11.1 Kaplan-Meier Estimator and Curve

Definition. Suppose a continuous random variable T has pdf $f(t)$ and cdf $F(t)$. The **survival function** $S(t) = 1 - F(t) = \mathbb{P}(T > t)$.

Note that the survival function uniquely defines the distribution. Moreover, the following relations hold: $f(t) = F'(t) = -S'(t)$, $F(t) = \int_{-\infty}^{t} f(u)du = 1 - S(t)$, and $S(t) = \int_{t}^{\infty} f(u)du = 1 - F(t)$.

We are interested in estimating empirically the survival function. Suppose the data consist of **survival times** or **times to event** for a certain number of individuals. The specificity of the data is that they may include censored observations. An observation is **censored** if it is know that the person survived (or hasn't experienced the event) up to certain time but nothing is known afterwards. It happens when an individual drops out of the study.
Definition. Suppose $t_1 < t_2 < \cdots < t_k$ are k distinct ordered survival times (or times to event). Note that there might be ties in the data: two or more events can occur at the same time. Also, along with the event, a censoring can occur at some of these times. Denote by $n_i, i = 1, \dots, k$, the number of individuals still alive (or those who have not experienced the event) shortly before time t_i (they are called **at-risk at time** t_i), and let e_i be the number of individuals who experienced the event at time t_i. The **Kaplan-Meier (KM) product-limit estimator** of the survival function is

$$\widehat{S}(t) = \prod_{i: t_i \leq t} \Big(1 - \frac{e_i}{n_i}\Big), \quad t \geq 0.$$

Derivation of the KM Estimator

Put $t_0 = 0$. Denote by $\pi_i = \mathbb{P}\big(T > t_i \,|\, T > t_{i-1}\big)$, $i = 1, \dots, k$.

The survival function at some fixed event time t_j may be written recursively as

$$\begin{aligned} S(t_j) &= \mathbb{P}(T > t_j) = \mathbb{P}(T > t_j \,|\, T > t_{j-1})\mathbb{P}(T > t_{j-1}) \\ &= \pi_j \, S(t_{j-1}) = \cdots = \prod_{i=1}^{j} \pi_i. \end{aligned}$$

The probabilities π_i's are estimated by the method of maximum likelihood. At any event time t_i, there are e_i individuals who experience the event at that time with probability $1 - \pi_i$ each, independently of all others, and there are $n_i - e_i$ individuals who are at risk but don't experience the event with probability π_i each, also independently of all others. Thus, the likelihood function has the form

$$L(\pi_1, \dots, \pi_j) = \prod_{i=1}^{j} (1-\pi_i)^{e_i} \pi_i^{n_i - e_i}.$$

Setting to zero the partial derivatives of the log-likelihood function $\ln L = \sum_{i=1}^{j} \big[e_i \ln(1-\pi_i) + (n_i - e_i) \ln \pi_i\big]$, we arrive at the system of normal equations

$$0 = \frac{\partial \ln L}{\partial \pi_i} = -\frac{e_i}{1-\pi_i} + \frac{n_i - e_i}{\pi_i}, \quad i = 1, \dots, j.$$

Thus, the maximum likelihood estimator $\widehat{\pi}_i$ of π_i solves

$$\frac{e_i}{1-\widehat{\pi}_i} = \frac{n_i - e_i}{\widehat{\pi}_i}, \quad i = 1, \dots, j.$$

The solution is

$$\widehat{\pi}_i = 1 - \frac{e_i}{n_i}, \quad i = 1, \dots, j.$$

Thus, for any event point t_j,

$$\widehat{S}(t_j) = \prod_{i=1}^{j} \Big(1 - \frac{e_i}{n_i}\Big).$$

For any time t such that $t_j < t < t_{j+1}$, $\widehat{S}(t)$, coincides with $\widehat{S}(t_j)$ since no events occur between times t_j and t.

Example. Times (in weeks) until remission for leukemia patients are recorded. They are 3, 5, 6+, 8, 8, 8+, 9, 12, 12+. The symbol "+" indicates that the observation was censored: either the patient dropped out of the study or hasn't experienced remission prior to the end of the study. The distinct times-to-event are $3, 5, 8, 9$, and 12. The calculations of the Kaplan-Meier estimator of the survival function are summarized in the following table.

time, t_i	at risk, n_i	event, e_i	survival rate, $1 - e_i/n_i$	estimator $\hat{S}(t_i)$
0	9	0	1	1
3	9	1	8/9	0.8889
5	8	1	7/8	0.7778
8	6	2	2/3	0.5185
9	3	1	2/3	0.3457
12	2	1	1/2	0.1728

Definition. The plot of the KM estimator against time is called the **Kaplan-Meier survival curve**. It is a step function with vertical lines corresponding to the event times. Times when censoring occurs are marked by some symbol (traditionally, a cross "x"). When censoring coincides with an event time, the convention is to put the symbol at the bottom of the step.

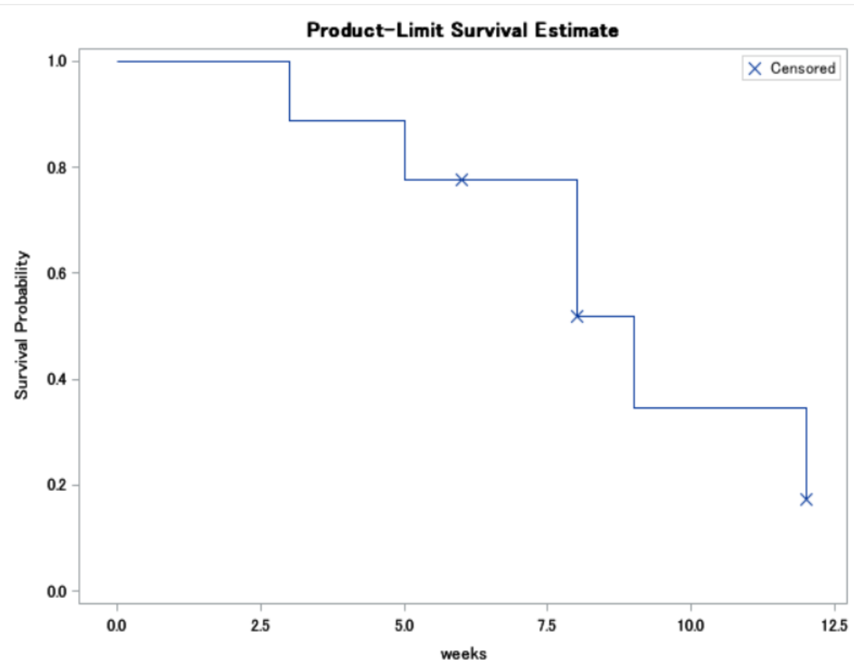
Example. In the above example, we use SAS and R to compute the KM estimator and to plot the KM curve.

In SAS:

```
data remission;
input weeks censored @@;
cards;
3 0 5 0 6 1 8 0 8 0 8 1 9 0 12 0 12 1
;

proc lifetest plots=(survival);
time weeks*censored(1);
run;
```

Product-Limit Survival Estimates						
weeks		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000		1.0000	0	0	0	9
3.0000		0.8889	0.1111	0.1048	1	8
5.0000		0.7778	0.2222	0.1386	2	7
6.0000	*	.	.	.	2	6
8.0000		.	.	.	3	5
8.0000		0.5185	0.4815	0.1759	4	4
8.0000	*	.	.	.	4	3
9.0000		0.3457	0.6543	0.1835	5	2
12.0000		0.1728	0.8272	0.1528	6	1
12.0000	*	.	.	.	6	0



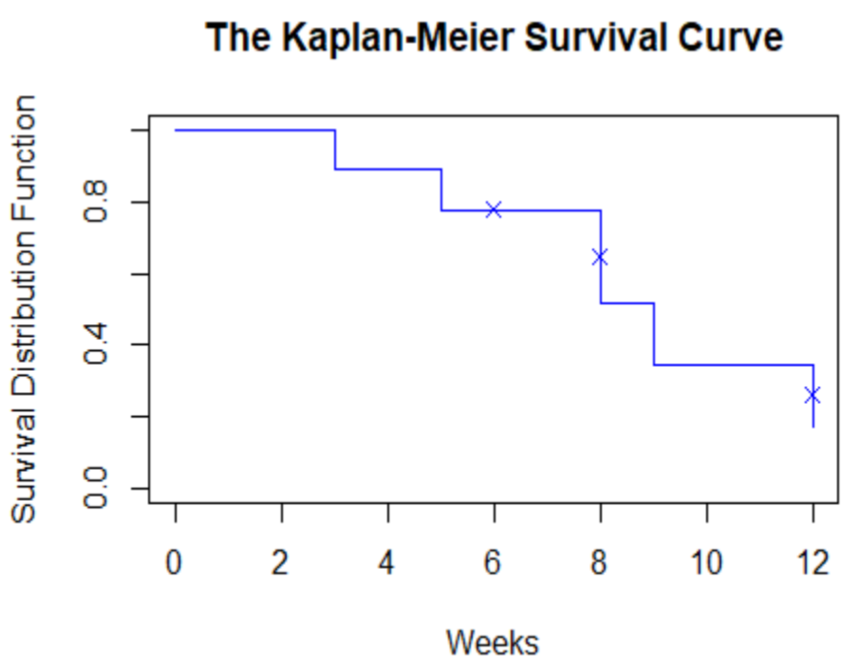
In R:
library(survival)
weeks<-c(3, 5, 6, 8, 8, 8, 9, 12, 12)

```
censored<-c(0, 0, 1, 0, 0, 1, 0, 0, 1)

#Surv() creates survival object
#survfit() produces KM estimator
weeks.surv <- survfit(Surv(weeks, censored==0)~ 1, se.fit=FALSE)
#no confidence band
summary(weeks.surv)
```

```
 time n.risk n.event survival
    3      9       1    0.889
    5      8       1    0.778
    8      6       2    0.519
    9      3       1    0.346
   12      2       1    0.173
```

```
#plotting KM survival curve
plot(weeks.surv, mark.time=TRUE, pch=4, col="blue", main="The Kaplan
-Meier Survival Curve", xlab="Weeks", ylab="Survival Distribution Function")
```



□