard function increases and we note a relatively steeper decline in the survival function. So, it is useful to use these two plots in combination, particularly because hazard functions can vary a lot and a lot more across time and it may be difficult to recognise a pattern. So, looking at those functions together can provide some understanding of the processes in the data.

So, to summarise, I have talked about censoring and it's important to remember that most analyses assume that censoring is happening because of non-informative mechanisms. If there is a suspicion that the mechanism of underlying censoring is informative, then more specific sophisticated type of models should be employed. I have illustrated and described life tables which summarise events’ occurrence over time and I have introduced the hazard function which is the conditional probability of an individual experiencing the event within the interval given that they did not experience it before. And then I have illustrated and talked about the survival function, which is the probability of surviving past a time interval for randomly selected individuals.

So, thank you very much for your attention and you can find more resources in the webpage of the National Centre for Research Methods together with exercises and other references.

Thank you. Bye.

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