**Survival Analysis outline**

Accommodates the survival analysis script in R

Today we cover Survival Analysis. Survival analysis is a branch of statistics that deals with the analysis of time-to-event data. The goal is to model and predict the time until an event of interest occurs, such as death or disease progression. Today we cover the basics of survival analysis and provide you some code in R.

We load some packages we will be using:

**Surv package**: Provides basic survival analysis functions, such as survival curves and Cox proportional hazards models.

**Flexsurv**: Offers more flexible survival modeling options, including a wide range of parametric survival distributions.

**SurvHE**: Tailored for health technology assessment (HTA), facilitating economic evaluations of health interventions.

**Darthtools:** Contains functions useful for survival analysis, particularly in health economics.

Input parameters, where we see some parameters are constant over time.

But the others are missing

Time-to-Event data

Time to event data is the fundamental data type in survival analysis, indicating the time until an event occurs.

Simples example would be an alive dead model.

However, often we don’t have only information about one transition but about two.

How would we call this?

**Competing Events**: Situations where more than one type of event can occur, such as transitioning to a sick state or death.

**Example**: if we start with 100 people being health. And we have at the end of the cycle 20 people in sick and 20 people in Dead. What is the probability of transitioning from Healthy to Sick?

Likely, to say 0.2. However, this is not entirely true, because the population at risk becomes smaller as soon as one person transitions.

*Draw on board what really happens: with both events happening*

[Question: would it make sense to here draw people getting sick, and people getting Dead and censored.

Meaning a mix of events that are happening. Or only looking at one event type]

1

2

3

4

5

6

Cycle

Competing events must be carefully accounted for, as the occurrence of one event excludes the person from being at risk of experiencing the other event. So we are not having a population of N, but of N-1

We are having this problem because we are using discrete time in our models. While in reality, time is continuous and therefore, in our model we have to deal with competing events.

One solution, is to make our cycle length as short as possible, but this comes with additional computational time.

How we are dealing with the situation is that we are splitting the data:

We say we know who experiences the event (20 transitioned to sick), and we assume the others were censored. Similarly, we do for those that died, we know who died and assume the other are censored. Therefore, we split the data. Or in fact your data came splitted.

Therefore, we have two datasets:

1. Individuals at risk of dying (df\_HS)
2. Individuals at risk of getting sick (df\_HD)

This separation allows us to model the competing risks explicitly during the modeling phase.

In our model, once an individual made a transition of one health state, they cannot transition to the other in the same cycle.

[Petros, is this reason an argument why we can assume our other people are censored? How dan we explain that this is reasonable? ]

Understanding the Dataset:

Our dataset contains individuals classified as having an event (1) or not having an event (0). As well as the time they were at risk of that event.

**Statistical Modeling -** Non-Parametric vs. Parametric Models:

First part of the code, shows a Kaplan meyer curve. And visualize our data.

Although this non-parametric model representation is while useful for certain analyses, it is not suitable for decision modeling. Why?? …

Due to their inability to extrapolate beyond observed data.

How would you do that when you have this figure? Would you continue the flat line?

And what at cycle 30? We have seen a decline before, but when and how would that happen?

And with our models, we like to extrapolate the trail data beyond the time of observation. Parametric models are solution for that. This can be done with base R, but packages, like flexsurv, offers a more flexible option. The parametric models, assume a specific distribution for the time-to-event data, allowing for extrapolation and more detailed analysis over time.

**Flexsurvreg** Function: Utilized for fitting parametric survival models, flexsurvreg assumes a particular distribution (e.g., Weibull) for the data, enabling the modeling of survival as a function of time and the extrapolation of future events.

We now have to let R know that we are dealing with survival data. The first argument is the time in our data called time. And the second is the event part (1 is an event, and 0 censored. In our data that is called status. It is possible to split by covariants like gender or treatment. But we don’t have that, and we set it to one ~1 and estimate the survival curve.

We see know what we saw before, but are now able to extrapolate for the time horizon we need in our model.

What we see, it that the rate in the beginning is somehow constant and after some cycles not one is transitioning anymore. So when you did not transition in the first years, you are likely to stay in that state but this data is not dealing with the other transition.

Sometimes, when you like to compare your data with for example data you have in Excel spreadsheets it is important to extract the survival probabilities for specific time points. With the *summary* function, you can extract the survival probabilities on the times points you specified. Which makes comparing model output with external data possible.

The good thing about this, is that now in R can can increase the “xlim” and extrapolate into the future.

In our example we tried to fit a Weibull distribution, but in fact we like to try many distribution and like to select the best one. Rather then going one by one, there is a “wrapper function” to do this using fit.models

We extract the AIC and BIC, which are measures on how well the survival model fits the observed data, but keep in mind only the observed data. It does not say anything about the extrapolated part. This you have to validate with clinical experts to discuss if an estimated survival would make sense with what they see in practice. The combination of this information should help you to decided the best fitting curve.

Right now we don’t have an expert

Caution should be exercised when interpreting survival curves, especially towards the end of the study period, where the sample size may be small, and observed trends may not hold in a larger population.