Lecture 33: Empirical Survivor Function
(Text Sections 10.3, 10.6)

When we wish to make inferences based on a model, we need methods to

1. Guide our choice of model
2. Check the appropriateness of this model for the given data.

To accomplish the first goal, we often rely on non-parametric methods (i.e. methods that are not based on models). One non-parametric tool for analyzing survival data is the empirical survivor function. The empirical survivor function is an estimate of the survivor function (the probability of survival past time \( y \)), which does not depend on distributional assumptions.

The most common way to estimate the survivor function non-parametrically is the Kaplan-Meier method. Two fundamental assumptions of this method are that survival times are independent, and that censoring times are independent of survival times (i.e. random dropout). The steps of this method are as follows:

1. Order the \( k \) observed survival times (uncensored data only) by increasing magnitude: \( 0 \leq y(1) \leq y(2) \cdots \leq y(k) \).
2. Define \( t_i \) as the \( i^{th} \) unique value in the series \( 0, y(1), y(2), \ldots, y(k), i = 1, \ldots, m \).
3. Define \( n_i \) as the number of subjects at risk (i.e. the number still alive and participating in the study) until just before time \( t_i, i = 1, \ldots, m \).
4. Define \( d_i \) as the number of subjects who fail at time \( t_i, i = 1, \ldots, m \).
5. For each \( i = 1, \ldots, m \), compute

\[
\hat{S}_i \equiv \prod_{j=1}^{i} \frac{n_j - d_j}{n_j}.
\]

This is equivalent to

\[
\hat{S}_i \equiv \left( \frac{n_i - d_i}{n_i} \right) \hat{S}_{i-1}
\]

for \( i > 1 \). This form is computationally simpler.

6. The estimate of the survivor function is then

\[
\hat{S}_Y(y) = \hat{S}_i, \quad t_i \leq y < t_{i+1}.
\]
Because of the structure of the Kaplan-Meier estimate, it is also called the \textit{product limit estimate}.

From the formula $\hat{S}_i = \left( \frac{n_i-d_i}{n_i} \right) \hat{S}_{i-1}$, we see that $\hat{S}_i$ is computed as the estimated probability of survival past time $t_{i-1}$ times the estimated probability of survival past time $t_i$ \textit{given} that the subject was still alive at time $t_{i-1}$.

Example: Remission times

The data set \texttt{remission} contains remission times of leukemia patients. There are two groups, one of which received a drug called mercaptopurine, the other of which received a placebo. Some treated patients’ data were censored, as indicated by the Status variable (1=uncensored, 0=censored). The question of interest is how survival time varies with treatment.

The boxplot of the survival times by treatment group shows that survival time seems to be longer in the treated group than in the untreated group. However, we need to be careful about interpreting this plot because many of the observations in the treated group are censored.

The Kaplan-Meier estimate of the survivor function takes the censoring into account, and hence is a better means of assessing the treatment effect.

The remission times (all uncensored) from the control group are as follows:

\begin{center}
\begin{tabular}{cccccccccccc}
 Controls & 1 & 1 & 2 & 2 & 3 & 4 & 4 & 5 & 5 & 8 & 8 & 8 & 11 & 11 & 12 & 12 & 15 & 17 & 22 & 23 \\
\end{tabular}
\end{center}

We can then calculate the Kaplan-Meier estimate:

\begin{center}
\begin{tabular}{cccc|c}
\hline
$t_i$ & $n_i$ & $d_i$ & $\frac{n_i-d_i}{n_i}$ & $\hat{S}_i$
\hline
0 & 21 & 0 & 1 & 1
1 & 21 & 2 & $\frac{19}{21}$ & $1 \cdot \frac{19}{21} = \frac{19}{21}$
2 & 19 & 2 & $\frac{17}{19}$ & $\frac{19}{21} \cdot \frac{17}{19} = \frac{17}{21}$
3 & 17 & 1 & $\frac{16}{17}$ & $\frac{17}{21} \cdot \frac{16}{17} = \frac{16}{21}$
\vdots & \vdots & \vdots & \vdots & \vdots
\hline
\end{tabular}
\end{center}

Note that, because there is no censoring in the control group, the Kaplan-Meier estimate reduces to the simple form $\hat{S}_i = \frac{21 - \sum_{j=1}^{i} d_j}{21}$.

In other words, the estimated probability of survival past time $t$ is the number of people alive just past time $t$ divided by the group size (21, in this case).

The remission times for the treated group are as follows (censored times indicated by \textasteriskcentered):
The Kaplan-Meier estimate is then:

\[ \hat{S}_i = 1 - \frac{d_i}{n_i} \]

In this case, where censoring is present, there is no simple form for \( \hat{S}_i \).

From the S-PLUS plot of the Kaplan-Meier estimates of \( S_Y(y) \) for each group, we see that survival times in the treated group appear to be longer than those in the untreated group.

Model Checking Using \( \hat{S}_Y(y) \)

Recall that, for the Weibull distribution:

\[
S_Y(y) = \exp \left[ - \left( \frac{y}{\theta} \right)^\lambda \right] \\
\log[- \log S_Y(y)] = \lambda \log y - \lambda \log \theta \\
\log H_Y(y) = \lambda \log y - \lambda \log \theta.
\]

In other words, if the survival times are Weibull distributed, we would expect \( \log H_Y(y) \) to be linear in \( \log y \).

The S-PLUS plot of the cumulative hazard vs. \( y \) (both on the log scale) for the two groups yields two approximately straight lines. Therefore, the Weibull assumption seems reasonable for these data.