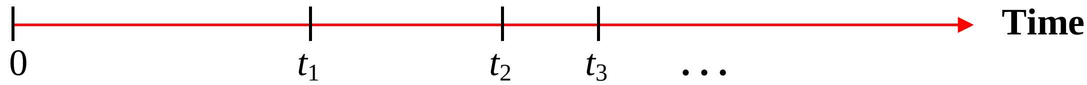
8.2 *Estimation:* Kaplan-Meier Product-Limit Formula



Let $t_1, t_2, t_3, \ldots$ denote the *actual* times of death of the n individuals in the cohort. Also let $d_1, d_2, d_3, \ldots$ denote the number of deaths that occur at each of these times, and let $n_1, n_2, n_3, \ldots$ be the corresponding number of patients remaining in the cohort. Note that $n_2 = n_1 - d_1$, $n_3 = n_2 - d_2$, etc. Then, loosely speaking, $S(t_2) = P(T > t_2)$ = "Probability of surviving beyond time t_2" depends *conditionally* on $S(t_1) = P(T > t_1)$ = "Probability of surviving beyond time t_1." Likewise, $S(t_3) = P(T > t_3)$ = "Probability of surviving beyond time t_3" depends *conditionally* on $S(t_2) = P(T > t_2)$ = "Probability of surviving beyond time t_2," etc. By using this recursive idea, we can iteratively build a numerical estimate $\hat{S}(t)$ of the true survival function $S(t)$. Specifically,

- For any time $t \in [0, t_1)$, we have $S(t) = P(T > t)$ = "Probability of surviving beyond time t" = 1, because no deaths have as yet occurred. Therefore, for all t in this interval, let $\hat{S}(t) = 1$.

  Recall (see 3.2): For any two events A and B, $\mathbf{P(A \text{ and } B) = P(A) \times P(B \mid A)}$.

  Let A = "survive to time t_1" and B = "survive from time t_1 to beyond some time t before t_2." Having *both* events occur is therefore equivalent to the event "A and B" = "survive to beyond time t before t_2," i.e., "$T > t$." Hence, the following holds.

- For any time $t \in [t_1, t_2)$, we have…

$$S(t) = P(T > t) = \underbrace{P(\text{survive in } [0, t_1))} \times \underbrace{P(\text{survive in } [t_1, t] \mid \text{survive in } [0, t_1))},$$

i.e,

$$\hat{S}(t) = 1 \times \frac{n_1 - d_1}{n_1}, \quad \text{or}$$

$$\hat{S}(t) = 1 - \frac{d_1}{n_1}.$$ Similarly,

- For any time $t \in [t_2, t_3)$, we have…

$$S(t) = P(T > t) = \underbrace{P(\text{survive in } [t_1, t_2))} \times \underbrace{P(\text{survive in } [t_2, t] \mid \text{survive in } [t_1, t_2))},$$

i.e,

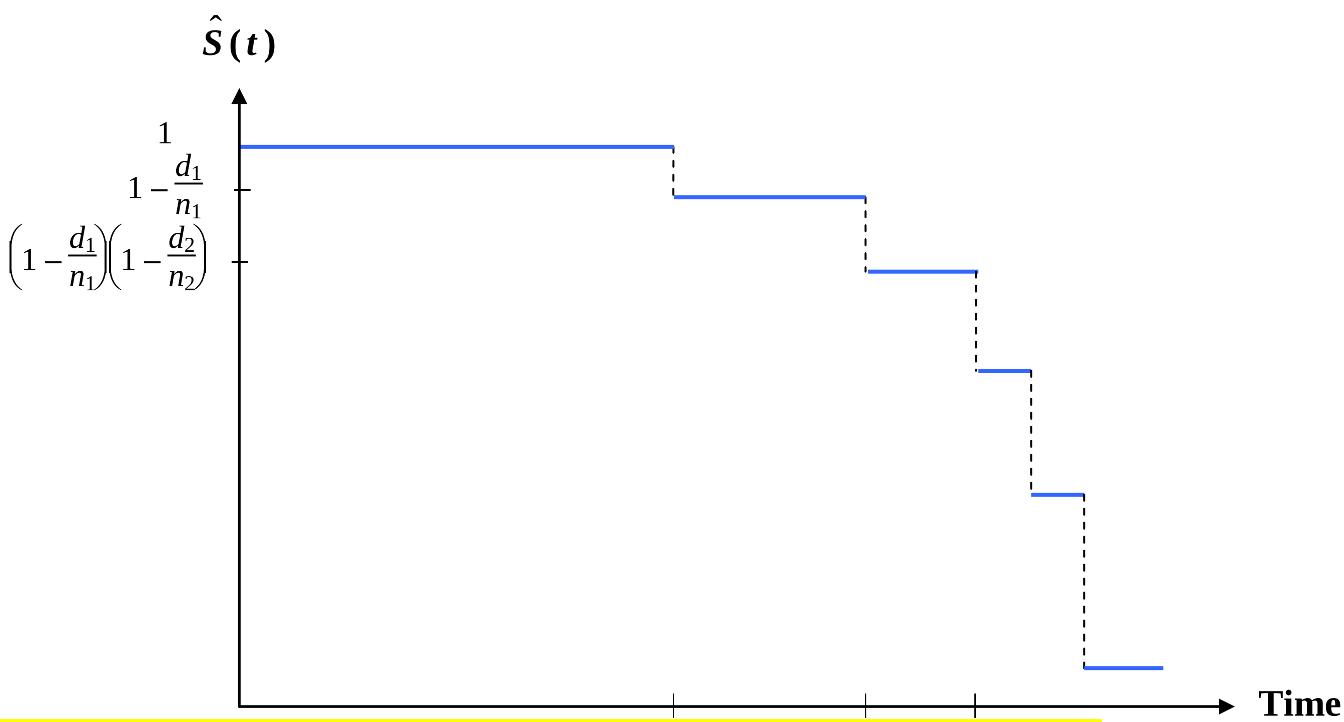
$$\hat{S}(t) = \left(1 - \frac{d_1}{n_1}\right) \times \frac{n_2 - d_2}{n_2}, \quad \text{or}$$

$$\hat{S}(t) = \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right),$$ etc.

In general, for $t \in [t_j, t_{j+1})$, $j = 1, 2, 3, \ldots$, we have…

$$\boxed{\hat{S}(t) = \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right)\cdots\left(1 - \frac{d_j}{n_j}\right) = \prod_{i=1}^{j}\left(1 - \frac{d_i}{n_i}\right).}$$

This is known as the **Kaplan-Meier estimator** of the survival function $S(t)$. (Theory developed in 1950s, but first implemented with computers in 1970s.) Note that it is *not continuous*, but only *piecewise-continuous* (actually, *piecewise-constant*, or "step function").



times of death:	0	t_1	t_2	t_3	. . .
# deaths:	0	d_1	d_2	d_3	. . .
# survivors:	$n_1 = n - 0$	$n_2 = n_1 - d_1$	$n_3 = n_2 - d_2$		. . .

<u>*Comment*</u>: The Kaplan-Meier estimator $\hat{S}(t)$ can be regarded as a point estimate of the survival function $S(t)$ at any time t. In a manner similar to that discussed in 7.2, we can construct 95% confidence intervals around each of these estimates, resulting in a pair of **confidence bands** that brackets the graph. To compute the confidence intervals, **Greenwood's Formula** gives an asymptotic estimate of the standard error of $\hat{S}(t)$ *for large groups*.

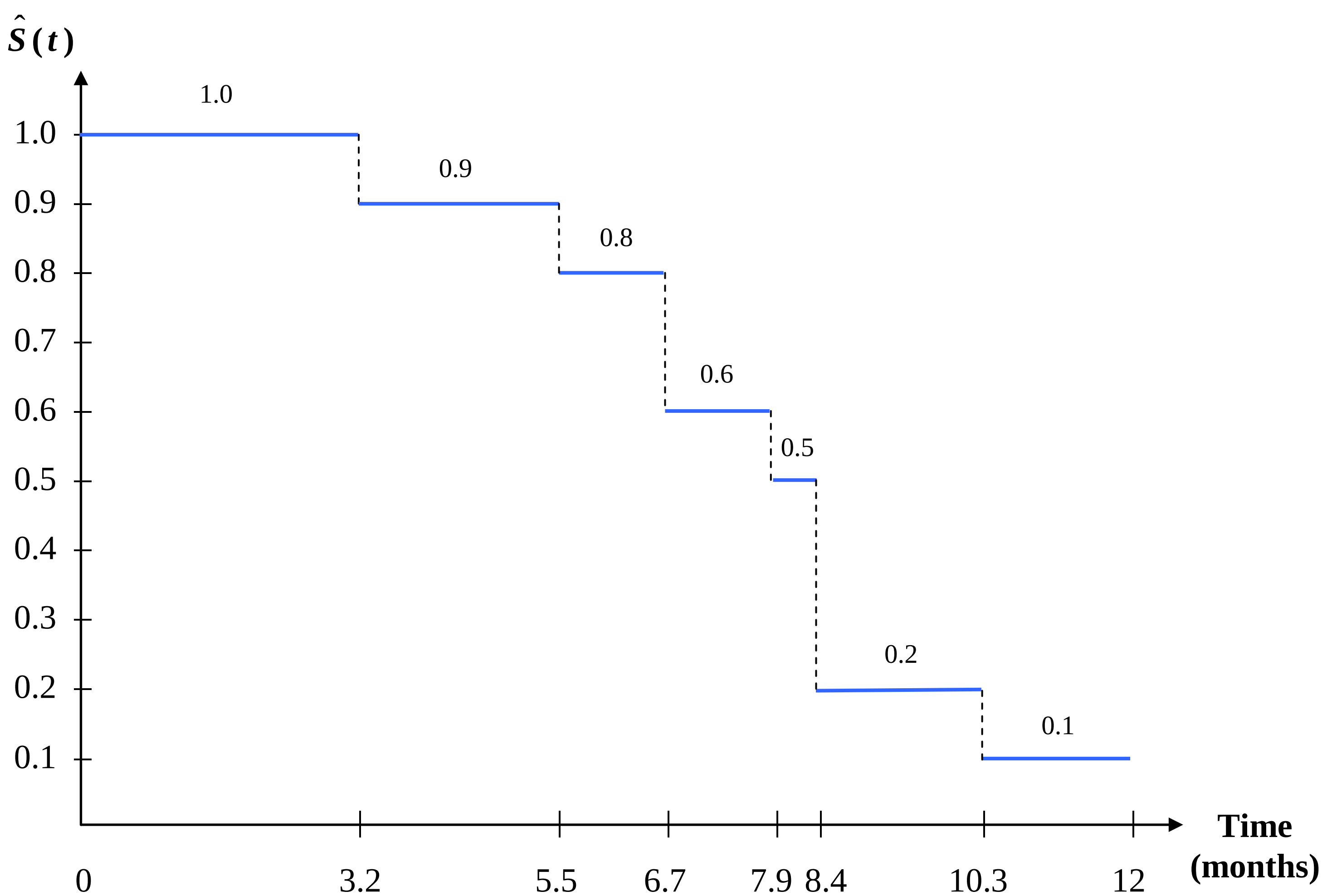
Example (cont'd): Twelve-month cohort study of $n = 10$ patients

Patient	t_i (months)
1	3.2
2	5.5
3	6.7
4	6.7
5	7.9
6	8.4
7	8.4
8	8.4
9	10.3
10	alive

→

Interval $[t_i, t_{i+1})$	n_i = # patients at risk at time t_i^-	d_i = # deaths at time t_i	$1 - \frac{d_i}{n_i}$	$\hat{S}(t)$
[0, 3.2)	10	0	1.00	1.0
[3.2, 5.5)	10 – 0 = 10	1	0.90	0.9
[5.5, 6.7)	10 – 1 = 9	1	0.89	0.8
[6.7, 7.9)	9 – 1 = 8	2	0.75	0.6
[7.9, 8.4)	8 – 2 = 6	1	0.83	0.5
[8.4, 10.3)	6 – 1 = 5	3	0.40	0.2
[10.3, 12)	5 – 3 = 2	1	0.50	0.1
Study Ends	2 – 1 = 1	0	1.00	0.1

t_i^- denotes a time *just prior* to t_i



Exercise: Prove algebraically that, <u>assuming no censored observations</u> (as in the preceding example), the Kaplan-Meier estimator can be written simply as $\hat{S}(t) = \dfrac{n_{i+1}}{n}$ for $t \in [t_i, t_{i+1})$, $i = 0, 1, 2, \ldots$ *<u>Hint</u>:* Use mathematical induction; recall that $\boldsymbol{n_{i+1} = n_i - d_i}$.

In light of this, now assume that the data consists of censored observations as well, so that $\boldsymbol{n_{i+1} = n_i - d_i - c_i}$.

<u>Example (cont'd)</u>:

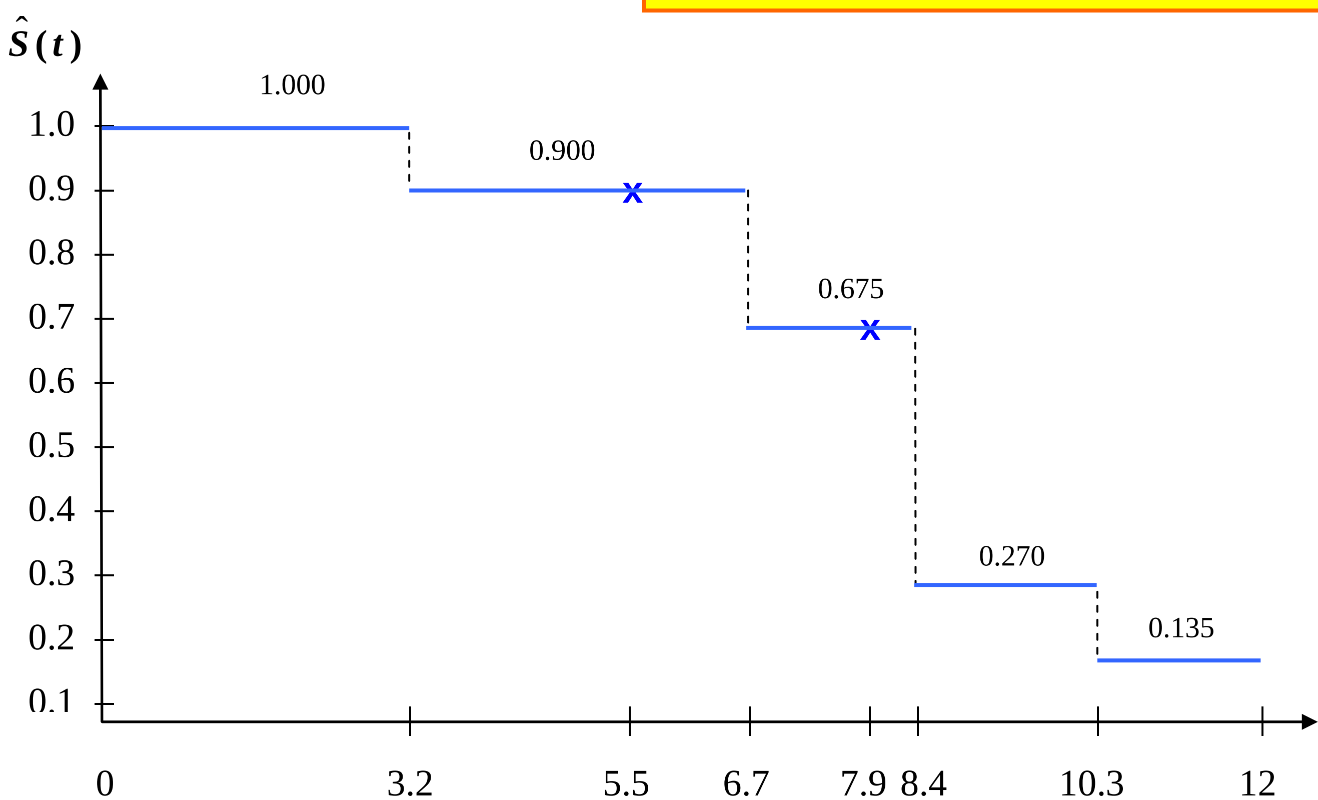
Patient	t_i (months)
1	3.2
2	5.5*
3	6.7
4	6.7
5	7.9*
6	8.4
7	8.4
8	8.4
9	10.3
10	alive

*censored

→

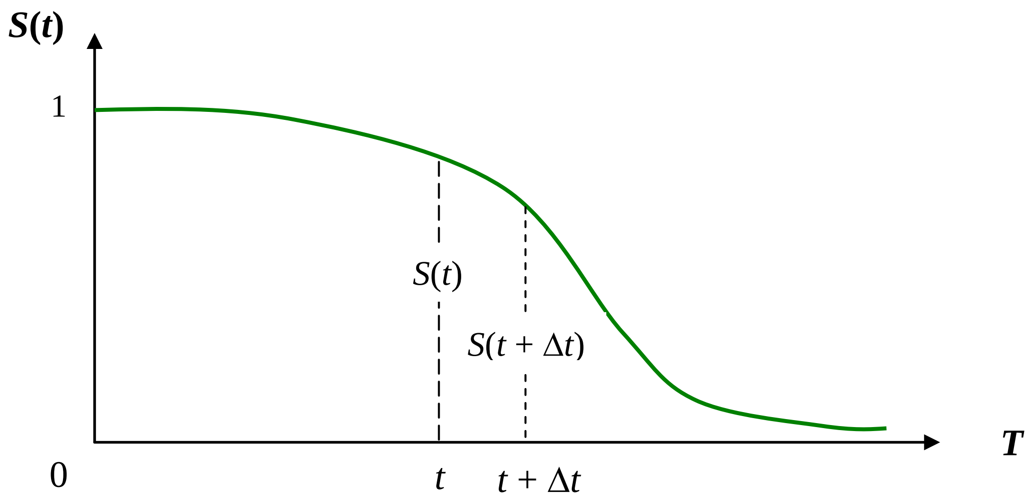
Interval $[t_i, t_{i+1})$	n_i = # at risk at time t_i^-	d_i = # deaths	c_i = # censored	$1 - \dfrac{d_i}{n_i}$	$\hat{S}(t)$
[0, 3.2)	10	0	0	1.00	1.000
[3.2, 6.7)	10 – 0 – 0 = 10	1	1	0.90	0.900
[6.7, 8.4)	10 – 1 – 1 = 8	2	1	0.75	0.675
[8.4, 10.3)	8 – 2 – 1 = 5	3	0	0.40	0.270
[10.3, 12)	5 – 3 – 0 = 2	1	0	0.50	0.135
Study Ends	2 – 1 – 0 = 1	0	0	1.00	0.135

Exercise: What would the corresponding changes be to the Kaplan-Meier estimator if Patient 10 died at the very end of the study?



Hazard Functions

Suppose we have a survival function $S(t) = P(T > t)$, where T = survival time, and some $\Delta t > 0$. We wish to calculate the <u>conditional</u> probability of survival to the later time $t + \Delta t$, *given* survival to time t.



$$P(\underbrace{\text{Survive in } [t, t+\Delta t)}_{t \le T < t + \Delta t} \mid \underbrace{\text{Survive after } t}_{T > t}) = \frac{P(t \le T < t + \Delta t)}{P(T > t)} = \frac{S(t) - S(t + \Delta t)}{S(t)}.$$

Therefore, dividing by Δt,

$$\frac{P(t \le T < t + \Delta t \mid T > t)}{\Delta t} = \frac{-1}{S(t)} \, \frac{S(t + \Delta t) - S(t)}{\Delta t}.$$

Now, take the limit of both sides as $\Delta t \to 0$:

$$h(t) = \frac{-1}{S(t)} S'(t) = -\frac{d\,[\ln S(t)]}{dt} \quad \Leftrightarrow \quad S(t) = e^{-\int_0^t h(u)\,du}$$

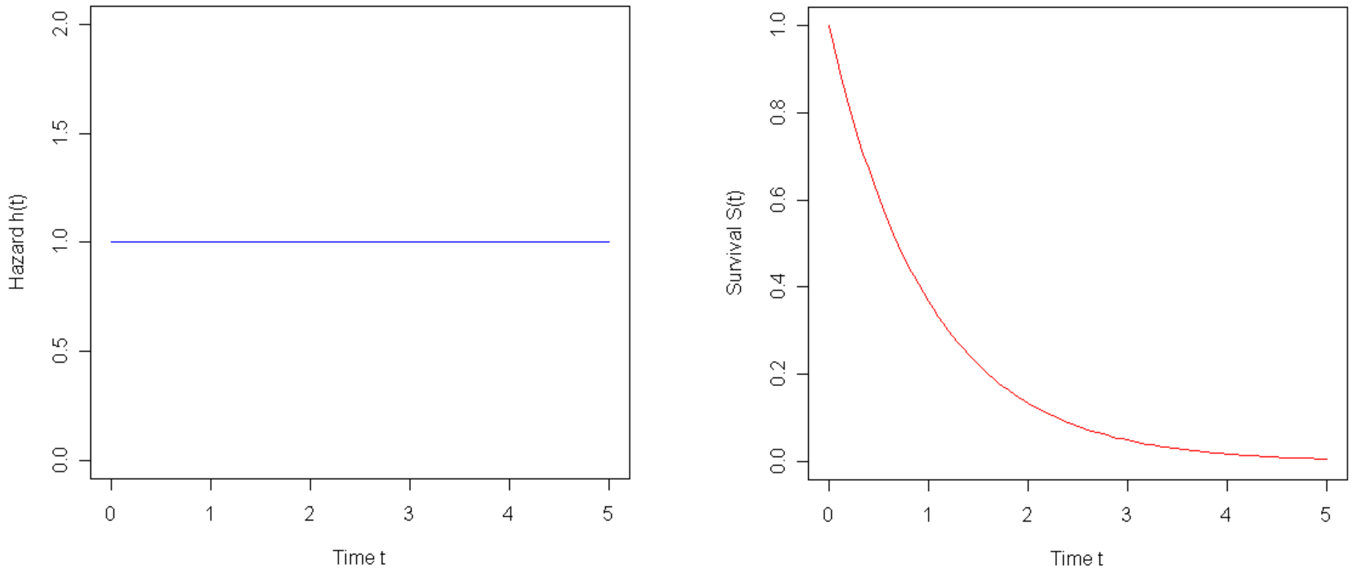
This is the **hazard function** (or **hazard rate**, **failure rate**), and *roughly* characterizes the "instantaneous probability" of dying at time t, in the above mathematical "limiting" sense. It is always ≥ 0 (Why? <u>*Hint*</u>*:* What signs are $S(t)$ and $S'(t)$, respectively?), but can be > 1, hence is not a true probability in a mathematically rigorous sense.

Exercise: Suppose two hazard functions are linearly combined to form a third hazard function: $c_1 h_1(t) + c_2 h_2(t) = h_3(t)$, for any constants $c_1, c_2 \ge 0$. What is the relationship between their corresponding log-survival functions $\ln S_1(t)$, $\ln S_2(t)$, and $\ln S_3(t)$?

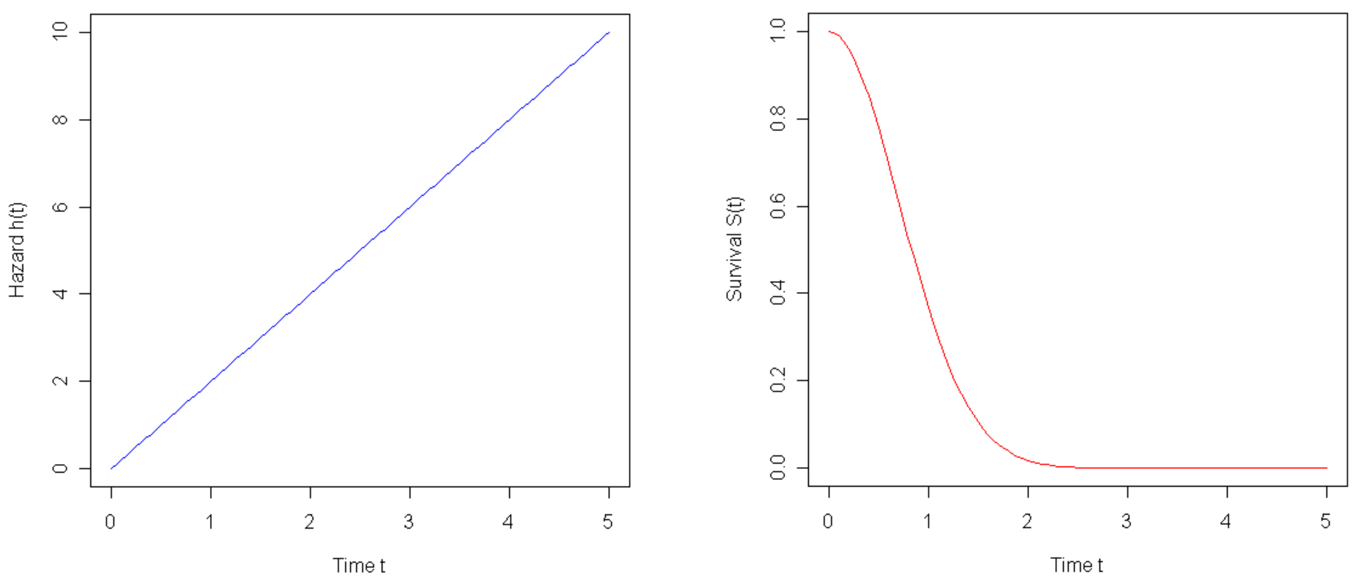
Its integral, $\int_0^t h(u)\,du$, is the **cumulative hazard** rate – denoted $H(t)$ – and increases (since $H'(t) = h(t) \ge 0$). Note also that $H(t) = -\ln S(t)$, and so $\hat{H}(t) = -\ln \hat{S}(t)$.

Examples: **(Also see last page of 4.2!)**

- If the hazard function is ***constant*** for $t \geq 0$, i.e., $h(t) \equiv \alpha > 0$, then it follows that the survival function is $S(t) = e^{-\alpha t}$, i.e., the **exponential model**. Shown here is $\alpha = 1$.



- More realistically perhaps, suppose the hazard takes the form of a more general **power function**, i.e., $h(t) = \alpha\beta t^{\beta-1}$, for "scale parameter" $\alpha > 0$, and "shape parameter" $\beta > 0$, for $t \geq 0$. Then $S(t) = e^{-\alpha t^{\beta}}$, i.e., the **Weibull model**, an extremely versatile and useful model with broad applications to many fields. The case $\alpha = 1$, $\beta = 2$ is illustrated below.



Exercise: Suppose that, for argument's sake, a population is modeled by the decreasing hazard function $h(t) = \dfrac{1}{t+c}$ for $t \geq 0$, where $c > 0$ is some constant. Sketch the graph of the survival function $S(t)$, and find the median survival time.