

The Kaplan-Meier Estimator as a Sum over Units

Malte C. Tichy

Siemens Gamesa Renewable Energy Deutschland GmbH

Beim Strohause 17-31

D-20097 Hamburg

2025-11-06

Abstract

A sum-wise formulation is proposed for the Kaplan-Meier product limit estimator of partially right-censored survival data. The derived representation permits to write the population's estimator as a sum over its individual units' semi-empirical estimators. This intuitive decomposition is applied to visualize the different contributions of failed and censored units to the overall population estimator.

1 Introduction: Survival Analysis and the Kaplan-Meier Estimator

The statistical survival problem (Moore 2016; Klein and Moeschberger 2003; Klein et al. 2014) naturally appears in many areas of science, business, medicine, and engineering. In these contexts, we encounter a non-negative continuous random variable which models, e.g., the survival time of a patient or the functional time of a technical component. The latter will be our working horse for illustration throughout this note: We will talk of functional (intact) or failed units. The event is then the failure of the unit, it occurs at age t^{fail} , some time after the unit's installment.

We consider a population of n units that were installed at different times in the past, and whose evolutions with time are now observed. Each unit j is observed from the moment of installation $\tau = 0$ until either its failure at t_j^{fail} , or until it becomes right-censored, either because it is still functional “now”, i.e. at the moment of the latest observation, or because it had left the dataset for some other reason in the past. For a given population with observed ages $\vec{t} = (t_1, \dots, t_n)$, the failure markers $\vec{\delta} = (\delta_1, \dots, \delta_n)$ indicate whether a particular unit j was censored at t_j ($\delta_j = 0, t_j < t_j^{\text{fail}}$) or was observed to fail at $t_j = t_j^{\text{fail}}$ ($\delta_j = 1$). For convenience of notation, we assume $t_1 < t_2 < \dots < t_n$. Ties $t_k = t_{k+1}$ can be resolved by formally adding a small delay ϵ to t_{k+1} . At each time instant t_k , there is exactly one unit that fails or is right-censored.

The non-parametric estimator for a population's cumulative failure distribution function had been used for a long time in the demographic and actuarial sciences (Gill 1980) until it was formalized by Kaplan and Meier (1958). For the following argument, it is convenient to write the Kaplan-Meier estimator at age τ of a population characterized by $\vec{t}, \vec{\delta}$ as:

$$F_{\text{KM}}(\tau; \vec{t}, \vec{\delta}) = 1 - \prod_{j=1}^n \left(1 - \frac{\delta_j \Theta(\tau - t_j)}{n - j + 1} \right), \quad (1)$$

where $\Theta(x)$ is the Heaviside step function with $\Theta(0) = 1$. This formulation can easily be related to the more common representation by noting that the number of units under risk at age $\tau = t_j$ is $n - j + 1$, and the number of failed units is δ_j .

The product representation in Eq. (1) resembles a competing-risk model (Kleinbaum and Klein 2012) that describes a population exposed to risks that occur at different ages t_j , with a probability of failure due to

the risk of $\delta_j/(n - k + 1)$. The Kaplan-Meier estimator is the predictive model for the data that maximizes the likelihood (Kalbfleisch and Prentice 2002).

As a product, the Kaplan-Meier estimator has a notably different formal representation compared to the empirical cumulative distribution function of a fully uncensored population:

$$F_{\text{KM}}(\tau; \vec{t}, \vec{\delta} = (1, \dots, 1)) = F_{\text{empirical}}(\tau; \vec{t}) = \frac{1}{n} \sum_{j=1}^n \Theta(\tau - t_j) \quad (2)$$

This sum-over-units representation permits a simple interpretation: The step function $\Theta(\tau - t_j)$ is the empirical cumulative failure distribution function (CDF) of the j th unit. Before age t_j , the unit was functional and the CDF value is zero; at and after t_j , it is known to have failed, and the CDF equals unity. There is no ambiguity about the fate of the j th unit, so the resulting CDF can be designated as *empirical*. The contribution of different units or sub-populations to the full population can then be easily and unambiguously identified: The value of the empirical estimator at a certain age τ , $F_{\text{empirical}}(\tau; \vec{t})$, is the fraction of units that have failed until that age, $F_{\text{empirical}}(\tau; \vec{t}) = \max(j | t_j \leq \tau) / n$.

This naturally raises the question of whether the Kaplan-Meier estimator (1) can also be written as a sum over units when right censoring obscures the fate of some of them, and whether a unit-level interpretation can be found. The goal of this short note is to present such a pedagogic formulation for Eq. (1) in full analogy to Eq. (2).

The motivation for finding and using such representation is twofold: On the one hand, statisticians often need to present and visualize data to non-technical audiences, and providing conceptually simple representations and explanations of the Kaplan-Meier estimate will ease communication. On the other hand, artifacts of the Kaplan-Meier plots may erroneously be interpreted as evidence for or against alignment of the dataset with a certain predictive model; de-composing the estimate into its contributing parts can shed light onto such situations.

1.1 Literature

The Kaplan-Meier estimator is the de facto textbook standard (Klein and Moeschberger 2003; Kleinbaum and Klein 2012; Kalbfleisch and Prentice 2002; Hosmer, Lemeshow, and May 2008) for estimating the cumulative failure probability in partially right-censored populations. Since its introduction, it has been the subject of extensive analysis and interpretation (Andersen and Borgan 1985; Gill 1980).

Numerous extensions and related estimators have been developed. For example, the cumulative hazard rate can be estimated by the method of (Nelson 1969) and (Aalen 1978) in close analogy to the cumulative failure distribution in Eq. (1).

Of particular relevance to the present note, a variety of alternative representations and interpretations of the Kaplan-Meier estimator have been proposed. For instance, it can be expressed as a function of empirical sub-survival functions for the uncensored and censored sub-populations (Peterson 1977). Redistribution-to-the-right methods provide simple heuristics for and interpretations of how to handle censored units: The key idea is to redistribute the probability mass of censored units to event times after their censoring time, reflecting the uncertainty about when censored units have eventually failed. The method was introduced by Efron (1967), where for a unit censored at age t_j , the probability mass associated with its potential failure is redistributed equally among all units still at risk at ages greater than t_j . A concrete implementation was proposed by Dinse (1985), the extension to interval censored data is treated by Betensky (2000).

The censoring process itself can also be formally viewed as another survival process that competes with the original failure mechanism. Following that line of thought, Satten and Datta (2001) expresses the Kaplan-Meier estimate as an average of step functions, with weights inversely proportional to the probability of censoring (Robins and Rotnitzky 1992). In this representation, we see how each observed failure is “enhanced” by the degree of censoring that is expected until that moment (Meira-Machado 2023).

2 Sum-Wise formulation of the Kaplan-Meier Estimator

We now propose an individual estimator for censored units, and show that the average of such unit-level estimators yields the Kaplan-Meier estimator of the population.

2.1 Semi-Empirical Unit-Level Cumulative Failure Probabilities

The survival status of an uncensored unit that fails at age t is described by a step function in age τ ,

$$\text{CDF}_{\text{uncensored}}(\tau; t) = \Theta(\tau - t), \quad (3)$$

that is, by the empirical distribution function with the unspectacular step-wise shape shown in Figure 1.

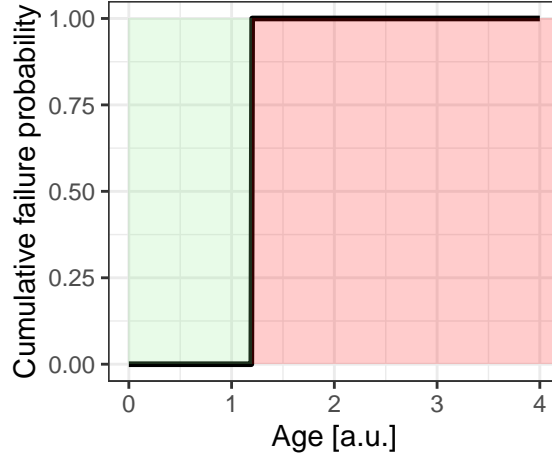


Figure 1: Empirical distribution function of a unit failing at age 1.2. The green shaded area visualizes an intact unit under observation, the red shaded area indicates a failed unit under observation.

In the absence of censoring, we possess full information about the subject's time evolution: It was under observation constantly, and its failure age is known to be $t^{\text{fail}} = 1.2$. Any suppression of information, be it by censoring or truncation, leaves a gap in the knowledge of the empirical CDF - for certain moments in age τ , the observer does not know with certainty whether the unit was intact ($F(\tau) = 0$) or broken ($F(\tau) = 1$), and needs to estimate the probability, based on some model. For right censoring, the last observed age t is then only a lower bound to the true event age: The unit might have failed immediately after leaving the dataset, or much later.

In general, a right-censored unit can be described by the cumulative distribution function **conditioned** on having survived until age $\tau = t$:

$$G_{\text{right censored}}(\tau; t, F) = \Theta(\tau - t) \frac{F(\tau) - F(t)}{1 - F(t)}, \quad (4)$$

where $F(\tau)$ is an imputation function that is necessary as modelling assumption to treat the unobserved time span. As visualized in Figure 2, the unit is observed to be functional until the censoring age t , after which we assume that its cumulative failure probability evolves as described by the imputation function F .

Within a population of many units, to avoid any dependency on a particular parametric modelling assumption F , one can use the Kaplan-Meier estimator itself as imputation function $F(t)$ for the j th unit, which yields

$$G_j^{\text{KM imputation}}(\tau; \vec{t}, \vec{\delta}) = \delta_j \Theta(\tau - t_j) + (1 - \delta_j) \Theta(\tau - t_j) \frac{F_{\text{KM}}(\tau; \vec{t}, \vec{\delta}) - F_{\text{KM}}(t_j; \vec{t}, \vec{\delta})}{1 - F_{\text{KM}}(t_j; \vec{t}, \vec{\delta})} \quad (5)$$

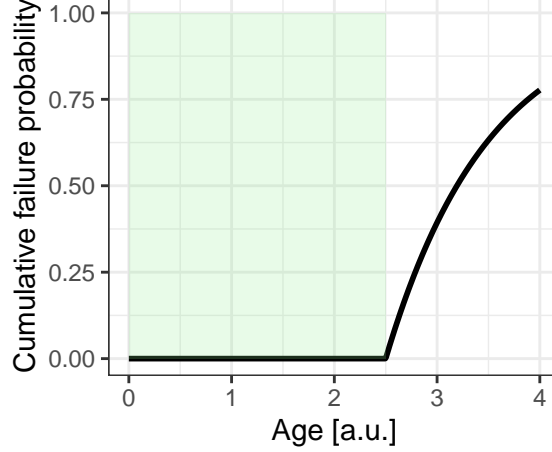


Figure 2: Semi-empirical distribution function of a unit censored at age 2.5. After having been under observation and undoubtedly intact until 2.5, it leaves the observation, and some failure model is assumed. The unshaded area represents a age period in which the status of the unit is uncertain.

In the following, we seek a more intuitive and interpretable representation of $G_j^{\text{KM imputation}}(\tau; \vec{t}, \vec{\delta})$, following the spirit behind the redistribution-to-the-right algorithms introduced by (Efron 1967; Dinse 1985; Betensky 2000).

2.2 Sum-Wise Formulation

We define a unit-wise Kaplan-Meier estimator $G_j(\tau; \vec{t}, \vec{\delta})$ as follows: When $\delta_j = 1$, the unit's failure age is observed at age t_j , such that the unit-wise estimator $G_j(\tau; \vec{t}, \vec{\delta})$ must match the empirical step-function $\Theta(\tau - t_j)$. When $\delta_j = 0$, we instead set $G_j(\tau; \vec{t}, \vec{\delta})$ to the Kaplan-Meier-estimator of the set of all units that are still functional and not censored at moment t_j .

That is, following a maximum-likelihood-logic, since we know that the unit has not failed yet, we disregard all other units that failed or were censored before ($k < j$), and keep the “later” units $k > j$ to estimate the expected behavior of unit j . That is, we set the unit-level Kaplan-Meier estimator to

$$G_j(\tau; \vec{t}, \vec{\delta}) = \delta_j \Theta(\tau - t_j) + (1 - \delta_j) F_{\text{KM}}(\tau; \vec{t}_{>j}, \vec{\delta}_{>j}) \quad (6)$$

where $\vec{t}_{>j} = (t_{j+1}, \dots, t_n)$, $\vec{\delta}_{>j} = (\delta_{j+1}, \dots, \delta_n)$ designate the subsets of units $k = j + 1, \dots, n$ with failure or censoring times after the failure or censoring time of unit j .

We assert that the average of such unit-wise defined estimators matches the population-level Kaplan-Meier estimator. Our proposition formally reads:

$$F_{\text{KM}}(\tau; \vec{t}, \vec{\delta}) = \frac{1}{n} \sum_{j=1}^n G_j(\tau; \vec{t}, \vec{\delta}) \equiv \frac{1}{n} \sum_{j=1}^n \left(\delta_j \Theta(\tau - t_j) + (1 - \delta_j) F_{\text{KM}}(\tau; \vec{t}_{>j}, \vec{\delta}_{>j}) \right), \quad (7)$$

which is equivalent to

$$1 - \prod_{j=1}^n \left(1 - \frac{\delta_j \Theta(\tau - t_j)}{n - j + 1} \right) = \frac{1}{n} \sum_{j=1}^n \left(\delta_j \Theta(\tau - t_j) + (1 - \delta_j) \left(1 - \prod_{k=j+1}^n \left(1 - \frac{\delta_k \Theta(\tau - t_k)}{n - k + 1} \right) \right) \right) \quad (8)$$

The left-hand-side is the well-known product-wise representation of the Kaplan-Meier-estimator, which is asserted to coincide with a unit-wise additive form on the right-hand side.

2.3 Proof

Eq. (8) can be proven by induction. For convenience, we set

$$L(\tau; \vec{\delta}, \vec{t}) = 1 - \prod_{j=1}^n \left(1 - \frac{\delta_j \Theta(\tau - t_j)}{n - j + 1} \right) \quad (9)$$

$$R(\tau; \vec{\delta}, \vec{t}) = \frac{1}{n} \sum_{j=1}^n \left(\delta_j \Theta(\tau - t_j) + (1 - \delta_j) \left(1 - \prod_{k=j+1}^n \left(1 - \frac{\delta_k \Theta(\tau - t_k)}{n - k + 1} \right) \right) \right) \quad (10)$$

It is to be proven that $L(\tau; \vec{\delta}, \vec{t}) = R(\tau; \vec{\delta}, \vec{t})$ for all $0 < t_1 < \dots < t_n$, all $\vec{\delta} \in \{0, 1\}^n$ and all $0 \leq \tau \leq t_n$.

Base case: For $n = 1$, the product in L and the sum in R collapse, and one easily sees that

$$L(\tau; \vec{\delta} = (\delta_1), \vec{t} = (t_1)) = \delta_1 \Theta(\tau - t_1) = R(\tau; \vec{\delta} = (\delta_1), \vec{t} = (t_1)) \quad (11)$$

Induction step $n \rightarrow n + 1$: Assuming that for a given set $(\vec{t}, \vec{\delta})$, the hypothesis is true, we add a new unit at index 0 with $t_0 < t_1$. This proof strategy does not restrict generality, since any set of $\vec{t}, \vec{\delta}$ can be constructed starting from the largest age t_n . In the induction step, we use $L(\tau; \vec{\delta}, \vec{t}) = R(\tau; \vec{\delta}, \vec{t})$ to prove $L(\tau; (\delta_0, \vec{\delta}), (t_0, \vec{t})) = R(\tau; (\delta_0, \vec{\delta}), (t_0, \vec{t}))$, where $(\delta_0, \vec{\delta}) = (\delta_0, \delta_1, \dots, \delta_n)$ and $(t_0, \vec{t}) = (t_0, t_1, \dots, t_n)$.

We can relate $L(\tau; (\delta_0, \vec{\delta}), (t_0, \vec{t}))$ to $L(\tau; \vec{\delta}, \vec{t})$:

$$\begin{aligned} L(\tau; (\delta_0, \vec{\delta}), (t_0, \vec{t})) &= 1 - (1 - L(\tau; \vec{\delta}, \vec{t})) \cdot \left(1 - \frac{\delta_0 \Theta(\tau - t_0)}{n + 1} \right) \\ &= L(\tau; \vec{\delta}, \vec{t}) + \frac{\delta_0 \Theta(\tau - t_0)}{n + 1} - L(\tau; \vec{\delta}, \vec{t}) \frac{\delta_0 \Theta(\tau - t_0)}{n + 1} \end{aligned} \quad (12)$$

Also $R(\tau; (\delta_0, \vec{\delta}), (t_0, \vec{t}))$ can be expressed with the help of $R(\tau; \vec{\delta}, \vec{t})$:

$$\begin{aligned} R(\tau; (\delta_0, \vec{\delta}), (t_0, \vec{t})) &= \frac{1}{n + 1} \sum_{j=0}^n \left(\delta_j \Theta(\tau - t_j) + (1 - \delta_j) \left(1 - \prod_{k=j+1}^n \left(1 - \frac{\delta_k \Theta(\tau - t_k)}{n - k + 1} \right) \right) \right) \\ &= \frac{1}{n + 1} \left(\delta_0 \Theta(\tau - t_0) + (1 - \delta_0) R(\tau; \vec{\delta}, \vec{t}) + n R(\tau; \vec{\delta}, \vec{t}) \right) \end{aligned} \quad (13)$$

We treat $\delta_0 = 0$ and $\delta_0 = 1$ separately.

- When a new censored unit is introduced at t_0 , we deal with $\delta_0 = 0$:

$$\begin{aligned} L(\tau; (\delta_0 = 0, \vec{\delta}), (t_0, \vec{t})) &\stackrel{(12)}{=} L(\tau; \vec{\delta}, \vec{t}) \\ &\stackrel{(\text{ind. hyp.})}{=} R(\tau; \vec{\delta}, \vec{t}) \\ &\stackrel{(13)}{=} R(\tau; (\delta_0 = 0, \vec{\delta}), (t_0, \vec{t})) \end{aligned} \quad (14)$$

- When a new failed unit is introduced at t_0 , we have $\delta_0 = 1$:

$$\begin{aligned} L(\tau; (\delta_0 = 1, \vec{\delta}), (t_0, \vec{t})) &\stackrel{(12)}{=} L(\tau; \vec{\delta}, \vec{t}) + \frac{\Theta(\tau - t_0)}{n + 1} - L(\tau; \vec{\delta}, \vec{t}) \frac{\Theta(\tau - t_0)}{n + 1} \\ &\stackrel{(\text{ind. hyp.})}{=} \frac{1}{n + 1} \left((n + 1) R(\tau; \vec{\delta}, \vec{t}) + \Theta(\tau - t_0) - R(\tau; \vec{\delta}, \vec{t}) \Theta(\tau - t_0) \right) \\ &\stackrel{(13)}{=} R(\tau; (\delta_0, \vec{\delta}), (t_0, \vec{t})) + \frac{1}{1 + n} \left(R(\tau; \vec{\delta}, \vec{t}) (1 - \Theta(\tau - t_0)) \right) \end{aligned} \quad (15)$$

Since $t_0 < t_1$, we have $R(\tau < t_0; \vec{\delta}, \vec{t}) = 0$, which implies $R(\tau; \vec{\delta}, \vec{t})(1 - \Theta(\tau - t_0)) = 0$ for all τ . This concludes the proof.

Since it is also easy to see that the to-be-proven equality holds for $\delta = (1, 1, \dots, 1)$ and $\delta = (0, 0, \dots, 0)$, an alternative proof idea is to do an induction over values of δ , i.e. to “swap” one value of δ_j from 0 to 1.

2.4 Consistency with Imputation Ansatz

It remains to be shown that the proposed unit-wise formulation in Eq. (6) matches the imputation Ansatz of Eq. (5). We thereby come to the proposition

$$F_{\text{KM}}(\tau; \vec{t}_{>j}, \vec{\delta}_{>j}) = \frac{F_{\text{KM}}(\tau; \vec{t}, \vec{\delta}) - F_{\text{KM}}(t_j; \vec{t}, \vec{\delta})}{1 - F_{\text{KM}}(t_j; \vec{t}, \vec{\delta})}, \quad (16)$$

for ages $\tau > t_j$.

This equality can be proven by invoking the product representation of the Kaplan-Meier estimator (1). We then find for the left-hand side of Eq. (16):

$$F_{\text{KM}}(\tau; \vec{t}_{>j}, \vec{\delta}_{>j}) = 1 - \prod_{k=j+1}^n \left(\frac{n+1-k-\delta_k \Theta(\tau - t_k)}{n+1-k} \right) \quad (17)$$

For the right-hand side of Eq. (16):

$$\frac{F_{\text{KM}}(\tau; \vec{t}, \vec{\delta}) - F_{\text{KM}}(t_j; \vec{t}, \vec{\delta})}{1 - F_{\text{KM}}(t_j; \vec{t}, \vec{\delta})} = 1 - \prod_{k=1}^n \left(\frac{n-k+1-\delta_k \Theta(\tau - t_k)}{n-k+1-\delta_k \Theta(t_j - t_k)} \right), \quad (18)$$

where, due to $\tau > t_j$, the factor in the last product matches unity for $k \leq j$, while for the remaining values $k > j$, the last term in the denominator vanishes. This proves the equality with the right hand side of Eq. (17). Alternatively, it is thinkable to induce over k , with the base case $k = 0$.

3 Visualization

3.1 Granular Example

The sum-wise formulation allows to construct intuitive visualizations and grasp which units are driving the value of the Kaplan-Meier estimator at certain ages.

As a simple example, consider observed ages $\vec{t} = (1, 2, 3, 4, 5, 6)$ and failure markers $\vec{\delta} = (0, 1, 0, 1, 1, 0)$. The resulting individual unit-level estimators of Eq. (6) are visualized in Figure 3. Note that the individual estimator of the very first unit (which is censored soon after being installed, with no other unit failing before) matches the overall Kaplan-Meier-estimator - what we can say about the likely fate of the first unit is determined by the behavior of the other units.

Another useful visualization is shown in Figure 4, where we can see how the different units contribute as summands to the overall estimator: Units whose failure was observed ($\delta_j = 1$) contribute with a constant summand to the overall Kaplan-Meier-estimator (units 2, 4, 5), whereas censored units have an increasing contribution that reflects our assumption how likely it is that they will fail (units 1, 3). Unit 6 has not failed at all until the largest observed age, such that it does not contribute at all to the cumulative failure probability; on the contrary, the existence of a “long survivor” has pushed down the overall value.

Thanks to the sum-wise representation, we see that the Kaplan-Meier estimator is the sum of a fully empirical component (the contribution of the units that actually failed) and a predicted contribution (the censored

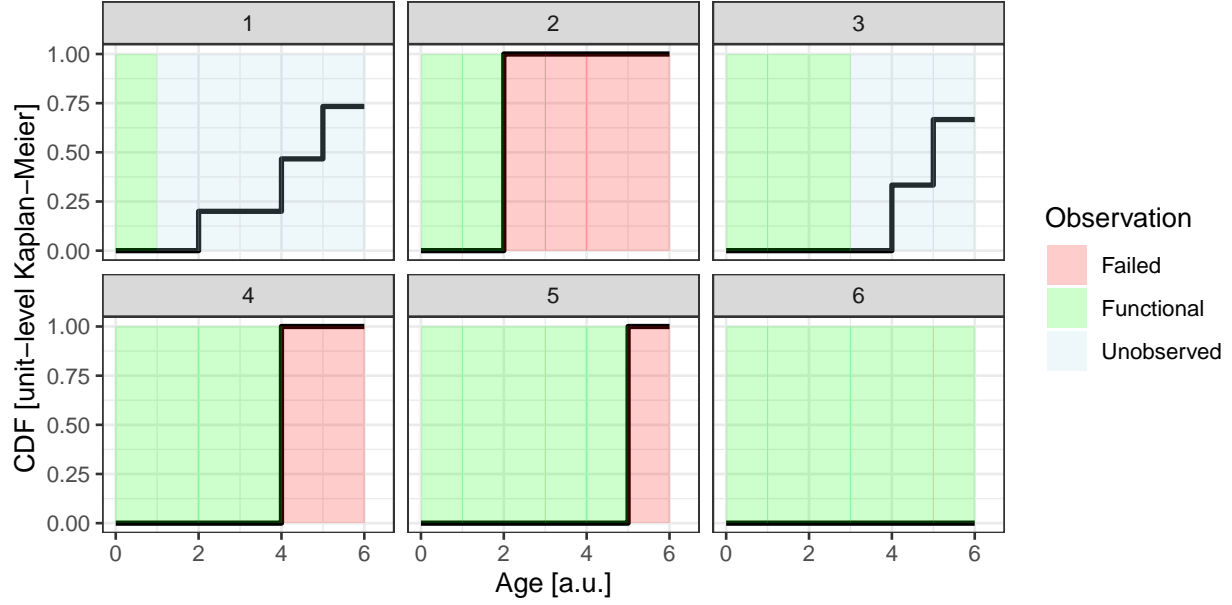


Figure 3: Unit-level estimators for simple dataset.

units, for which a certain failure probability is assumed). Therefore, even if the value of the estimator remains unchanged if we removed the first unit, the contributions would shift: The overall value is then borne more strongly by the remaining ones. If, on the other hand, we added 20 units that are immediately censored after their creation, the resulting Kaplan-Meier estimator would be turned from a mostly-empirical to a mostly-predictive nature – even if its predicted values do not change.

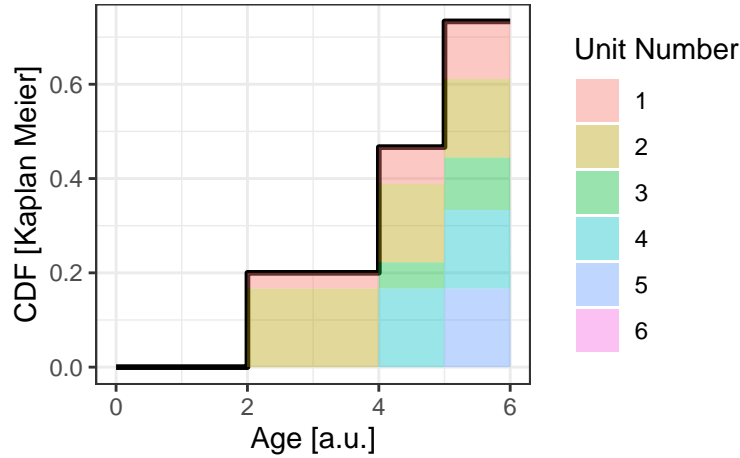


Figure 4: Sum of normalized unit-level estimators, which yields the population-level Kaplan-Meier estimator. The shaded areas represent the fraction of the population CDF that can be attributed to the individual unit.

3.2 Population Example

As a further example how the contributions of censored and observed units interplay, we simulate 100 units that fail under a Weibull process with shape parameter 1.4 and scale parameter 1. Censoring is applied under a Weibull-distributed censoring age with shape 1 and scale 1.5.

The resulting Kaplan-Meier estimator and its separation into empirical and predicted contributions are shown in Figure 5.

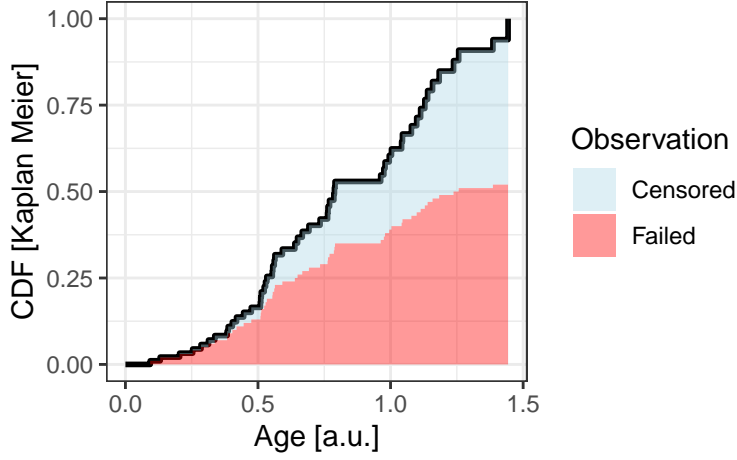


Figure 5: Kaplan-Meier estimator of a population of 100 units, the shading color reflects the status of the individual units that contribute to the overall estimator.

The fraction of the Kaplan-Meier estimator that is borne by truly observed failures changes with age, and reflects the degree of censoring.

3.3 Interpretation

As shown by these two examples, the unit-level additive representation of the Kaplan-Meier estimator Eq. (8) offers an intuitive vehicle to analyse and communicate a given survival dataset of partially observed units. By decomposing the population-level estimator into contributions from individual units, we gain a granular view of how observed failures and censored units shape the cumulative failure curve. Analysts can then assess whether the estimator is primarily data-driven or imputation-driven at different unit ages. This is particularly helpful in situations in which a population of some initial size n is observed for a long time, for example, a fixed set of expensive and long-living technical components.

The Kaplan-Meier assumption turns out to be self-consistent on unit level: The unit-level estimator for a unit censored at t_j reflects the information that is available to predict that unit's fate, namely the behavior of all units with observed ages larger than t_j . The maximum-likelihood forecast for that unit is then, unsurprisingly, again the Kaplan-Meier estimate of that sub-population.

Consistently, given a certain population, adding new trivially unobserved units with $t_j = 0, \delta_j = 0$ leaves the population-level Kaplan-Meier estimator unchanged – no new information about the survival behavior of the component is added by starting up new units. The unit-level estimators of the new upstarted units become the Kaplan-Meier estimate of the previous population. The meaning of the Kaplan-Meier-estimator of that full population becomes a different one, however, as it pivots from an empirical description of what happened to a prediction of what will happen.

4 Conclusions and Outlook

The unit-level decomposition and the representation in Eq. (8) provide a transparent and interpretable framework for understanding the mechanics of the Kaplan-Meier estimator, enhancing both technical analysis and stakeholder communication. The sum-wise formulation is consistent with the imputation ansatz in Eq. (4). In this context, the Kaplan-Meier estimator serves as a self-consistent fallback imputation prediction $f(\tau), F(\tau)$.

While the Kaplan-Meier-estimator is parameter-free, it is not at all assumption-free: For any censored unit j , the CDF remains at its current value until the next failure event at $t_{k>j}$, regardless of how far into the future that event may occur. This assumption is quite strong, and it leads to paradoxical and problematic situations. Depending on the situation, other assumptions for the “fall-back” failure probabilities $f(\tau)$, $F(\tau)$ can be more appropriate. Again, the decomposition makes the mechanics of the Kaplan-Meier estimate more transparent, and can help in judging to which extent the Kaplan-Meier estimator is appropriate.

Several desiderata immediately emerge: Uncertainty bands are typically generated via bootstrapping or via Greenwood’s formula (Greenwood 1926), and the question arises whether a distinction between statistically induced and censoring induced components is possible. In this note, we have only considered right-censored units for which the Kaplan-Meier estimate possesses the closed formula (1), the generalization to uniquely left-censored units (Gomez et al. 1992) is straightforward.

The question naturally arises whether the recursive approach and the unit-wise representation could help in finding representations of estimators for mixed populations with left-, right- and interval-censoring, and/or left- and right truncation. This endeavor might be quite demanding, since no closed form of the Turnbull estimator (Turnbull 1976a, 1976b) for interval-censored populations is available. Note, however, that the unit-level formulation is also the fixed point of this iterative procedure for unit-level estimators: Given a set of unit-level estimators $H_j^{(k)}$ ($j = 1..n$) in the k th iteration of the algorithm, the next iteration yields

$$H_j^{(k+1)}(\tau; \vec{t}, \vec{\delta}) = \delta \cdot \Theta(\tau - t_j) + (1 - \delta) \frac{H^{(k)}(\tau; \vec{t}, \vec{\delta}) - H^{(k)}(t_j; \vec{t}, \vec{\delta})}{1 - H^{(k)}(t_j; \vec{t}, \vec{\delta})}. \quad (19)$$

Perhaps the unit-wise perspective can thereby also help in more complex censoring and truncation situations.

Finally, similarly to the sum-representation of the Kaplan-Meier estimator of Eq. (7), an analogous representation of the hazard in the Nelson-Aalen estimator (Nelson 1969; Aalen 1978) could be envisaged. Given the importance of finite-size and partially censored and / or truncated survival datasets in many domains of science and technology, further work that facilitates the interpretation and visualization of the related estimators will be highly valuable and appreciated.

Acknowledgements

The author would like to thank Josu Aguirrebeitia Celaya, Anna Linke and Tim Mazzotta for helpful comments on the manuscript, and acknowledges insightful discussions with Bärbel Angersbach, Kun Marhadi, Lucas Mäde and Michael Revesz.

References

- Aalen, Odd. 1978. “Nonparametric estimation of partial transition probabilities in multiple decrement models.” *Ann. Stat.* 6 (3): 534–45.
- Andersen, Per Kragh, and Ørnulf Borgan. 1985. “Counting Process Models for Life History Data: A Review.” *Scand. J. Stat.* 12 (2): 97–158.
- Betensky, Rebecca A. 2000. “Redistribution algorithms for censored data.” *Stat. Probab. Lett.* 46 (11): 385–89.
- Dinse, Gregg E. 1985. “An Alternative to Efron’s Redistribution-of-Mass Construction of the Kaplan-Meier Estimator.” *Am. Stat.* 39 (4): 299–300.
- Efron, Bradley. 1967. “The Two Sample Problem with Censored Data.” In *Berkeley Symp. Math. Stat. Prob.*, 831–53.
- Gill, R. D. 1980. “Censoring and Stochastic Integrals.” Amsterdam: Mathematisch Centrum.
- Gomez, Guadalupe, Olga Julià, Frederic Utzet, and Melvin L. Moeschberger. 1992. “Survival Analysis For Left Censored Data.” In *Surviv. Anal. State Art*, 269–88. https://doi.org/10.1007/978-94-015-7983-4_16.
- Greenwood, M. 1926. “The natural duration of cancer.” *Rep. Public Health Med. Subj. (Lond)*. 33: 1–26.

- Hosmer, David W., Stanley Lemeshow, and Susanne May. 2008. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Wiley.
- Kalbfleisch, John D., and Ross L. Prentice. 2002. *The Statistical Analysis of Failure Time Data*. Wiley Series. John Wiley & Sons.
- Kaplan, E. L., and Paul Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *J. Am. Stat. Assoc.* 53 (282): 457–81.
- Klein, John P., Hans C Van Houwelingen, Joseph G Ibrahim, and Thomas H Scheike, eds. 2014. *Handbook of Survival Analysis*. CRC Press.
- Klein, John P., and Melvin L. Moeschberger. 2003. *Survival Analysis - Techniques for Censored and Truncated Data*. Edited by K. Dietz, M. Gail, K. Krickeberg, J. Samet, and A. Tsiatis. Springer New York Berlin Heidelberg.
- Kleinbaum, David G, and Mitchel Klein. 2012. *Survival Analysis*. Springer New York Berlin Heidelberg.
- Meira-Machado, Luís. 2023. "The Kaplan-Meier Estimator: New Insights and Applications in Multi-state Survival Analysis." In *Comput. Sci. Its Appl. - ICCSA 2023 Work. Athens, Greece, July 3–6, 2023, Proceedings, Part IX*, 129–39.
- Moore, Dirk F. 2016. *Applied Survival Analysis Using R*. Springer Nature Switzerland.
- Nelson, Wayne. 1969. "Hazard Plotting for Incomplete Failure Data." *J. Qual. Technol.* 1 (1): 27–52.
- Peterson, Arthur V. 1977. "Expressing the Kaplan-Meier Estimator as a Function of Empirical Subsurvival Functions." *J. Am. Stat. Assoc.* 72 (360a): 854–58.
- Robins, James M., and Andrea Rotnitzky. 1992. "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers." In *AIDS Epidemiol. Methodol. Issues*, edited by Nicholas P. Jewell, Klaus Dietz, and Vernon T. Farewell, 297–331. Springer Science + Business Media, LLC.
- Satten, Glen A., and Somnath Datta. 2001. "The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average." *Am. Stat.* 55 (3): 207–10.
- Turnbull, Bruce W. 1976a. "Nonparametric Estimation of a Survivorship Function with Doubly Censored Data." *J. Am. Stat. Assoc.* 69 (345): 169–73.
- . 1976b. "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data." *J. R. Stat. Soc. Ser. B* 38 (3): 290–95.