Survival Analysis

Alexis Destefano (she/her), working with
Ethan Ancell and Sherry Ren

Survival analysis is a branch of statistics centered on time to event data. That is, it attempts to determine the amount of time that will elapse from the beginning of a study (or other specified event) to the occurrence of a particular event of interest. For example, a researcher might want to take the time to event data of a group's first usage of illegally obtained pain pills to their first usage of heroin. In this scenario, the "time" starts at the first usage of pain pills and is measured until the "event" --the first usage of heroin. Survival analysis is particularly useful in biostatistics, but can be applied to any analysis of data measuring time numerically and focused on one outcome that ends that time measurement.

Each observation in a survival analysis dataset is comprised of certain elements, including τ_i, which indicates true "death" time. Survival analysis is predicated upon the concept of censored data, which means data for which we don't have a time at which the event of interest took place, *if* it took place–that is unknown to us as well. Data may be censored for numerous reasons, including the researcher losing contact with this particular data point, the participant representing this point of data dying before the event of interest occurs, or the participant simply never experiencing the event of interest during the course of the study. We don't know that the data point would never have experienced the event had they not "died" prematurely, so our best numerical representation of their survival is time $\tilde{\tau}_i$, or the last time at which this data point was noted to be "alive." If the observation dies before censorship, $\tilde{\tau}_i$ is still that last time that the data point is noted to be alive. It is the minimum of τ_i , the true censoring or death time, and C_i, the time at which the data is noted to have "died."

The survival function, denoted $S(t)$, is a graphical representation of survival probability as the study progresses. It is the complement of the CDF: $S(t) = 1 - F(t)$. The CDF in the context of any study represents the probability of experiencing the event of interest as time goes on, thus the function is constantly increasing or maintaining its value according to the y-axis. The survival function, however, represents the probability of survival at every point of time on the x-axis. It decreases at every point in time at which an observation in the dataset experiences the event. Censored events don't cause the function to drop, but they do influence the size of the drop at the point in time occurring after the time of censoring. The survival curve is useful because unlike other graphical analyses of data (such as linear and logistic regression), it accounts for censored observations. It better estimates the probability of not experiencing the event of interest occurring based on both the timing and occurrence (because the event may not necessarily occur) of observations, better handling the questions of "if" and "when."

Another component of survival analysis is the hazard function, which represents the risk of experiencing an event at the points in time denoted on the x-axis. It is closely related to the survival function in that it represents the rate of change of the survival function. This means that it can increase, decrease, or stay steady based on the activity of the survival function as it

continues along the x-axis. The hazard function remains low on the y-axis when the survival function is experiencing a slow decline. When the survival function has a steep decline, however, the hazard function spikes, thus at that point the risk of "death" has increased significantly. Though this concept wasn't a focus of our study of survival analysis, it is still useful for understanding time to event data.

In a real data set, the only elements we have access to are $\tilde{\tau}_i$ and C_i. As explained above, τ_i is the *true* moment at which the data point i was censored or died, but we only have the last point at which the data point was still in the study ($\tilde{\tau}_i$) and the first moment in the study that we weren't able to track it (C_i). With these two values, we can estimate the survival function $S(t)$. There are two estimate methods of interest: the naive estimator and the Kaplan Meier estimator.

The naive estimator is pictured in the following image. Verbally put, it is the number of objects in the study that we are still monitoring (that is, they haven't yet experienced the event nor have they been censored), divided by the number of objects that haven't yet been censored at time t. "Naive" is an apt name for this formula, as it ignores the observations that have a censoring time preceding time t.

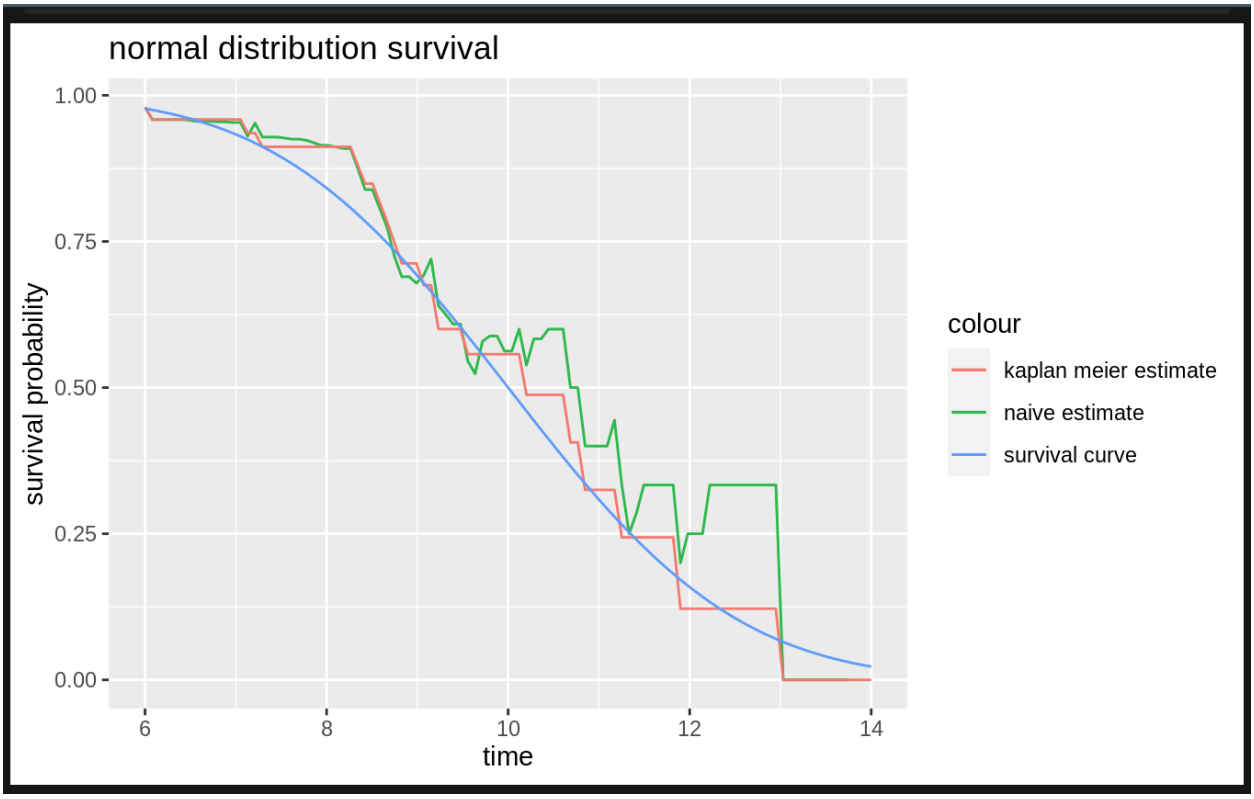
$$\hat{S}_{\text{naive}}(t) = \frac{1}{m(t)} \sum_{k:c_k \geq t} X_k = \frac{|\{1 \leq k \leq n \,:\, \tilde{\tau}_k \geq t\}|}{|\{1 \leq k \leq n \,:\, c_k \geq t\}|} = \frac{|\{1 \leq k \leq n \,:\, \tilde{\tau}_k \geq t\}|}{m(t)}$$

The better estimator is the aforementioned Kaplan Meier estimate, also pictured below and inspired by the naive estimator. It handles censoring better than the naive estimator, especially when there is a significant amount of censored data. The Kaplan Meier estimate has a recursive component to it. To calculate the Kaplan Meier estimate at time t–$S(t)$–take one minus the proportion of events at time t among those at risk and multiply that value by the probability of survival past the previous event time ($S(t)_{prev}$). Because of that last value being a previous calculation of $S(t)$, we are continually multiplying the result of this formula through the study's beginning time measurement.

$$\begin{aligned}
S(t) &= P(T > t) \\
&= P(T > t | T \geq t) P(T \geq t) \\
&= [1 - P(T \leq t | T \geq t)]\, P(T \geq t) \\
&= [1 - P(T = t | T \geq t)]\, P(T \geq t) \\
&= [1 - P(T = t | T \geq t)]\, P(T > t_{\text{prev}}) \\
&= [1 - P(T = t | T \geq t)]\, S(t_{\text{prev}})
\end{aligned}$$

Using the normal distribution, for which we know the *true* survival curve since there are no censored points, my group and I plotted its true survival function against its naive and Kaplan Meier estimates. It is evident that the Kaplan Meier estimator (in red) is a better approximation

of the normal distribution's true survival curve than is the naive estimator. This is the case in most real data sets as well.



It was a privilege to work on this research project with Ethan and Sherry. This quarter's work has led to a deeper understanding of statistical methodology and its applications, and a newfound interest of mine in biostatistics.