Survival Analysis

**What is Survival Analysis**: [**Survival analysis**](https://en.wikipedia.org/wiki/Survival_analysis) is used to study the **time** until some **event** of interest (often referred to as **death**) occurs [1]. Time could be measured in years, months, weeks, days, etc. The event could be anything of interest. It could be an actual death, a birth, a Pokemon Go server crash, etc.

**What will we actually want to do: (Methodology)**

In this project we are interested in how long drafted NFL players are in the league, so the event of interest will be the retirement of drafted NFL players. The duration of time leading up to the event of interest can be called the **survival time**.

Some of the players in this analysis are still active players (e.g. Aaron Rodgers, Eli Manning, etc.), so we haven't observed their retirement (the event of interest). Those players are considered **censored**. While we have some information about their career length (or survival time), we don't know the full length of their career. This specific type of censorship, one in which we do not observe end of the survival time, is called **right-censorship**. The methods developed in the field of survival analysis were created in order to deal with the issue of censored data. In this project we will use one such method, called the [Kaplan-Meier estimator](https://en.wikipedia.org/wiki/Kaplan%E2%80%93Meier_estimator) [2][3], to estimate the survival function and construct the survival curve for an NFL career.

In a word the survival function just gives us the probability that someone survives longer than (or at least as long as) a specified value of time. So in the context of our analysis, it will provide us the probability that an NFL career lasts longer than 3 years.

What is survival function?

The [survival function](https://en.wikipedia.org/wiki/Survival_function), S(t), of a population is defined as follows:

S(t) = Pr(T>t)

Capital T is a [random variable](https://www.khanacademy.org/math/probability/random-variables-topic/random-variables-prob-dist/v/random-variables) that represents a subject's survival time. In our case T represents an NFL player's career length. Lower case t represents a specific time of interest for T. In our analysis the t represents a specific number of years played. In other words the survival function just gives us the probability that someone survives longer than (or at least as long as) a specified value of time, t. So in the context of our analysis, S(3) will provide us the probability that an NFL career lasts longer than 3 years.

What is Kaplan-meier estimator?

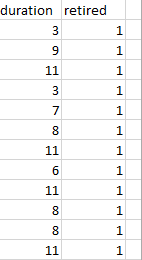
To estimate the survival function of NFL players we will use the Kaplan-Meier estimator. The Kaplan-Meier estimator is defined by the following product (from the [lifelines documentation](https://lifelines.readthedocs.io/en/latest/Intro%20to%20lifelines.html#estimating-the-survival-function-using-kaplan-meier)):

C:\Users\personal\Pictures\kmformula.PNG

Where di are the number of death events at time t and ni is the number of subjects at risk of death just prior to time t.

Data:

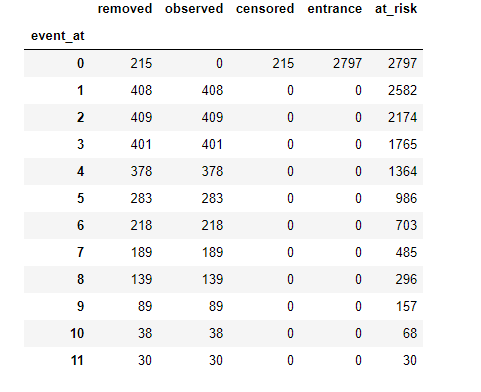
Data collection is from NFL player draft table. (For example 1 year players data is shown in excel file.)



The columns of interest for our analysis are the Duration and Retired columns. The Duration column represents the number of years a player played in the NFL. The Retired column represents whether the player retired from the NFL or not. 1 indicates that he is retired, while 0 indicates that he is still an active player.

Estimating the survival function for NFL player:

To estimate the survival function of NFL players we will be using the [lifelines library](https://lifelines.readthedocs.io/en/latest/index.html) [4]. It provides a user friendly interface for survival analysis using Python. To calculate the Kaplan-Meier estimate we will need to create a KaplanMeierFitter object. After fitting our data we can access the event table that contains a bunch of information regarding the subjects (the NFL players) at each time period.

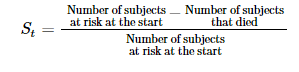


Here event\_at is the time period (t). Obsereved value is equal to retired value of subjects.

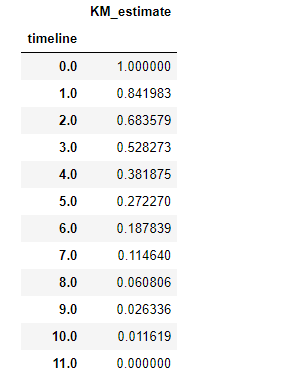
The removed column contains the number of observations removed during that time period, whether due to death (the value in the observed column) or censorship. So the removed column is just the sum of the observed and censorship columns. The entrance column tells us whether any new subjects entered the population at that time period. Since all the players we are studying start at time=0time=0 (the moment they were drafted), the entrance value is 2797 at that time and 0 for all other times.

The at\_risk column contains the number of subjects that are still alive during a given time. The value for at\_risk at time=0time=0, is just equal to the entrance value. For the remaining time periods, the at\_risk value is equal to the difference between the time previous period's at\_risk value and removed value, plus the current period's entrance value. For example for time=1, the number of subject's at risk is 2582 which is equal to 2797 (the previous at\_risk value) - 215 (the previous removed value) + 0 (the current period's entrance value).

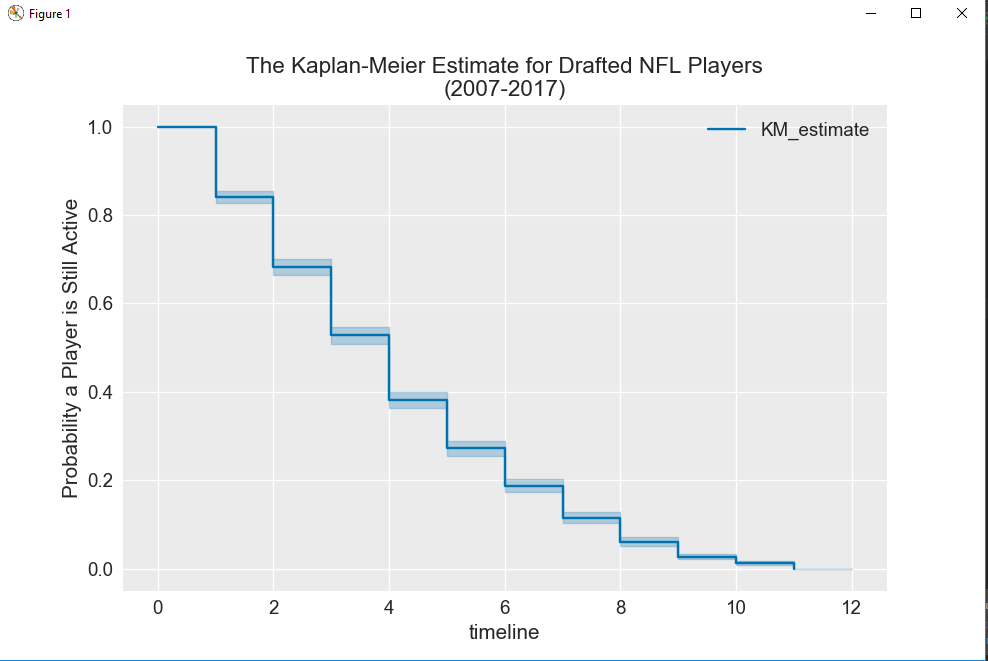
We can define the survival probability for an individual time period as follows:



KM estimation table



Plotting the Kaplan-Meier estimate (along with its confidence intervals) is pretty straight forward. All we need to do is call the plot method.



 Each horizontal line represents the probability that a player is still active after a given time t. For example, when t=0, the probability that a player is still active after that point is about 100% and when t=1, the point is about 84%.

Reference:

[1] <https://www.cscu.cornell.edu/news/statnews/stnews78.pdf>

[2] <http://www.garfield.library.upenn.edu/classics1983/A1983QS51100001.pdf>

[3]<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/>

[4]<https://lifelines.readthedocs.io/en/latest/index.html>