The Kaplan-Meier (Product Limit) Estimate
STA312 Fall 2023

\(^1\)See last slide for copyright information.
Objective: To estimate the survival function without making any assumptions about the distribution of survival time.

If there were no censoring, it would be easy.

Use the empirical distribution function: the proportion of observations less than or equal to $t$.

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I\{t_i \leq t\}$$

Then let $\hat{S}_n(t) = 1 - \hat{F}_n(t)$
Consider times $t_0 = 0, t_1, t_2, \ldots$, maybe minutes or days.
Let $q_j =$ the probability of failing at time $t_j$, given survival to time $t_{j-1}$.
This is the idea behind the hazard function.
$p_j = 1 - q_j =$ the probability of surviving past time $t_j$, given survival past time $t_{j-1}$.

\[
p_j = \frac{P(T > t_j | T > t_{j-1})}{P(T > t_{j-1})} = \frac{P(T > t_j, T > t_{j-1})}{P(T > t_{j-1})} = \frac{P(T > t_j)}{P(T > t_{j-1})} = \frac{S(t_j)}{S(t_{j-1})}
\]
\[ p_j = \frac{S(t_j)}{S(t_{j-1})} \]

Probability of surviving past time \( t_j \), given survival past time \( t_{j-1} \)

With \( S(t_0) = S(0) = 1 \),

- \( p_1 = \frac{S(t_1)}{S(t_0)} = \frac{S(t_1)}{1} = S(t_1) \)
- \( p_2 = \frac{S(t_2)}{S(t_1)} \)
- \( p_3 = \frac{S(t_3)}{S(t_2)} \)
- Continuing . . .
- \( p_k = \frac{S(t_k)}{S(t_{k-1})} \)

Then,

\[
\begin{align*}
p_1 & \quad p_2 & \quad p_3 & \quad \cdots & \quad p_k \\
= & \quad S(t_1) & \quad \frac{S(t_2)}{S(t_1)} & \quad \frac{S(t_3)}{S(t_2)} & \quad \cdots & \quad \frac{S(t_k)}{S(t_{k-1})} \\
= & \quad S(t_k)
\end{align*}
\]
\[ S(t_k) = \prod_{j=1}^{k} p_j \]

Estimate \( S(t_k) \) by estimating the \( p_j \).
- Let \( d_j \) be the number of deaths at time \( t_j \).
- Let \( n_j \) be the number of individuals at risk before time \( t_j \).
- Anyone censored before time \( t_j \) is no longer at risk.
- Estimated probability of failure at time \( t_j \) is \( \hat{q}_j = \frac{d_j}{n_j} \).

\[
\hat{p}_j = 1 - \hat{q}_j = \frac{n_j - d_j}{n_j}
\]

\[
\hat{S}(t_k) = \prod_{j=1}^{k} \hat{p}_j
\]

\[
\hat{S}(t) = \prod_{t_j \leq t} \hat{p}_j
\]
Working toward a standard error for $\hat{S}(t) = \prod_{t_j \leq t} \hat{p}_j$

Large-sample Distribution Theory

- $\hat{p}_j = 1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j}$ is a sample proportion – a sample mean.
- It is the proportion of individuals eligible at risk for failure at time $t$, who did not fail.
- Mean of independent Bernoullis (conditionally on $n_j$).
- $E(\hat{p}_j) = p_j$, $Var(\hat{p}_j) = \frac{p_j (1-p_j)}{n_j}$
- $\hat{p}_j \sim N(p_j, \frac{p_j (1-p_j)}{n_j})$ by the Central Limit Theorem.
- This is for large $n_j$. 
Recall
Theorem based on the delta method of Cramér

Let \( \theta \in \mathbb{R}^k \). Under the conditions for which \( \hat{\theta}_n \) is asymptotically \( N_k(\theta, V_n) \) with \( V_n = \frac{1}{n} \mathbf{I}(\theta)^{-1} \), let the function \( g : \mathbb{R}^k \rightarrow \mathbb{R} \) be such that the elements of \( \dot{g}(\theta) = \left( \frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_k} \right) \) are continuous in a neighbourhood of the true parameter vector \( \theta \). Then

\[
g(\hat{\theta}) \sim N \left( g(\theta), \dot{g}(\theta) V_n \dot{g}(\theta)^\top \right).
\]

Note that the asymptotic variance \( \dot{g}(\theta) V_n \dot{g}(\theta)^\top \) is a matrix product: \((1 \times k)\) times \((k \times k)\) times \((k \times 1)\).

The standard error of \( g(\hat{\theta}) \) is \( \sqrt{\dot{g}(\hat{\theta}) V_n \dot{g}(\hat{\theta})^\top} \).
Specializing the delta method to the case of a single parameter
Yielding the univariate delta method

Let $\theta \in \mathbb{R}$. Under the conditions for which $\hat{\theta}_n$ is asymptotically $N(\theta, v_n)$ with $v_n = \frac{1}{n} I(\theta)$, let the function $g(x)$ have a continuous derivative in a neighbourhood of the true parameter $\theta$. Then

$$g(\hat{\theta}) \sim N\left(g(\theta), g'(\theta)^2 v_n\right).$$

The standard error of $g(\hat{\theta})$ is $\sqrt{g'(\hat{\theta})^2 \hat{v}_n}$, or $\left|g'(\hat{\theta})\right| \sqrt{\hat{v}_n}$. 
\[ \hat{S}(t) = \prod_{t_j \leq t} \hat{p}_j \text{ with } \hat{p}_j = \frac{n_j - d_j}{n_j} \sim N \left( p_j, \frac{p_j(1-p_j)}{n_j} \right) \]

- Sums are easier to work with than products.
- \( \log \hat{S}(t) = \sum_{t_j \leq t} \log \hat{p}_j \)
- Using the one-variable delta method, \( \log \hat{p}_j \sim N(\log p_j, \frac{1-p_j}{n_j p_j}) \)
- Sum of normals is normal (asymptotically, too).
- \( E(\sum_{t_j \leq t} \log \hat{p}_j) \approx \sum_{t_j \leq t} \log p_j = \log \prod_{t_j \leq t} p_j = \log S(t) \)

\[
\text{Var} \left( \sum_{t_j \leq t} \log \hat{p}_j \right) \approx \sum_{t_j \leq t} \text{Var}(\log \hat{p}_j) \\
= \sum_{t_j \leq t} \frac{1 - p_j}{n_j p_j}
\]
Asymptotic Distribution of \( \log \hat{S}(t) = \sum_{t_j \leq t} \log \hat{p}_j \)

\[
\log \hat{S}(t) \sim N \left( \log S(t), \sum_{t_j \leq t} \frac{1 - p_j}{n_j p_j} \right)
\]

- This is a stepping stone to the distribution of \( \hat{S}(t) \).
- Use the univariate delta method again.
- Univariate delta method says that if \( T_n \sim N(\theta, \nu_n) \) then \( g(T_n) \sim N \left( g(\theta), \nu_n[g'( \theta)]^2 \right) \).
- Here, \( T_n = \log \hat{S}_n(t) \), \( \theta = \log S(t) \) and \( g(x) = e^x \).
- \( g'(\theta) = e^\theta = e^{\log S(t)} = S(t) \). So,

\[
\hat{S}(t) \sim N \left( S(t), S(t)^2 \sum_{t_j \leq t} \frac{1 - p_j}{n_j p_j} \right)
\]
Standard error of $\hat{S}(t)$

Used in the denominator of $Z$-tests and $\hat{S}(t) \pm 1.96 \text{se}$

$$\hat{S}(t) \sim N \left( S(t), S(t)^2 \sum_{t_j \leq t} \frac{1-p_j}{n_j p_j} \right)$$

- Of course we don’t know $S(t)$ or $p_j$ in the variance.
- So use estimates.
- Estimate $S(t)$ with $\hat{S}(t)$, and estimate $p_j$ with $\hat{p}_j = \frac{n_j-d_j}{n_j}$.
- The resulting estimated asymptotic variance is $\hat{S}(t)^2 \sum_{t_j \leq t} \left( \frac{d_j}{n_j(n_j-d_j)} \right)$
- This is expression (3.1.2) on p. 27 of the text.
- The standard error of $\hat{S}(t)$ is $\hat{S}(t) \sqrt{\sum_{t_j \leq t} \left( \frac{d_j}{n_j(n_j-d_j)} \right)}$.
- In R’s survival package, the default confidence interval for the Kaplan-Meier estimate uses this standard error.
Distribution theory for the Kaplan Meier estimate (asymptotic normality, standard error etc.) has been presented the way it was originally developed.

The derivation is partly sound, but it has some holes.

More recently, viewing number of failures up to a point as a counting process (stochastic processes, STA348 and beyond) has cleaned the whole thing up.

Results are the same, but now the proofs are rigorous.

There was some guesswork in the development of these ideas, but the main guesses were right.
This slide show was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The \LaTeX{} source code is available from the course website:

http://www.utstat.toronto.edu/brunner/oldclass/312f23