# Introduction to Survival Analysis - part1

## Transcript

Video: <https://youtu.be/PKrlbjWjWVg>

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

Hello. I'm Oliver Perra, and welcome to these resources where I'm going to introduce Survival Analysis. This is the first of three presentations.

 In this presentation, I'm going to introduce the type of questions that can be answered by using Survival Analysis. Then we'll highlight that common statistical methods we use every day are inadequate to answer the type of questions that Survival Analysis can answer. And finally, I will highlight the prerequisite for running Survival Analysis, particularly focusing on the type of information, the data and the decisions we need to take ensure that we can meaningfully answer resource questions using Survival Analysis. I will then focus on the concepts and principles underlying Survival Analysis without presenting models and equations, which I will introduce in the other presentations among these resources.

 Researchers in very different fields from economics to medical sciences or from psychology to engineering are often interested in investigating the occurrence of events. So, when does a specific event take place? And for example, we may be interested in studying the onset of puberty. When do the physical and hormonal changes that lead to sexual maturity start to appear in a population of young people? Or another example in the field of psychology or psychiatry involves studying the first instance of thoughts about suicide that individuals experience in their life. A final example involves studying when individuals get a full-time employment after leaving education, if they ever get a full-time employment.

 So, what do these examples all have in common? What is the common thread in these examples? In all these examples have a common thread that is that the fact that researchers are interested in investigating two key questions, whether the event of interest takes place and when does the event take place. For example, within a large population, psychiatrists may find that only a relatively small proportion of individuals have experienced suicidal thoughts, so they might be interested in investigating the factors that are associated with their occurrence of this event associated with the first onset of suicidal thoughts. Within those that experience suicidal thoughts, there may be variability in the onset of these thoughts where some people experience those for the first time in adolescence, others may experience them later in life and many will not experience them at all. So, there might be differences in the timing of the first ever suicidal thought between gender or there may be other characteristics that are significantly associated with an earlier onset of the first ever suicidal thought. The point is that across different disciplines, we are interested often in studying wherever an event took place, and if it did take place, when did it take place? And after these two key questions, we may also start to ask questions that are more specific, for example,. are women more likely to display an earlier onset of suicidal thoughts than men, and so on.

 Well, how can we investigate whether and when questions, questions that concern whether and when a target event takes place? The common statistical methods that we use are inadequate to deal with these types of data. Let's look at this fiction example I have created to get an idea. Here the topic is whether and when participants first experience suicidal thoughts. In this fictional data, there are five participants. All of them were enrolled in the study when they were aged 14 years and now if we look at this data, Participants 1, 2 and 3 in the bottom of the picture experience suicidal thoughts. Participant 3 experienced these thoughts for the first time at an early age of 14, Participant 2 at a later age when they were aged 25 years. But then we have two more participants, Participant 4 and 5 in the top of this graph who never experienced suicidal thoughts while they were in the study. The problem, which is not uncommon, is that Participant 5 left the study at age 20 while Participant 4 left the study at age 31. So, we don't know whether suicidal thoughts ever occurred in Participant 4 and 5 and let alone we know when did they occur, if ever.

 So how can we, for example, estimate the medium age of onset of suicidal thoughts? We could consider only those that did experience suicidal thoughts among the three people here that experienced suicidal thoughts. The average age of onset would be 19 years, but this would be discarding information from the other participants that never experienced suicidal thoughts and left the study. So, this is not an adequate summary of our data.

 We could come up with some arbitrary rule that says where, for example, we may assume that participants that left the study experience suicidal thoughts after they left the study. So, the age of onset for the two participants that left the study before we could record the onset of suicidal thoughts would be 32 and 21 years for Participant 4 and 5 respectively. But in this scenario, the average age of onset would be 22 years, but these are some just are not just tenable. Many people never experience suicidal thoughts in their life.

 So, the point really is that in this type of data, where we have people leaving a study before we can record the onset of the event, we don't have straightforward ways to describe the age of onset or the timing of onset of the events we are interested in.

 And that is where Survival Analysis comes to help. Survival Analysis is a set of statistical methods that provides answers to research questions that concern whether and when events take place. And these methods are really quite flexible. So, for example, they can be applied to different types of designs, different study designs, such as experimental studies or randomised controlled trials, longitudinal studies, whether the data are being collected prospectively or have been collected retrospectively.

 Those methods, Survival Analysis, can be applied to events that only happen once, for example, puberty onset or can be applied to events that are recurring. For example, suicidal thoughts can reoccur over some people's lifetime and we can investigate the reoccurrence of these events. Furthermore, Survival Analysis can be applied to negative events like death, onset of disease, relapse, but can equally be applied to positive events, so to speak, for example, promotions at work, the timing of an infant's first words, and so on.

 The Survival Analysis label is there because these methods were developed to investigate the timing of death of patients with different conditions, but these methods can effectively be applied to different topics and the main requirement is that the research questions concern whether and when a target event takes place.

 In the remainder of this presentation, I'm going to focus on the prerequisites for running Survival Analysis. So, what type of data and what type of decisions we need to make in order to run Survival Analysis?

 And the first prerequisite is that Survival Analysis should have a clearly defined target event. And this means that the target event should be defined as a transition from one state to another state. In the example I had used before, individuals were in state one if they had never experienced any thought about suicide but transitioned into a second state as soon as they had thought about suicide for the first time.

 Or another example is ever been married, people that have never been married are in state one until they marry and therefore move into state two. From these examples I should be clear that the states that make up a target event in Survival Analysis should be mutually exclusive and exhaustive. That means that no individual can be in more than one state at any given time. Either and individual had never had suicidal thoughts or had experienced suicidal thoughts before, or an individual either has been married or is married currently, or has never been married, so an individual would be either or the other, either in one state or the other state.

 Exhaustive states mean that all the individuals should be in one or the other of the possible states. All individuals will have either experienced suicidal thoughts or never experienced them.

 Survival Analysis is often applied to the dichotomous target events where there are only two possible states that are alive, never married versus ever married. However, these methods can also be applied to events that can have more than two states. For example, a person may be employed, self-employed or unemployed or retired. The point is that all these states should be exhaustive, all individuals should be in one or another of these categories and the states should be mutually exclusive. An individual will be in only one category at any given time. Someone employed cannot be unemployed at the same time, for example.

Also, transitions into one state can take place only once. For example, the transition from alive to dead is irreversible. But in other cases, individuals may transition between states several times. For example, if we were interested in the occurrence of suicidal thoughts over time rather than the fast onset of suicidal thoughts, then suicidal thoughts can occur several times in the lifetime of some individuals or, for example, some individuals may move from being unemployed to being employed several times. The key point is that at any given time, individuals should be in one or another state and they all should be in one of the event states possible.

 Another requisite for Survival Analysis is identifying a point that starts the clock in the study, a so-called beginning of time. In other words, the researchers must identify a point where all the individuals in the population we are studying are in the same state of the target event. For example, when we are studying life expectancy in a population, all live births start life in one state being alive and from that point on they can transition into the alternative state of being dead. So, in similar studies, birth is the beginning of time. The beginning of time is therefore the point where all the individuals in the study have not yet experienced the target event, but they are potentially at risk of experiencing it.

 In some studies, the beginning of time is the same for all participants in the sense the clock starts at the same time for all participants. For example, in studies that consider age as the time variable, birth is the beginning of time for all participants and is the time where all participants have not yet experienced the event but are eligible to experience it as the example I made before is that as soon as a live birth is born, they are eligible to experience the detrimental event of dying.

 In other studies, the beginning of time may be linked to a precipitating event. For example, if we were interested in whether and when someone is promoted in their job, the beginning of time may be the time when an individual starts in the job. So, this will be different across participants and in this fictional example on the right, for example, participants who started a job at 18 years, Participant 1 started their job at 25 years, the time variable in this study is not age, but rather time in employment.

 So, in many studies, the start time may also be arbitrary or detected by convenience. And this is fine, but it is important and it's not a problem for Survival Analysis, but it is important to remember that the starting time in a study must be unrelated to event occurrence. It must be not related or linked to the target event we are studying. It should be independent of the target event.

 It is then important to specify a metric for time. What are the units of time the researchers will use to record the passing of time? Many events of interest can potentially be recorded using a very precise time measurement. For example, if we conducted a diary study of suicidal thoughts and asked participants to record the onset of suicidal thoughts using an app, we could potentially record the occurrence of suicidal thoughts in minutes, so we will know in what day and what hour and minute of the day did a participant have their first suicidal thoughts?

 Another example is that a paediologist that record lifetime use death certificates that can provide measures of lifetime in days rather than months or years so that measurements can be quite precise. However, there are often limits to the precision of time measurements, due often to practical problems. Some of these problems may be linked with the way data are recorded, especially when researchers use administrative data or data collected by others. For example, hospital records may record the occurrence of some interventions in weeks rather than days.

 Other limitations may be due to the nature of the events researchers are interested in. For example, if we are interested in the timing of academic promotions, promotions are usually awarded yearly, so we will necessarily use years as the matrix, not months or days. Other limitations may be due to the study design and particularly when we collect data retrospectively and ask participants to recall when did a specific event take place, unless it was a very salient event, participants may remember the month when a certain event took place or maybe the season or even the year. So, difficulties in recording the specific and precise timing of an event may force researchers to use a coarser time unit.

 The main point, however, is that time should be recorded in the smallest possible unit that is relevant to the event of interest that is enabling the researchers to detect the onset of the event of interest. For example, when we study the timing of a child's first words, measuring this in years would be a very coarse metric. It will not be very useful to know that most children say their first words in their second year of life rather than the first one, because in the first words spoken by a child in months is much more adequate and is going to provide information about individual differences and so on. So, there are two main types of time measurements. Continuous as when we can measure lifetime in days or discrete as when we measure the onset of suicidal thoughts in years.

 But this distinction is also important for a further reason. When we measure time continuously, the probability of two individuals experiencing the target event at the same time will be low or very low. Here in the example on the left, the red dots represent the timing of target events for different participants. When two individuals experience the target event at the same time, this is called a tie. And when we measure time continuously, ties are less likely and we will observe few of these ties. But when time is measured in discrete and more coarse units, ties will be inevitable.

 But this is important because modern methods in Survival Analysis such as Cox regression, also known as the proportional hazard modelling, assume that time is recorded continuously and therefore ties should be far and few between.

 Applying these methods, more modern methods, to data that have many ties then means really stretching the models to a point where their inference may be unreliable. So, this is an important point to remember if we are using more sophisticated models that have some assumptions that cannot be tenable, cannot be satisfied.

 So, to summarise, I've highlighted that Survival Analysis is a statistical method that can answer questions that concern whether and when an event of interest takes place and I've highlighted the prerequisites for Survival Analysis. We first need to define event occurrence defining the event as a transition from one state to another. There may be more than two states, but at a minimum we need to have a definition of two mutually exclusive and exhaustive states and be able to observe how individuals transition from one state to another.

 We also then need to identify the beginning of time, so we need to identify when we need to start a study and we need to in particular make sure that the starting point of a study should not be related to the target event of interest.

 And then we need to specify a metric for measuring time. This can be quite precise, minutes, even seconds, or it may be coarser, years or quinquennials and so on. The point is that we need to consider the smallest possible unit that is relevant to the events of interest. The smallest unit that can provide information that can differentiate different individuals and the processes that are relevant in the onset of the events.

 So, thank you very much for your attention and don't forget to look at the webpage of the National Centre for Research Methods for other resources. And you will find other presentations on Survival Analysis in these resources including exercises. So, thank you very much again. 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