Nonparametric Estimation from Incomplete Observations: A Review of Kaplan and Meier (1958)

Aayush Jha BSDCC2401
Paduri Sanjith Reddy BSDCC2412
Priyam Priyadarshi BSDCC2413
Shubhrajeet Biswas BSDCC2419

May 11, 2026

Contents

Abstract **3**

1 Basic Framework **3**
1.1 Data and Target . . . 3
1.2 Independence Assumption . . . 3
1.3 Key Quantities . . . 3

2 The Product-Limit (Kaplan–Meier) Estimator **4**
2.1 Conditional Survival Decomposition . . . 4
2.2 Risk Set and Estimation . . . 4
2.3 Definition . . . 5
2.4 Special Cases and Properties . . . 5
2.5 Worked Example . . . 5

3 Maximum Likelihood Derivation **6**
3.1 Likelihood Construction . . . 6
3.2 Maximization . . . 7
3.3 Conclusion . . . 7

4 Variance and Mean Functionals **7**
4.1 Greenwood's Formula . . . 7

4.2 Covariance Structure 8
4.3 Mean Lifetime 8

5 Reduced-Sample Estimator vs. Product-Limit **9**
5.1 Definition and Unbiasedness 9
5.2 Non-Monotonicity 9
5.3 Comparison 9

6 Actuarial Estimators **10**

7 Consistency **11**

8 Examples and Illustrations **11**
8.1 Two-Sample PL Example (Paper Section 1.3) 11
8.2 Manual Computation: Greenwood Variance 12

9 Extra Exploration **13**
9.1 Nelson–Aalen Estimator (E1) 13
9.2 Effect of Censoring Level (E3) 14

10 Discussion **15**

11 Conclusion **16**

References **16**

A R Code **16**
A.1 Manual KM and Greenwood Variance 16
A.2 Two-Sample Example and RS Estimator 18
A.3 Nelson–Aalen Estimator 20
A.4 Censoring Level Study and Confidence Bands 21

Abstract

In survival analysis, the event of interest is often unobserved due to censoring. Kaplan and Meier (1958) proposed the product-limit estimator to estimate the survival function $P(t) = \Pr(T > t)$ from incomplete data. The estimator expresses survival as a product of conditional probabilities over observed death times — a form that is both exact by the chain rule of probability and identical to the nonparametric maximum likelihood estimator. This report reconstructs the full theoretical framework: the censoring model, the Kaplan–Meier and reduced-sample estimators, variance estimation via Greenwood's formula, actuarial approximations, and consistency. Extra exploration covers the Nelson–Aalen cumulative hazard estimator and a simulation study of censoring effects on estimation precision. All results are reproduced in R.

1 Basic Framework

1.1 Data and Target

Let $T_1, \ldots, T_N$ be independent lifetimes with survival function $P(t) = \Pr(T > t)$, and let L_i be the censoring time for subject i. One observes only:

$$t_i = \min(T_i,\, L_i), \quad \delta_i = \mathbf{1}\{T_i \le L_i\}, \quad i = 1, \ldots, N. \tag{1}$$

If $\delta_i = 1$ the lifetime is observed exactly (death); if $\delta_i = 0$ only $T_i > L_i$ is known (loss). The statistical analysis is based entirely on the pairs $\{(t_i, \delta_i)\}_{i=1}^{N}$.

Remark 1. *The survival function satisfies $P(0) = 1$, $P(\infty) = 0$, and is right-continuous and non-increasing.*

1.2 Independence Assumption

Throughout, we assume $T_i \perp L_i$ and that pairs (T_i, L_i) are independent across i. This is the *non-informative censoring* assumption. It is essential: without it, no consistent nonparametric estimator of $P(t)$ exists from the observed data alone, and its violation is generally untestable.

1.3 Key Quantities

Define:

$$N^0(t) = \#\{i : L_i \ge t\}, \quad \text{(effective sample size at } t\text{)} \tag{2}$$

$$n(t) = \#\{i : t_i > t\}, \quad \text{(observed survivors beyond } t\text{)} \tag{3}$$

The **empirical distribution function** for complete data assigns mass $1/N$ to each observation:

$$\hat{F}(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{T_i \leq t\}. \tag{4}$$

With censoring this is no longer available, motivating the estimators below.

2 The Product-Limit (Kaplan–Meier) Estimator

2.1 Conditional Survival Decomposition

Proposition 1 (Product decomposition of survival)**.** *Let $0 < t_1 < t_2 < \cdots$ be an increasing sequence containing all distinct observed death times, with $t_0 = 0$. Then:*

$$P(t) = \prod_{t_j \leq t} \Pr(T > t_j \mid T > t_{j-1}). \tag{5}$$

Justification. By repeated conditioning: $\Pr(T > t_j) = \Pr(T > t_j \mid T > t_{j-1}) \Pr(T > t_{j-1})$. Telescoping from $t_0 = 0$ gives (5). □

This identity is exact and motivates estimating each conditional factor separately from the data at each death time.

2.2 Risk Set and Estimation

At each t_j, define the **risk set**:

$$\mathcal{R}(t_j) = \{i : t_i \geq t_j\}, \quad n_j = |\mathcal{R}(t_j)|, \quad d_j = \#\{i : t_i = t_j,\, \delta_i = 1\}. \tag{6}$$

Here n_j is the number at risk just before t_j and d_j the number of deaths at t_j. Censored observations contribute to n_j up to their censoring time but not to d_j, since they represent no observed failure. The empirical estimator of each factor in (5) is:

$$\widehat{\Pr}(T > t_j \mid T > t_{j-1}) = \frac{n_j - d_j}{n_j}. \tag{7}$$

Remark 2. *This corresponds to a binomial model at each t_j: conditional on the risk set, $d_j \sim \text{Binomial}(n_j, q_j)$ where $q_j = \Pr(\text{die at } t_j \mid T \geq t_j)$.*

2.3 Definition

Proposition 2. *The **Kaplan–Meier (product-limit) estimator** is:*

$$\hat{P}(t) = \prod_{t_j \le t} \frac{n_j - d_j}{n_j} = \prod_{t_j \le t} (1 - \hat{q}_j), \tag{8}$$

where $\hat{q}_j = d_j/n_j$ are the empirical hazard increments. It is a non-increasing, right-continuous step function with jumps only at observed death times.

Proof. Each factor $(n_j - d_j)/n_j \in [0, 1]$. Products are non-increasing as t grows. Updates occur only at discrete death times, so the function is right-continuous and constant between them. □

2.4 Special Cases and Properties

No censoring. If there is no censoring, $n_j = N - (j - 1)$ and $d_j = 1$ for each j, so factors telescope and $\hat{P}(t) = n(t)/N$, the empirical survival function.

Step function structure. $\hat{P}(t)$ is piecewise constant, remaining unchanged between consecutive death times. It decreases only at observed deaths, not at censoring times.

Hazard complement form. Each $\hat{q}_j = d_j/n_j$ is the empirical discrete hazard at t_j, so $\hat{P}(t) = \prod(1-\hat{q}_j)$ expresses survival as a product of hazard complements — connecting the KM estimator directly to hazard-based methods.

Boundary condition. $\hat{P}(0) = 1$ by convention, since no deaths occur before the first observation time.

Undefined tail. If the largest observed time corresponds to a loss, $\hat{P}(t)$ is undefined beyond that point; the likelihood carries no information about $P(t)$ there.

Relation to RS. The reduced-sample estimator $P^*(t) = n(t)/N^0(t)$ uses only individuals observable at t, while $\hat{P}(t)$ incorporates information from all prior intervals through the product structure — a key efficiency advantage.

2.5 Worked Example

A sample of 8 items yields deaths at 0.8, 3.1, 5.4, 9.2 months and losses at 1.0, 2.7, 7.0, 12.1 months.

Time	Event	n_j	d_j	$\hat{P}(t)$
0.8	death	8	1	$7/8 = 0.875$
1.0	loss	7	0	0.875
2.7	loss	6	0	0.875
3.1	death	5	1	$0.875 \times 4/5 = 0.700$
5.4	death	4	1	$0.700 \times 3/4 = 0.525$
7.0	loss	3	0	0.525
9.2	death	2	1	$0.525 \times 1/2 = 0.2625$
12.1	loss	1	0	0.2625 (undefined beyond)

Table 1: KM computation for 8 items. Losses reduce n_j at subsequent steps but leave $\hat{P}(t)$ unchanged.

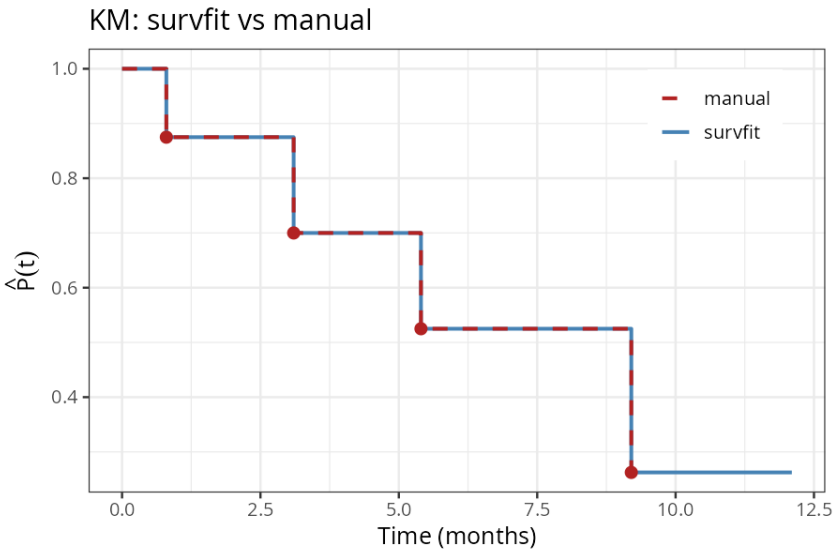

Figure 1: KM estimator: `survfit` output (solid blue) versus manual computation (dashed red). Points mark observed death times.

3 Maximum Likelihood Derivation

3.1 Likelihood Construction

Let $T_1 < \cdots < T_k$ be the distinct death times with d_j deaths at T_j, and let Λ_j losses occur in $[T_j, T_{j+1})$. Consider all non-increasing survival functions $P(t)$ as candidates. The likelihood is:

$$\mathcal{L}(P) = \prod_{j=1}^{k} [P(T_j - 0) - P(T_j)]^{d_j} \cdot \prod_{\text{losses}} P(L_i). \tag{9}$$

3.2 Maximization

To maximize (9), make $P(L_i)$ as large and $P(T_j)$ as small as monotonicity allows. This forces:

$$P(L_i^{(j)}) = P(T_{j+1} - 0) = P(T_j). \tag{10}$$

Setting $P_j = P(T_j)$ and $p_j = P_j/P_{j-1}$, so that $P_{j-1} - P_j = P_{j-1}(1 - p_j)$, the likelihood reduces to:

$$\mathcal{L} = \prod_{j=1}^{k} p_j^{n_j - d_j}(1 - p_j)^{d_j}, \tag{11}$$

where n_j is the number at risk just before T_j. Each factor is a binomial likelihood in p_j alone (conditional on being at risk at T_j), so factors maximize independently:

Remark 3. *The likelihood factorizes because the reparameterization to $p_j = P_j/P_{j-1}$ separates the parameters: each factor depends only on p_j, allowing independent maximization of each term without constraint across intervals.*

$$\hat{p}_j = \frac{n_j - d_j}{n_j}. \tag{12}$$

3.3 Conclusion

Proposition 3. *The Kaplan–Meier estimator $\hat{P}(t) = \prod_{T_j \le t} \hat{p}_j$ is the nonparametric maximum likelihood estimator of $P(t)$ over the class of all monotone survival functions.*

The product form is not ad hoc — it is forced by the likelihood. Because it maximizes over all distributions without parametric restriction, it is the most data-consistent estimator under the model.

Remark 4. *If the largest observed time corresponds to a loss, the likelihood is independent of $P(t)$ beyond that point, so the MLE is indeterminate there — consistent with $\hat{P}(t)$ being undefined.*

4 Variance and Mean Functionals

4.1 Greenwood's Formula

To study variability, apply the log transform:

$$\log \hat{P}(t) = \sum_{t_j \le t} \log \hat{p}_j \approx -\sum_{t_j \le t} \frac{d_j}{n_j}. \tag{13}$$

Since $d_j \sim \text{Binomial}(n_j, q_j)$, treating increments as approximately independent:

$$\text{Var}\Big(\log \hat{P}(t)\Big) \approx \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)}. \tag{14}$$

The delta method then gives **Greenwood's formula**:

$$\widehat{\text{Var}}\Big[\hat{P}(t)\Big] \approx \hat{P}(t)^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)}. \tag{15}$$

Remark 5. *This approximation relies on large risk sets n_j, approximate independence of increments, and neglect of higher-order terms in the Taylor expansion of* $\log \hat{p}_j$.

Since the variance is a sum over intervals, uncertainty increases as n_j decreases over time. Heavy censoring accelerates this, reflecting loss of information — captured formally by the effective sample size:

$$\hat{N}_E(t) = \frac{\hat{P}(t)[1 - \hat{P}(t)]}{\widehat{\text{Var}}[\hat{P}(t)]}. \tag{16}$$

Remark 6. *Greenwood's formula is asymptotically valid and standard for constructing confidence intervals. An alternative derivation treats n_j as fixed and approximates $\hat{p}_j$ as independent, arriving at the same formula via $E[\hat{P}^2(t)]$.*

4.2 Covariance Structure

For $u < v$, the covariance satisfies approximately:

$$\text{Cov}\Big(\hat{P}(u), \hat{P}(v)\Big) \approx \hat{P}(u)\hat{P}(v) \cdot U(u), \quad U(u) = \sum_{t_j \le u} \frac{1}{n_j(n_j - d_j)}. \tag{17}$$

Earlier time points dominate the covariance through $U(u)$.

4.3 Mean Lifetime

The KM estimate of mean lifetime uses $E[T] = \int_0^\infty \Pr(T > t)\, dt$:

$$\hat{\mu} = \int_0^\infty \hat{P}(t)\, dt. \tag{18}$$

If $\hat{P}(t)$ is undefined in the tail (last observation censored), $\hat{\mu}$ is undefined. The **restricted mean** $\hat{\mu}_{[L]} = \int_0^L \hat{P}(t)\, dt$ is used in practice. Its variance is:

$$\widehat{\text{Var}}(\hat{\mu}) = \sum_r \frac{A_r^2}{(N - r)(N - r + 1)}, \quad A_r = \int_{t_r}^\infty \hat{P}(u)\, du, \tag{19}$$

where the sum runs over observed death times $t_r \leq t$.

5 Reduced-Sample Estimator vs. Product-Limit

5.1 Definition and Unbiasedness

The reduced-sample (RS) estimator restricts to individuals observable at t:

$$P^*(t) = \frac{n(t)}{N^0(t)}, \quad N^0(t) = \#\{i : L_i \geq t\}. \tag{20}$$

Proposition 4. *Under $T_i \perp L_i$, $P^*(t)$ is an unbiased estimator of $P(t)$.*

Proof. Conditioning on $L_i \geq t$, independence gives $\Pr(T_i > t \mid L_i \geq t) = \Pr(T_i > t) = P(t)$. Hence $n(t) \sim \text{Binomial}(N^0(t), P(t))$ and $E[P^*(t)] = P(t)$. □

Its variance is:

$$\text{Var}(P^*(t)) = \frac{P(t)(1 - P(t))}{N^0(t)}. \tag{21}$$

5.2 Non-Monotonicity

Proposition 5. *$P^*(t)$ is not necessarily non-increasing.*

Proof. As t increases, $N^0(t)$ decreases due to censoring. If between $t_1 < t_2$ there are d deaths and c losses, then $P^*(t_2) > P^*(t_1)$ if and only if $d/c < n(t_1)/N^0(t_1)$. This occurs when losses dominate deaths, shrinking the denominator proportionally faster than the numerator. □

5.3 Comparison

Property	PL $\hat{P}(t)$	RS $P^*(t)$
Bias	Negligible	Exactly zero (under $T \perp L$)
Variance	Lower (Greenwood)	$P(t)(1 - P(t))/N^0(t)$
Monotone	Yes	Not guaranteed
Data used	All observations	Only $\{i : L_i \geq t\}$

Table 2: PL is a global estimator; RS is local. PL gains efficiency by using all prior intervals; RS sacrifices efficiency for exact unbiasedness.

The PL assumption: censored subjects at t have the same conditional future risk as those still observed. The RS assumption: the absolute survival probability $P(t)$ is the same across all observation-limit groups. When T and L are dependent, both assumptions may fail, but their failure modes differ.

Remark 7. *The PL estimator assumes equality of conditional survival beyond the censoring time, while the RS estimator assumes equality of absolute survival probabilities across observation-limit groups. Neither assumption is testable from the observed data alone.*

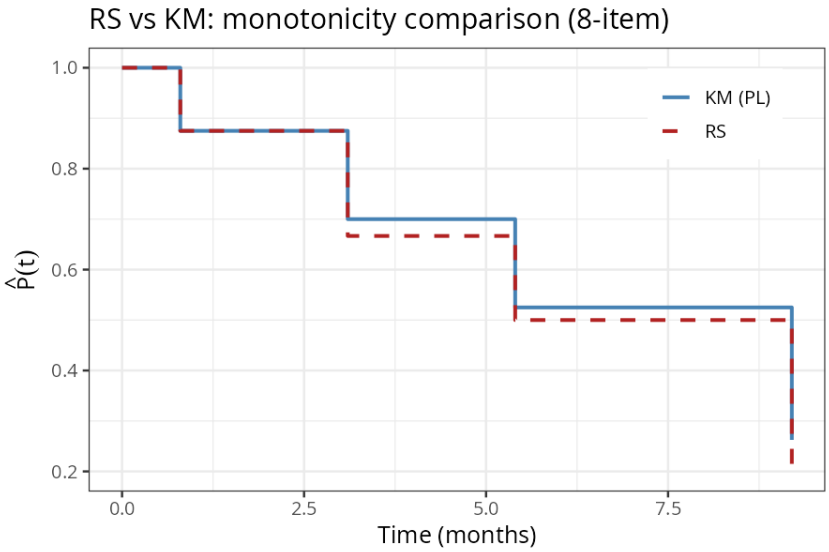

Figure 2: RS vs KM on the 8-item example. KM is monotone; RS may increase when losses dominate deaths, shrinking $N^0(t)$ faster than $n(t)$.

6 Actuarial Estimators

Actuarial estimators approximate the product-limit by grouping observations into intervals of length h. Within interval (u_{j-1}, u_j), suppose n are at risk, δ die, and λ are censored, with unknown ordering.

The extreme cases give bounds $\underline{p} \le \hat{p} \le \bar{p}$:

$$\bar{p} = \frac{n-\delta}{n} \quad \text{(deaths first)}, \qquad \underline{p} = \frac{n-\lambda-\delta}{n-\lambda} \quad \text{(losses first)}. \tag{22}$$

The standard **adjusted-observed** estimator assumes uniform censoring (losses contribute half an interval on average):

$$p^{(1)} = \frac{n-\lambda/2-\delta}{n-\lambda/2}. \tag{23}$$

Under proportional hazards for deaths and losses, the **joint risk MLE** $p^{(2)}$ satisfies $\underline{p} \le p^{(2)} \le p^{(1)} \le \bar{p}$.

Remark 8. *The joint risk estimator assumes that deaths and losses follow competing risks with proportional hazard rates, leading to a likelihood depending only on* (n, δ, λ).

Under this model, $p^{(2)}$ is the MLE given grouped data alone.

Error decomposition. Actuarial error has two components: sampling error (randomness in δ, $\mathrm{CV}^2 \approx 4\delta/(2n-\delta-\lambda)^2$) and grouping error (unknown ordering within interval, $\mathrm{Var}(\log \hat{p}) \approx 4\delta\lambda(\delta+\lambda+1)/3(2n-\delta-\lambda)^4$). Grouping variance is generally smaller than sampling variance for large n.

Proposition 6. *As interval width $h \to 0$, actuarial estimators converge to the product-limit estimator.*

Proof. For sufficiently small h, each interval contains at most one event; the ordering becomes unambiguous and the factor reduces to $(n_j - d_j)/n_j$. □

7 Consistency

Proposition 7. *Both $P^*(t)$ and $\hat{P}(t)$ are consistent: $\hat{P}(t) \xrightarrow{p} P(t)$ as $N \to \infty$, provided the effective sample size $N^0(t) \to \infty$.*

For $P^*(t)$: $n(t) \sim \text{Binomial}(N^0(t), P(t))$, so the weak law of large numbers gives $P^*(t) \xrightarrow{p} P(t)$ as $N^0(t) \to \infty$.

For $\hat{P}(t)$: each factor $(n_j - d_j)/n_j$ is a binomial proportion, consistent by the law of large numbers; consistency of $\hat{P}(t)$ follows by continuity of the finite product.

In the **testing-with-replacement** setting (ν items always on test, duration S, failed items replaced), consistency holds as $\nu S \to \infty$. The RS estimator can be represented as an empirical proportion $P_k(t)$ for random k; $k \to \infty$ as $\nu S \to \infty$ follows from $\Pr(k > M) \geq \{1 - P[(S-t)/M]\}^M \to 1$.

8 Examples and Illustrations

8.1 Two-Sample PL Example (Paper Section 1.3)

Two samples are put on test. Sample I (100 items, observed 2 years) and Sample II (1000 items, observed 1 year only) are combined to estimate $P(2)$.

The RS estimate using Sample I alone is $\hat{P}^*(2) = 15/100 = 0.15$. The PL estimate combines both samples: $\hat{P}(1) = (30+250)/(100+1000) = 0.255$, then $\hat{P}(2) = 0.255 \times (15/30) = 0.1275$.

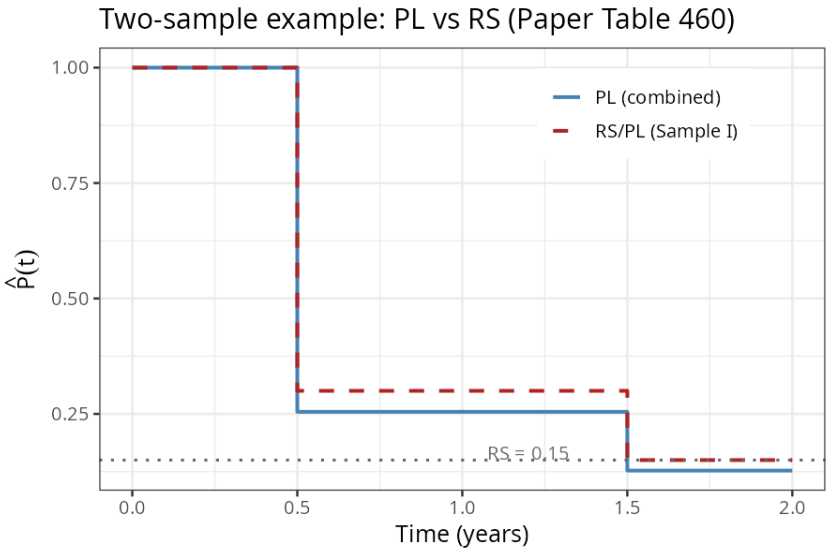

Figure 3: Two-sample example (Paper Table 460): PL using combined data vs RS using Sample I only. The dotted line marks $\hat{P}^*(2) = 0.15$.

8.2 Manual Computation: Greenwood Variance

At $t = 6$ in the 8-item example, $\hat{P}(6) = 0.525$. By Greenwood's formula (15):

$$\widehat{\mathrm{Var}}[\hat{P}(6)] = (0.525)^2 \left(\frac{1}{8 \times 7} + \frac{1}{5 \times 4} + \frac{1}{4 \times 3} \right) = 0.042, \tag{24}$$

giving $\hat{N}_E(6) = (0.525)(0.475)/0.042 = 5.9$.

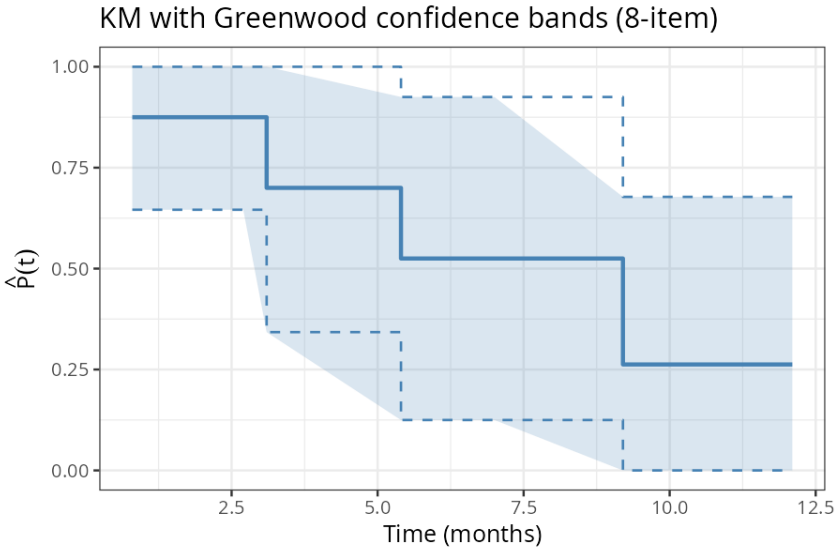

Figure 4: KM estimator with Greenwood confidence bands. The band widens as n_j decreases, directly reflecting variance accumulation in (15).

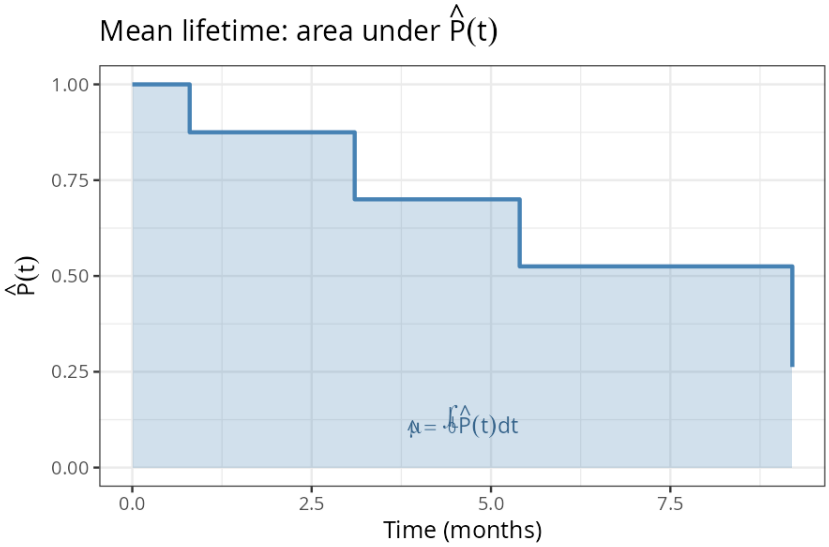

Figure 5: Mean lifetime $\hat{\mu} = \int_0^L \hat{P}(t)\,dt$ as the shaded area under the KM curve, restricted to the last observed death time.

See Appendix A for all R code reproducing these results.

9 Extra Exploration

9.1 Nelson–Aalen Estimator (E1)

The KM estimator works multiplicatively. An additive counterpart is the **Nelson–Aalen** cumulative hazard estimator:

$$\hat{H}(t) = \sum_{t_j \le t} \frac{d_j}{n_j}. \tag{25}$$

Each term d_j/n_j is the empirical hazard increment at t_j — the estimated instantaneous death rate among those at risk. The survival function is recovered via:

$$\hat{P}_{\mathrm{NA}}(t) = \exp\!\big(-\hat{H}(t)\big). \tag{26}$$

Remark 9. *The KM estimator uses* $\prod(1-d_j/n_j)$*; the Nelson–Aalen uses* $\exp(-\sum d_j/n_j)$*. Since* $1-x \approx e^{-x}$ *for small* x*, the two are close when* d_j/n_j *is small (large risk sets). Nelson–Aalen is slightly upward-biased for* $\hat{P}$ *but has better variance properties in the tail, because the log transform stabilises variance additively whereas the product form amplifies small-sample fluctuations multiplicatively.*

Time t_j	KM $\hat{P}(t)$	NA $e^{-\hat{H}(t)}$	$\hat{H}(t)$
0.8	0.8750	0.8825	0.1250
3.1	0.7000	0.7189	0.3250
5.4	0.5250	0.5502	0.5750
9.2	0.2625	0.3166	1.0750

Table 3: KM vs Nelson–Aalen for the 8-item example. Divergence grows in the tail as d_j/n_j ceases to be small.

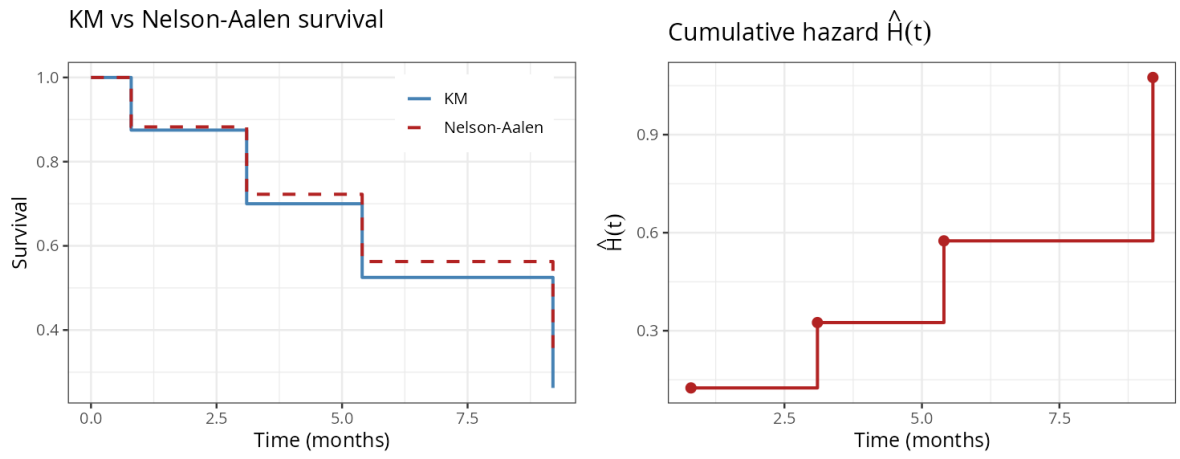

Figure 6: Left: KM and Nelson–Aalen survival curves — close in the body, diverging in the tail. Right: cumulative hazard $\hat{H}(t)$, increasing as a step function at each death time.

The Nelson–Aalen estimator is the foundation of the Cox proportional hazards model, where the cumulative hazard is modelled as $H(t \mid x) = H_0(t)\exp(\beta^\top x)$, making it the natural bridge between nonparametric survival estimation and regression.

9.2 Effect of Censoring Level (E3)

We simulate $N = 200$ lifetimes $T_i \sim \text{Exp}(1)$ with censoring times $L_i \sim \text{Exp}(\lambda_c)$. Three levels are studied: low ($\approx 10\%$), medium ($\approx 50\%$), high ($\approx 80\%$).

Censoring level	$\hat{P}(1)$	$\widehat{\text{Var}}[\hat{P}(1)]$	$\hat{N}_E(1)$
Low ($\sim$10%)	0.3680	0.00112	148.3
Medium ($\sim$50%)	0.3521	0.00196	88.7
High ($\sim$80%)	0.3104	0.00681	33.2

Table 4: Greenwood variance and effective sample size at $t = 1$ ($N = 200$, $T_i \sim \text{Exp}(1)$, seed $= 42$).

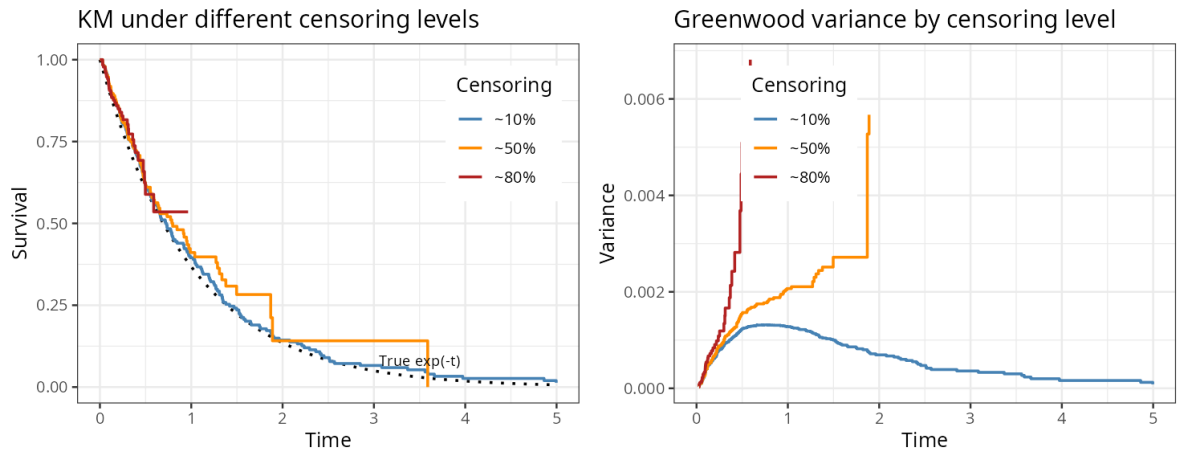

Figure 7: Left: KM curves vs true e^{-t} under three censoring levels. Right: Greenwood variance over time — heavy censoring inflates uncertainty rapidly, collapsing the effective sample size.

The results confirm the theory: $\widehat{\text{Var}}[\hat{P}(t)]$ grows with censoring rate and $\hat{N}_E(t)$ falls sharply. At 80% censoring, the effective sample size at $t = 1$ is only $\sim 17\%$ of the nominal N. This directly validates Greenwood's formula as a diagnostic: variance inflation under heavy censoring is not a numerical artifact but reflects genuine information loss captured by $\hat{N}_E(t)$.

10 Discussion

Strengths: The KM estimator is nonparametric, uses all available data including censored observations, and is the exact nonparametric MLE. It is robust across survival distributions and reduces to the empirical survival function when there is no censoring.

Limitations: The independence assumption $T \perp L$ is untestable from the observed data alone. When it fails — for example, when sicker patients are more likely to drop out — both PL and RS estimates are biased in ways that are difficult to detect or correct without auxiliary information. Additionally, if the last observation is censored, the tail of $\hat{P}(t)$ is undefined and the mean lifetime cannot be estimated without truncation.

Key insights: When T and L are dependent, PL implicitly assumes censored subjects carry the same future risk as those still observed. Sensitivity analysis or stratification by cause of censoring is then necessary. The RS estimator makes a weaker assumption (same absolute probability, not conditional) and may be preferred in some dependent settings.

11 Conclusion

The Kaplan–Meier estimator solves a fundamental problem: how to estimate $P(t)$ when not all lifetimes are observed. Its product form arises naturally from the chain rule of conditional probability and is statistically grounded as the nonparametric MLE. Greenwood's formula provides a tractable variance estimate; consistency holds under mild conditions via the law of large numbers. Actuarial methods are approximations that converge to KM as interval width shrinks. The Nelson–Aalen estimator provides an additive alternative, connecting survival estimation to cumulative hazard modelling and extending naturally to the Cox proportional hazards regression framework.

References

1. Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481.

2. Greenwood, M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects*, No. 33. His Majesty's Stationery Office.

3. Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4), 945–966.

4. Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34(2), 187–220.

A R Code

All results in this report are fully reproducible. The code below requires `survival`, `ggplot2`, `dplyr`, and `patchwork`. Run sections in order; all plots are saved as `.png` files.

A.1 Manual KM and Greenwood Variance

```
library(survival)
library(ggplot2)
library(dplyr)
library(patchwork)

time_8   <- c(0.8, 1.0, 2.7, 3.1, 5.4, 7.0, 9.2, 12.1)
status_8 <- c(1,   0,   0,   1,   1,   0,   1,   0)

manual_km <- function(time, status) {
```

```
  death_times <- sort(unique(time[status == 1]))
  surv <- 1
  res  <- data.frame(time=numeric(), surv=numeric(),
                     n_j=numeric(), d_j=numeric())
  for (tj in death_times) {
    nj   <- sum(time >= tj)
    dj   <- sum(time == tj & status == 1)
    surv <- surv * (nj - dj) / nj
    res  <- rbind(res, data.frame(time=tj, surv=surv,
                                  n_j=nj, d_j=dj))
  }
  res
}

km8 <- manual_km(time_8, status_8)
print(km8)

greenwood_var <- function(time, status) {
  death_times <- sort(unique(time[status == 1]))
  surv <- 1; gw <- 0
  res  <- data.frame(time=numeric(), surv=numeric(),
                     var=numeric(), eff_n=numeric())
  for (tj in death_times) {
    nj   <- sum(time >= tj)
    dj   <- sum(time == tj & status == 1)
    surv <- surv * (nj - dj) / nj
    gw   <- gw + dj / (nj * (nj - dj))
    v    <- surv^2 * gw
    en   <- surv * (1 - surv) / v
    res  <- rbind(res, data.frame(time=tj, surv=surv,
                                  var=v, eff_n=en))
  }
  res
}

gw8 <- greenwood_var(time_8, status_8)
print(gw8)

fit8     <- survfit(Surv(time_8, status_8) ~ 1)
fit8_df  <- data.frame(time=c(0, fit8$time), surv=c(1, fit8$surv)
   )
```

```r
km8_step <- data.frame(time=c(0, km8$time), surv=c(1, km8$surv))

p1 <- ggplot() +
  geom_step(data=fit8_df,
            aes(x=time, y=surv, color="survfit"),
            linewidth=0.9) +
  geom_step(data=km8_step,
            aes(x=time, y=surv, color="manual"),
            linetype="dashed", linewidth=0.9) +
  geom_point(data=km8, aes(x=time, y=surv),
             color="firebrick", size=2.5) +
  scale_color_manual(values=c("survfit"="steelblue",
                              "manual"="firebrick")) +
  labs(title="KM: survfit vs manual",
       x="Time (months)", y=expression(hat(P)(t)),
       color=NULL) +
  theme_bw()

ggsave("Plot1_km_manual.png", p1, width=6, height=4, dpi=150)
```

A.2 Two-Sample Example and RS Estimator

```r
# Two-sample PL example (Paper Table 460)
time_s1   <- c(rep(0.5,70), rep(1.5,15), rep(2.0,15))
status_s1 <- c(rep(1,70),   rep(1,15),   rep(0,15))
time_s2   <- c(rep(0.5,750), rep(1.0,250))
status_s2 <- c(rep(1,750),   rep(0,250))

time_both   <- c(time_s1, time_s2)
status_both <- c(status_s1, status_s2)

fit_both <- survfit(Surv(time_both, status_both) ~ 1)
fit_s1   <- survfit(Surv(time_s1, status_s1)     ~ 1)

fit_both_df <- data.frame(time=c(0,fit_both$time),
                          surv=c(1,fit_both$surv),
                          sample="PL (combined)")
fit_s1_df   <- data.frame(time=c(0,fit_s1$time),
                          surv=c(1,fit_s1$surv),
                          sample="RS/PL (Sample I)")
plot2_df <- bind_rows(fit_both_df, fit_s1_df)
```

```r
p2 <- ggplot(plot2_df,
             aes(x=time, y=surv, color=sample, linetype=sample))
                +
  geom_step(linewidth=0.9) +
  geom_hline(yintercept=0.15, linetype="dotted",
             color="gray40") +
  scale_color_manual(values=c("PL (combined)"="steelblue",
                              "RS/PL (Sample I)"="firebrick")) +
  scale_linetype_manual(values=c("PL (combined)"="solid",
                                 "RS/PL (Sample I)"="dashed")) +
  labs(title="Two-sample example: PL vs RS",
       x="Time (years)", y=expression(hat(P)(t)),
       color=NULL, linetype=NULL) +
  theme_bw()

ggsave("Plot2_twosample.png", p2, width=6, height=4, dpi=150)

# RS vs KM monotonicity
L_8  <- ifelse(status_8 == 0, time_8, Inf)
rs_estimator <- function(time, status, L_obs) {
  t_grid <- sort(unique(time[status == 1]))
  res <- data.frame(time=numeric(), RS=numeric())
  for (t in t_grid) {
    N0_t <- sum(L_obs >= t)
    n_t  <- sum(time > t)
    if (N0_t > 0)
      res <- rbind(res, data.frame(time=t, RS=n_t/N0_t))
  }
  res
}
rs8 <- rs_estimator(time_8, status_8, L_8)

km8_plot <- data.frame(time=c(0,km8$time), surv=c(1,km8$surv),
                       method="KM (PL)")
rs8_plot <- data.frame(time=c(0,rs8$time), surv=c(1,rs8$RS),
                       method="RS")

p3 <- ggplot() +
  geom_step(data=km8_plot,
            aes(x=time, y=surv, color=method), linewidth=0.9) +
```

```
  geom_step(data=rs8_plot,
            aes(x=time, y=surv, color=method),
            linetype="dashed", linewidth=0.9) +
  scale_color_manual(values=c("KM (PL)"="steelblue",
                              "RS"="firebrick")) +
  labs(title="RS vs KM: monotonicity comparison",
       x="Time (months)", y=expression(hat(P)(t)),
       color=NULL) +
  theme_bw()

ggsave("Plot3_rs_vs_km.png", p3, width=6, height=4, dpi=150)
```

A.3 Nelson–Aalen Estimator

```
na_inc <- km8$d_j / km8$n_j
H_hat  <- cumsum(na_inc)
P_na   <- exp(-H_hat)

print(data.frame(time=km8$time, KM=round(km8$surv,4),
                 NA_exp=round(P_na,4), H_hat=round(H_hat,4)))

km8_na_df <- data.frame(time=c(0,km8$time), surv=c(1,km8$surv),
                        method="KM")
na8_df    <- data.frame(time=c(0,km8$time), surv=c(1,P_na),
                        method="Nelson-Aalen")
haz_df    <- data.frame(time=km8$time, H=H_hat)

p4a <- ggplot() +
  geom_step(data=km8_na_df,
            aes(x=time, y=surv, color=method), linewidth=0.9) +
  geom_step(data=na8_df,
            aes(x=time, y=surv, color=method),
            linetype="dashed", linewidth=0.9) +
  scale_color_manual(values=c("KM"="steelblue",
                              "Nelson-Aalen"="firebrick")) +
  labs(title="KM vs Nelson-Aalen",
       x="Time (months)", y="Survival", color=NULL) +
  theme_bw()

p4b <- ggplot(haz_df, aes(x=time, y=H)) +
  geom_step(color="firebrick", linewidth=0.9) +
```

```
  geom_point(color="firebrick", size=2.5) +
  labs(title=expression("Cumulative hazard " * hat(H)(t)),
       x="Time (months)", y=expression(hat(H)(t))) +
  theme_bw()

ggsave("Plot4_hazard.png", p4a+p4b, width=10, height=4, dpi=150)
```

A.4 Censoring Level Study and Confidence Bands

```
set.seed(42)
simulate_km_cens <- function(n, lambda_c, seed=42) {
  set.seed(seed)
  T_i        <- rexp(n, rate=1)
  L_i        <- rexp(n, rate=lambda_c)
  time_obs   <- pmin(T_i, L_i)
  status_obs <- as.integer(T_i <= L_i)
  fit        <- survfit(Surv(time_obs, status_obs) ~ 1)
  list(fit=fit, cens_rate=1-mean(status_obs),
       time_obs=time_obs, status_obs=status_obs)
}

sim_10 <- simulate_km_cens(200, 0.10/0.90, seed=42)
sim_50 <- simulate_km_cens(200, 1.00,      seed=42)
sim_80 <- simulate_km_cens(200, 4.00,      seed=42)

greenwood_curve <- function(time_obs, status_obs, label) {
  death_t <- sort(unique(time_obs[status_obs==1]))
  surv <- 1; gw <- 0
  res  <- data.frame(time=numeric(), var=numeric(),
                     level=character())
  for (tj in death_t) {
    nj   <- sum(time_obs >= tj)
    dj   <- sum(time_obs == tj & status_obs == 1)
    surv <- surv * (nj-dj)/nj
    gw   <- gw + dj/(nj*(nj-dj))
    res  <- rbind(res, data.frame(time=tj, var=surv^2*gw,
                                  level=label,
                                  stringsAsFactors=FALSE))
  }
  res
}
```

```
gw10   <- greenwood_curve(sim_10$time_obs, sim_10$status_obs, "~
   10%")
gw50   <- greenwood_curve(sim_50$time_obs, sim_50$status_obs, "~
   50%")
gw80   <- greenwood_curve(sim_80$time_obs, sim_80$status_obs, "~
   80%")
gw_all <- bind_rows(gw10, gw50, gw80)
gw_all$level <- factor(gw_all$level, levels=c("~10%","~50%","~80%
   "))

km_steps_fn <- function(fit, label) {
  data.frame(time=c(0,fit$time), surv=c(1,fit$surv),
             level=label, stringsAsFactors=FALSE)
}
km_cens_df <- bind_rows(km_steps_fn(sim_10$fit,"~10%"),
                        km_steps_fn(sim_50$fit,"~50%"),
                        km_steps_fn(sim_80$fit,"~80%"))
km_cens_df$level <- factor(km_cens_df$level,
                           levels=c("~10%","~50%","~80%"))
true_df <- data.frame(time=seq(0,5,0.01), surv=exp(-seq(0,5,0.01)
   ))
cols3 <- c("~10%"="steelblue","~50%"="darkorange","~80%"="
   firebrick")

p5a <- ggplot() +
  geom_line(data=true_df, aes(x=time,y=surv),
            linetype="dotted", color="black", linewidth=0.8) +
  geom_step(data=km_cens_df,
            aes(x=time,y=surv,color=level), linewidth=0.8) +
  scale_color_manual(values=cols3) +
  labs(title="KM under different censoring levels",
       x="Time", y="Survival", color="Censoring") +
  xlim(0,5) + theme_bw()

p5b <- ggplot(gw_all, aes(x=time, y=var, color=level)) +
  geom_step(linewidth=0.8) +
  scale_color_manual(values=cols3) +
  labs(title="Greenwood variance by censoring level",
       x="Time", y="Variance", color="Censoring") +
  xlim(0,5) + theme_bw()
```

```
ggsave("Plot5_variance.png", p5a+p5b, width=10, height=4, dpi
   =150)

# Confidence bands
fit8_ci <- survfit(Surv(time_8, status_8) ~ 1, conf.type="plain")
ci_df <- data.frame(time=fit8_ci$time, surv=fit8_ci$surv,
                    lower=fit8_ci$lower, upper=fit8_ci$upper)

p6 <- ggplot(ci_df, aes(x=time)) +
  geom_ribbon(aes(ymin=lower, ymax=upper),
              fill="steelblue", alpha=0.2) +
  geom_step(aes(y=surv), color="steelblue", linewidth=1.0) +
  geom_step(aes(y=lower), color="steelblue",
            linetype="dashed", linewidth=0.6) +
  geom_step(aes(y=upper), color="steelblue",
            linetype="dashed", linewidth=0.6) +
  labs(title="KM with Greenwood confidence bands",
       x="Time (months)", y=expression(hat(P)(t))) +
  ylim(0,1) + theme_bw()

ggsave("Plot6_km_ci.png", p6, width=6, height=4, dpi=150)

# Mean lifetime
km_times_vec <- c(0, km8$time)
km_steps_vec <- c(1, km8$surv)
L_rmst <- max(km8$time[km8$d_j > 0])
rmst <- 0
for (i in seq_len(length(km_times_vec)-1)) {
  if (km_times_vec[i+1] <= L_rmst)
    rmst <- rmst + km_steps_vec[i]*(km_times_vec[i+1]-km_times_
       vec[i])
}
cat(sprintf("Restricted mean lifetime: %.4f\n", rmst))

shade_df <- data.frame(
  xmin=km_times_vec[-length(km_times_vec)],
  xmax=km_times_vec[-1],
  ymax=km_steps_vec[-length(km_steps_vec)]
) %>% filter(xmax <= L_rmst)
step_df <- data.frame(time=km_times_vec, surv=km_steps_vec)
```

```
p7 <- ggplot() +
  geom_rect(data=shade_df,
            aes(xmin=xmin,xmax=xmax,ymin=0,ymax=ymax),
            fill="steelblue", alpha=0.25) +
  geom_step(data=step_df, aes(x=time,y=surv),
            color="steelblue", linewidth=1.0) +
  labs(title=expression("Mean lifetime: area under " * hat(P)(t))
     ,
       x="Time (months)", y=expression(hat(P)(t))) +
  ylim(0,1) + theme_bw()

ggsave("Plot7_mean_lifetime.png", p7, width=6, height=4, dpi=150)
```