Hazard Rate Functions

Examples via Phase-Type Distributions

Definition. If T is an *absolutely continuous* non-negative random variable, its hazard rate function $h(t),\ t \geq 0$, is defined by

$$h(t) \ = \ \frac{f(t)}{S(t)}, \quad t \geq 0,$$

where $f(t)$ is the density of T and $S(t)$ is the survival function:
$S(t) = \int_t^\infty f(u)du = P\{T > t\}$.
Note that $\mathrm{P}\{T \leq t + \Delta \,|\, T > t\} \approx h(t) \cdot \Delta$.

If T is a *discrete* non-negative random variable that takes values $t_1 < t_2 < \ldots$ with corresponding probabilities $\{p_i,\ i \geq 1\}$, then its hazard-sequence $\{h(t_i)\}$ is defined by

$$h(t_i) \ = \ \frac{p_i}{\sum_{j \geq i} p_j} \ = \ \frac{p_i}{S(t_i-)}, \quad i \geq 1.$$

Note that $\mathrm{P}\{T = t_i \,|\, T > t_{i-1}\} = h(t_i)$.

Why estimate the hazard rates of service times or patience?

- The hazard rate is a *dynamic* characteristic of a distribution.
  (One of the main goals of our note is to demonstrate this statement).
- The hazard rate is a more precise "fingerprint" of a distribution than the cumulative distribution function, the survival function, or density (for example, unlike the density, its tail need not converge to zero; the tail can increase, decrease, converge to some constant etc.)
- The hazard rate provides a tool for comparing the tail of the distribution in question against some "benchmark": the exponential distribution, in our case.
- The hazard rate arises naturally when we discuss "strategies of abandonment", either rational (as in Mandelbaum & Shimkin) or ad-hoc (Palm).

Why do phase-type distributions constitute a convenient class of models for service times ? As discussed in class:

- dense;
- structurally informative;
- meta theorem: homogeneous unpaced human service\task durations are exponential.

Why is it convenient to illustrate the concept of hazard rate via phase-type examples?

- Small number of phases suffices to illustrate the various modes of hazard-rate behavior.
- Simple intuitive explanations of hazard-rate patterns can be demonstrated. (In contrast, try to develop intuition for the hazard rates of normal or lognormal random variables!)

Limitations: *Which patterns of hazard rate cannot be illustrated by phase-type distributions?*
Answer. We shall see below that the hazard rate of a phase-type distribution has a limit as $t \to \infty$. This limit can be shown to be neither 0 nor ∞. Hence, phase-type distributions can not belong to heavy-tail distributions with hazard rates that converge to zero (recall Pareto) or to distributions with hazard rates that converge to infinity (recall the Normal distribution).

Hazard-rate representation for Phase-Type distributions

Let T be phase-type distributed. Animate T by an absorbing Markov jump-process $X = \{X_t,\ t \geq 0\}$, on a finite state-space S, with an absorbing state Δ. Then the hazard-rate function of T, $h_T(t)$, has the representation:

$$\boxed{h_T(t) \ = \ \sum_{i \in S} q_{i\Delta} \mathrm{P}\{X_t = i | T > t\}, \quad t \geq 0}$$

where $q_{i\Delta}$ is the transition (absorption) rate from state i, that is

$$\mathrm{P}\{X_{t+\epsilon} = \Delta | X_t = i\} \ = \ q_{i\Delta} \cdot \epsilon + o(\epsilon),\ i \in S.$$

The representation above demonstrates the *dynamic approach* to the hazard rate of phase-type distributions: the hazard rate at time t is determined by the conditional distribution of the underlying Markov process X.

For convenience, denote

$$P_i(t) \ = \ \mathrm{P}\{X_t = i | T > t\}, \quad t \geq 0, \quad i \in S.$$

Remark. As $t \uparrow \infty$, the functions $\{P_i(t),\ i \in S\}$ converge to, what is called, the *quasi-stationary* distribution of X. It can be expressed in terms of eigen-values related to the matrix Q (generator of X, restricted to S), and gives rise to a representation for the limit

$$h_T(\infty) = \sum_{i \in S} q_{i\Delta} P_i(\infty)\ .$$

In the examples that follow, $P_i(\infty)$ will be calculated directly.

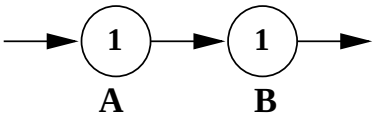
General description of our (static) simulation.

We consider four examples of phase-type distributions. For each example, 10,000 independent realizations were simulated in Excel. The theoretical hazard rates were plotted and compared against estimates of the hazard rate, based on the simulation data. (The method used for hazard rate estimation is described in the Technical Appendix, at the end of the handout.)

In the examples below, the probabilities $P_i(t)$ for all non-absorbing states $i \in S$ were calculated explicitly. We then tried to illuminate the connection between $P_i(t)$ and the hazard rate, based on the representation above.

Example 1. Increasing hazard rate.

Assume that the service consists of several exponential phases in sequence (e.g. first, a customer issues a request, then a server provides service). As a simple service-time model, consider the **Erlang (Gamma) distribution** with two phases, each distributed exp(1):

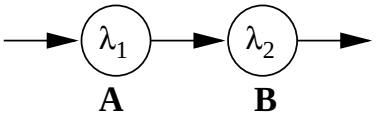


Survival function	Density	Hazard rate	Expectation
$te^{-t} + e^{-t}$	te^{-t}	$\frac{t}{t+1}$	2

$P_A(t)$	$P_B(t)$	Hazard rate
$\frac{1}{t+1}$	$\frac{t}{t+1}$	$P_B(t)$

Intuition to Figure 1. The hazard rate is close to zero near zero since the probability to complete two exponential tasks in a short time is negligible. As time increases, the probability $P_B(t)$ that the service is at the second phase increases to one. Therefore, the hazard rate converges to the parameter of the second phase.

Remark. One must actually be careful with the above interpretation. Indeed, assume that the service time has the representation :



for arbitrary $\lambda_1, \lambda_2 > 0$, $\lambda_1 \neq \lambda_2$. Then it can be shown that

$$P_A(t) \;=\; \frac{e^{-\lambda_1 t}}{\frac{\lambda_2}{\lambda_2-\lambda_1}e^{-\lambda_1 t} - \frac{\lambda_1}{\lambda_2-\lambda_1}e^{-\lambda_2 t}} \longrightarrow \left(1 - \frac{\lambda_1}{\lambda_2}\right)^+ , \quad \text{as } t \to \infty \; ;$$

$$P_B(t) \;=\; \frac{\lambda_1(e^{-\lambda_1 t} - e^{-\lambda_2 t})}{\lambda_2 e^{-\lambda_1 t} - \lambda_1 e^{-\lambda_2 t}} \longrightarrow \left(1 \wedge \frac{\lambda_1}{\lambda_2}\right) , \quad \text{as } t \to \infty \; .$$

The hazard rate

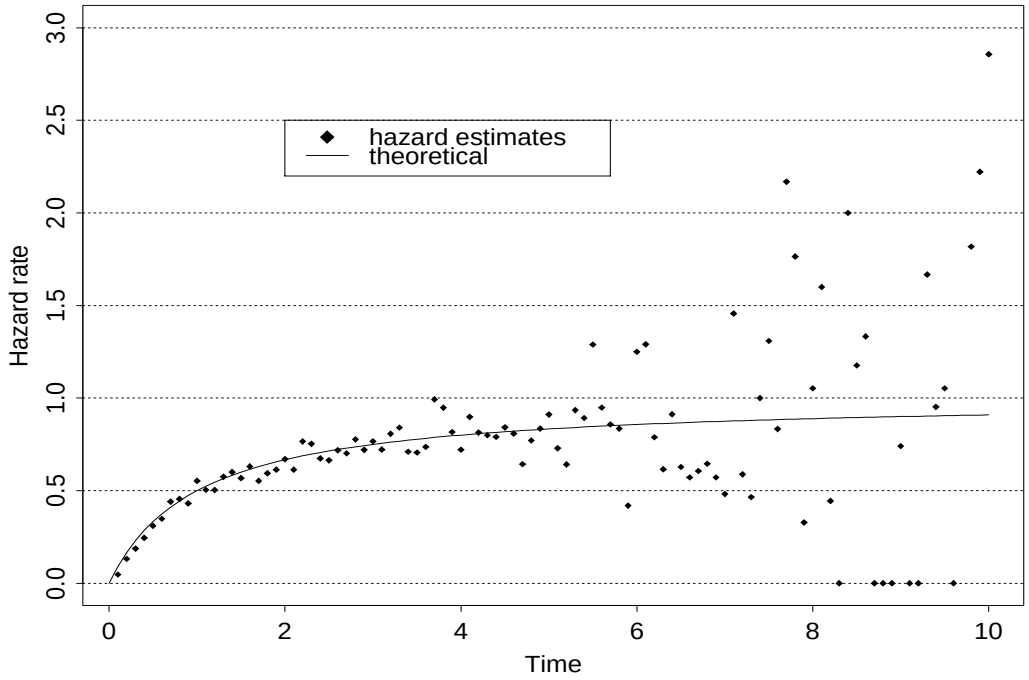
$$h(t) = \lambda_2 P_B(t) \longrightarrow \lambda_1 \wedge \lambda_2 \; , \quad \text{as } t \to \infty \; ,$$

converges to the minimum of the two rates. (Why?) (This statement can be proved for generalized-Erlang distribution with any number of phases, namely $h(\infty) = \min\{\lambda_i\}$.)

Remarks.

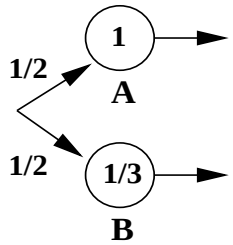
- Using L'Hopital rule one can obtain $P_B(t) = \frac{\lambda_1 t}{1+\lambda_1 t}$ for $\lambda_1 = \lambda_2$.
- Differentiating $P_B(t)$ shows that this function is strictly increasing for any λ_1, λ_2.

Figure 1
Example of increasing hazard rate
Erlang distribution



Example 2. Decreasing hazard rate.

There may be several types of customers, each with an exponential service time. **The hyper-exponential distribution** is a natural model in this case. For example, consider the two-phase case with rates 1 and 1/3 respectively:

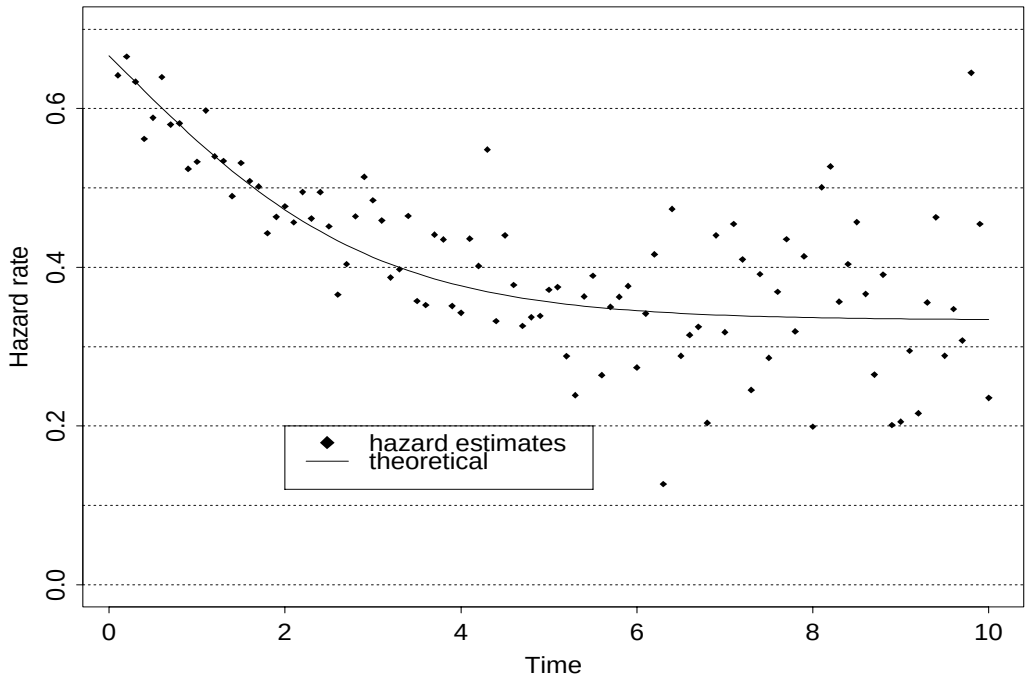


Survival function	Density	Hazard rate	Expectation
$\frac{1}{2}(e^{-t} + e^{-t/3})$	$\frac{1}{2}e^{-t} + \frac{1}{6}e^{-t/3}$	$\frac{\frac{1}{3} + e^{-2t/3}}{1 + e^{-2t/3}}$	2

$P_A(t)$	$P_B(t)$	Hazard rate
$\frac{e^{-2t/3}}{1+e^{-2t/3}}$	$\frac{1}{1+e^{-2t/3}}$	$P_A(t) + \frac{1}{3}P_B(t)$

Intuition to Figure 2. The hazard rate near zero is close to 2/3: the average of the two rates of the individual phases. (This is clear by our representation.) If the service has not been completed for a long time, the probability $P_B(t)$ that the phase with the longest expectation (rate 1/3) had been "chosen" converges to one. Hence, the hazard rate converges to 1/3.

Figure 2
Example of decreasing hazard rate
Hyperexponential distribution



General hyperexponential distribution.
Consider a hyperexponential distribution with k phases, initial probabilities $p_1, \ldots, p_k$ and rates $\lambda_1, \ldots, \lambda_k$. Then the survival function $S(t)$ and density $f(t)$ are equal to

$$S(t) \;=\; \sum_{i=1}^{k} p_i e^{-\lambda_i t}; \qquad f(t) \;=\; \sum_{i=1}^{k} \lambda_i p_i e^{-\lambda_i t} \;.$$

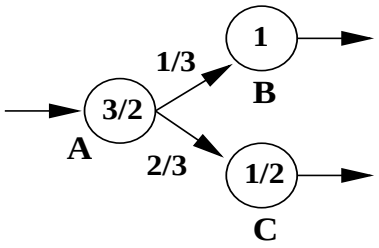
The function $h(t)$ is strictly decreasing. This can be proved by noting that, in fact, $S(t) = E[e^{-\lambda t}]$ and $f(t) = E[\lambda e^{-\lambda t}]$, where λ is thought of as a random variable that takes the values $\{\lambda_i\}$ with probabilities $\{p_i\}$. Proving monotonicity then reduces to applying Cauchy-Shwartz to the derivative of $h(t)$.

It is also straightforward to show that the limit of the hazard rate:

$$\lim_{t\to\infty} h(t) \;=\; \min\{\lambda_1, \ldots, \lambda_k\} \;.$$

Example 3. "Hump" hazard rate.

Assume that service consists of two parallel exponential tasks (say, one is performed by a human server and another by a computer). Service ends when both tasks have been completed. Then the total service duration is the **maximum** of two exponential random variables. Assume rates 1 and 1/2. This distribution can be represented in the following phase-type form:



Note that the duration of phase A is the minimum between the tasks' durations.

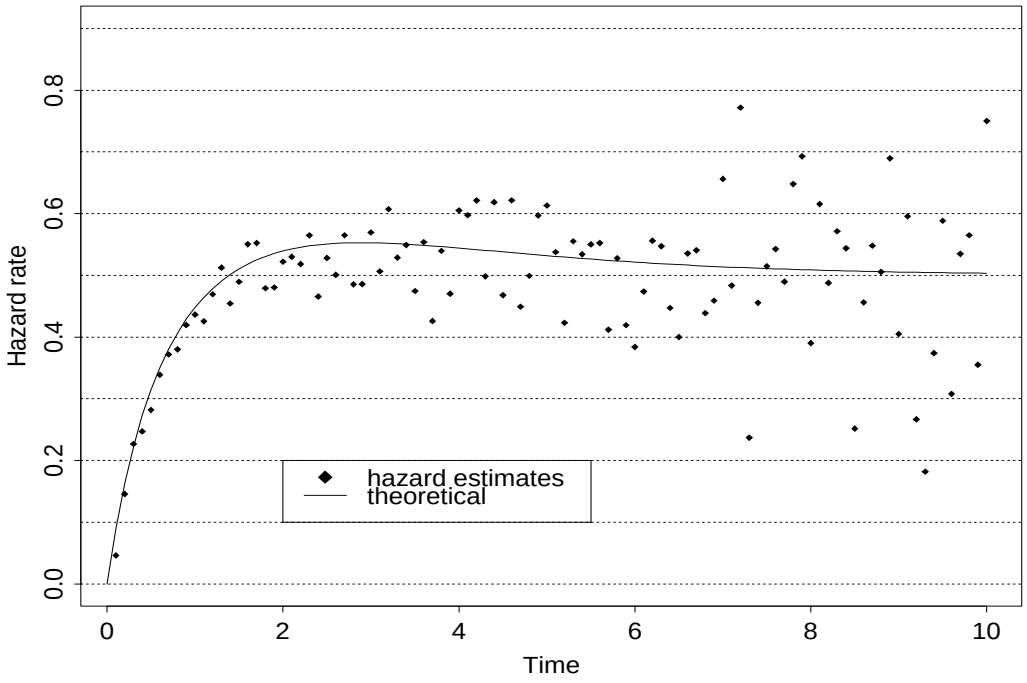
Survival function	Density	Hazard rate	Expectation
$e^{-t/2} + e^{-t} - e^{-3t/2}$	$\frac{1}{2}e^{-t/2} + e^{-t} - \frac{3}{2}e^{-3t/2}$	$\frac{\frac{1}{2}e^{-t/2} + e^{-t} - \frac{3}{2}e^{-3t/2}}{e^{-t/2} + e^{-t} - e^{-3t/2}}$	$\frac{7}{3}$

$P_A(t)$	$P_B(t)$	$P_C(t)$	Hazard rate
$\frac{e^{-3t/2}}{e^{-t/2} + e^{-t} - e^{-3t/2}}$	$\frac{e^{-t} - e^{-3t/2}}{e^{-t/2} + e^{-t} - e^{-3t/2}}$	$\frac{e^{-t/2} - e^{-3t/2}}{e^{-t/2} + e^{-t} - e^{-3t/2}}$	$P_B(t) + \frac{1}{2}P_C(t)$

Intuition to Figure 3. The hazard rate at zero is equal to zero as in Example 1 (both $P_B(t)$ and $P_C(t)$ are equal to zero at $t = 0$).
Then the hazard rate increases until some maximum. ($P_A(t)$ becomes small: there is a high

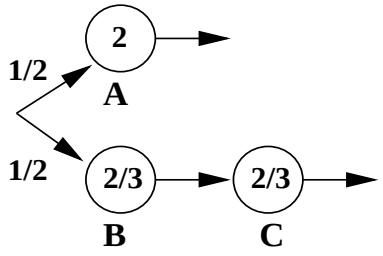
probability that the shortest task is over; $P_B(t)$ is relatively large.)
Finally, the hazard rate decreases to the minimal rate, as in Example 2 ($P_C(t)$ converges to one: it is more and more likely that the maximum is attained at the phase with the longest expectation).

Figure 3
Example of "hump" hazard rate
Maximum of exponential random variables



Example 4. "Bathtub" hazard rate.

Assume that there are two types of customers. Customers of the first type are forwarded to an alternative server after a short exponential check-up. The service of the second type can be expressed by the Erlang model of Example 1. In this case, we model the service time using the **Erlang mixture** of an exponential random variable (in our example, the rate is equal to 2) and the Erlang random variable (two phases, each with rate 2/3).



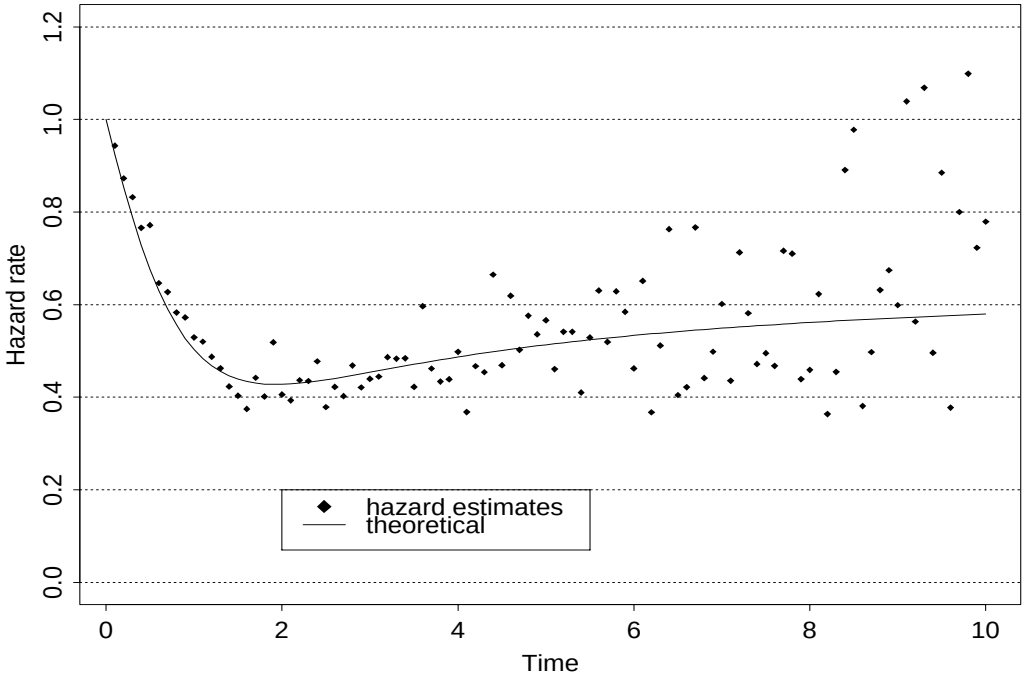
Survival function	Density	Hazard rate	Expectation
$\left(\frac{1}{2}+\frac{1}{3}t\right)e^{-2t/3}+\frac{1}{2}e^{-2t}$	$\frac{2}{9}te^{-2t/3}+e^{-2t}$	$\frac{\frac{2}{9}te^{-2t/3}+e^{-2t}}{\left(\frac{1}{2}+\frac{1}{3}t\right)e^{-2t/3}+\frac{1}{2}e^{-2t}}$	$\frac{7}{4}$

$P_A(t)$	$P_B(t)$	$P_C(t)$	Hazard rate
$\frac{\frac{1}{2}e^{-2t}}{\frac{1}{2}e^{-2t}+\left(\frac{1}{2}+\frac{1}{3}t\right)e^{-2t/3}}$	$\frac{\frac{1}{2}e^{-2t/3}}{\frac{1}{2}e^{-2t}+\left(\frac{1}{2}+\frac{1}{3}t\right)e^{-2t/3}}$	$\frac{\frac{1}{3}te^{-2t/3}}{\frac{1}{2}e^{-2t}+\left(\frac{1}{2}+\frac{1}{3}t\right)e^{-2t/3}}$	$2P_A(t)+\frac{2}{3}P_C(t)$

Intuition to Figure 4. Since $P_A(0) = 1/2$, the hazard rate at zero is equal to half of the service rate at phase A. As t increases, it becomes clear that we deal with the second service type (sequential phases B and C). The behavior of the hazard rate then becomes Erlang-like (see Example 1).

Remark. The "bathtub" hazard rate is widely applicable in reliability theory.

Figure 4
Example of "bathtub" hazard rate
Erlang mixture



Technical Appendix

Estimation of the Hazard Rate

The hazard rate is assumed to be constant on successive time intervals of length 0.1 between 0 and 10 (100 intervals overall). Formally, interval j is $\left(\frac{j-1}{10}, \frac{j}{10}\right]$, $j = 1, 2, \ldots, 100$.

The hazard estimate $\hat{h}_j$ for interval number j is calculated using the following formula:

$$\hat{h}_j \;=\; \frac{d_j}{b_j\left(r_{j-1} - \frac{1}{2}d_j\right)}\ ,$$

where
d_j = *number of events* (service terminations) in interval number j;
r_{j-1} = *number at risk* at the beginning of interval number j (number of services that have not terminated yet at time $\frac{j-1}{10}$);
b_j = length of interval number j (0.1 for all intervals, in our case).

The following provides some intuition for the above formula:
Let n denote the sample size. Then $\frac{d_j}{b_j \cdot n}$ is a reasonable estimate of the average density in interval number j and $\frac{r_{j-1} - 0.5 \cdot d_j}{n}$ is an approximation for the survival function in the center of this interval.

Technical Remark:
Proof Outline for the Representation:

Recall that T is the absorption time of X (first hitting time of the absorbing state $\triangle$); $f_T(t)$ is the density of T and $h_T(t)$ its hazard rate function.

$$\boxed{h_T(t) \;=\; \sum_{i \in S} q_{i\Delta} \mathrm{P}\{X_t = i | T > t\}, \quad t \geq 0}$$

A direct proof of the representation can be based on the relations:

$$\{X_{t+\epsilon} = \Delta\} \;=\; \{T \leq t + \epsilon\}\ , \quad \forall \epsilon > 0; \quad \{X_t \neq \triangle\} = \{T > t\}\ , \ \forall t > 0;$$

$$\{X_t = i\} \;\subset\; \{T > t\}\ , \quad \forall i \in S;$$

These are applied to show that

$$\begin{aligned} f_T(t) \cdot \epsilon + o(\epsilon) \;&=\; \mathrm{P}\{t < T \leq t + \epsilon\} \\ &=\; \sum_{i \in S} \mathrm{P}\{X_{t+\epsilon} = \Delta | X_t = i\} \mathrm{P}\{X_t = i\} \\ &=\; [\sum_{i \in S} q_{i\Delta} P\{X_t = i\}] \cdot \epsilon + o(\epsilon)\ . \end{aligned}$$

Technical Remark: The "truth" behind the above representation is
Dynkin's Formula:

$$\mathrm{E}f(X_\tau) - \mathrm{E}f(X_0) \;=\; \mathrm{E}\int_0^\tau Qf(X_s)ds \;,$$

for any stopping time τ. (Recall that Q is the generator of X.)
Applying Dynkin's formula at $\tau = t$, for $f(x) = 1_{\{\Delta\}}(x)$, yields (after some algebra)

$$\mathrm{P}\{T \le t\} \;=\; \sum_{i\in S} q_{i\Delta}\int_0^t P\{X_u = i\}du \;,$$

which is equivalent to our representation.

Technical Remark:
Martingale Representations of the Hazard Rate

Let T be any non-negative random variable. Then the stochastic process

$$M_t \;=\; 1_{T\le t} - \int_0^t 1_{u<T}dH(u) \;=\; 1_{T\le t} - \int_0^{t\wedge T} dH(u) \;=\; 1_{T\le t} - H(t\wedge T), \quad t \ge 0,$$

is a martingale. Here $H(\cdot)$ is the *integrated* hazard rate of T.

In the absolutely-continuous case,

$$H(t) \;=\; \int_0^t h(u)du \;=\; -\ln S(t) \;,$$

in which we have assumed $F(0) = 0$. Hence,

$$M_t \;=\; 1_{T\le t} + \ln S(t\wedge T), \quad t \ge 0.$$

(We ignore here technicalities associated with $S(t) = 0$.)

In the general case,

$$H(t) \;=\; \int_0^t \frac{1}{S(u-)}dF(u), \quad t \ge 0,$$

which expresses the hazard in terms of the distribution. This is equivalent to
$dF(t) = S(t-)dH(t)$, from which one deduces that

$$S(t) \;=\; 1 - F(t) \;=\; 1 - \int_0^t S(u-)dH(u), \quad t \ge 0,$$

namely $dS(t) = -S(t-)dH(t)$, with $S(0) = 1$. We have obtained an equation for S in terms of H, the unique solution for which turns out to be

$$S(t) \;=\; 1 - F(t) \;=\; \prod_{0\le u\le t}[1 - \Delta H(u)]\exp(-\int_0^t dH_c(u)) \;.$$

Here $\Delta H(u) = H(u) - H(u-)$ (the jumps of H) and $H_c(t) = H(t) - \sum_{0\le u\le t}\Delta H(u)$ (the continuous part of H).

References.

Daley, D.J., Vere-Jones, D.
"*An Introduction to the Theory of Point Processes*", Springer-Verlag, 1988. (See Sections 4.6, 13.2).

Last, G., Brandt, A.
"*Marked Point Processes on the Real Line: The Dynamic Approach*", Springer-Verlag, 1995. (See Appendix A5, Section 1.2, Section 1.6).

Aven, T., Jensen, U.
"*Stochastic Models in Reliability*", Springer-Verlag, 1999. (Section 2.2 is relevant, about Basci Notions of Aging.)