Statistical Analysis of Reliability and Survival data

Lecture notes

Academic year: 2019 - 2020

Example: Turbine disk

(Nelson (1982))

In an industrial trial, the lifetime is followed for 206 turbine disks. The data are observed in 100-hour intervals and after each inspection the number of failed disks is recorded.

Extra difficulty: the experiment was stopped after 2100 hours and some disks were still running at that time (censored).

Example: Computer

A computer is build up out of many parts.

If we know the life expectancy of the hard disk, does this tell us sometime about the life expectancy of the computer ?

Example: Leukemia

(Klein and Moeschberger (2003))

Bone marrow transplants are a standard treatment for acute leukemia. In this multi center trial, patients were followed from the moment they had a bone marrow transplant until the leukemia returns. The researchers investigated the time until relapse and looked for factors that influenced this time.

Some patients did not have a relapse during the study period or died without relapsing.

Example: Vaginal cancer

(Kalbfleisch and Prentice (2002))

In a study on carcinogenesis, two groups of rats were exposed to a carcinogen DMBA. One group was keeped in a germ-free environment while the other group was not. In both group the number of days was recorded until these rats died of vaginal cancer. The researchers investigated whether the time until death was different in both groups.

However four rats did not die from cancer or were still alive at the end of the study.

Example: Recidivism

(Schmidt and Witte (1988))

The North Caroline Department of Correction conducted two studies on recidivism. They looked at the time until a prisoner, who was released from a North Carolina prison, returned to a NC prison.

The goal of these studies was to find possible predicting factors like race, age at release, drug or alcohol problems, marital status, ...

We note that some released prisoners will never return.

Example: Liability claims

(Klugman and Rioux (2006))

At an insurance company, liability claims are made after some damage has occurred. There are different policies which have deductibles of 100, 250 or 500 euro and maximum payments of 1000, 3000 or 5000 euro. This insurance company wants to know how much it should pay each year and is interested in the distribution of the claim sizes.

Because they only pay a maximum payment of 5000, it is impossible to express the true amount of the damage.

Example: Wages

To gain insight into the economical status of a region, you may want to look at the income of every person in this region.

How would you deal with unemployment?

In reliability and survival analysis, we are interested in a non-negative random variable T ($T \geq 0$). This variable T can be discrete with values $\{0, 1, 2, \ldots\}$ or continuous on $(0, \infty)$.

This variable is known under a lot of different names:

- failure time
- lifetime
- time until an event
- loss
- ...

Sometimes we cannot fully observe this random variable T but only observe some boundaries for this time. This is called censoring.

Example: Leukemia

(Klein and Moeschberger (2003))

Let's go back to the starting examples...

Bone marrow transplants are a standard treatment for acute leukemia. In this multi center trial, patients were followed from the moment they had a bone marrow transplant until the leukemia returns. The researchers investigated the time until relapse and looked for factors that influenced this time.

Some patients did not have a relapse during the study period or died without relapsing.

Example: Vaginal cancer

(Kalbfleisch and Prentice (2002))

In a study on carcinogenesis, two groups of rats were exposed to a carcinogen DMBA. One group was keeped in a germ-free environment while the other group was not. In both group the number of days was recorded until these rats died of vaginal cancer. The researchers investigated whether the time until death was different in both groups.

However four rats did not die from cancer or were still alive at the end of the study.

Example: Recidivism

(Schmidt and Witte (1988))

The North Caroline Department of Correction conducted two studies on recidivism. They looked at the time until a prisoner, who was released from a North Carolina prison, returned to a NC prison.

The goal of these studies was to find possible predicting factors like race, age at release, drug or alcohol problems, marital status, ...

We note that some released prisoners will never return.

Conclusion:

In each of the examples, we cannot fully observe the time until a certain event. Due to different practical reasons, we only observe in the examples a lower bound of the true time.

This is called right censoring.

In general, any situation in which you cannot fully observe a time until an event but only observe some boundaries for this time is called censoring.

In the examples, there is another non-negative random variable C ($C \geq 0$) which we call the censoring variable and which obscures the observation of T.

For the moment, we only consider right censoring. There are three types of censoring.

- Type I or fixed censoring.
- Type II censoring
- Type III or random censoring.

Type I or fixed censoring

Let $t_c \in \mathbb{R}$ be a fixed time point and take a sample lifetimes $T_1, \ldots, T_n$.

We only observe a lifetime T_i if it is smaller than t_c, otherwise we get this fixed time point.

Hence, we get a sample $Y_1, \ldots, Y_n$ where

$$Y_i = \begin{cases} T_i & \text{, if } T_i \leq t_c \\ t_c & \text{, if } T_i > t_c \end{cases} \quad \text{, } i = 1, \ldots, n.$$

Example: Study stopped at a fixed time.

Type II censoring

Let $r < n$ with $r \in \mathbb{N}$ and denote by $T_{(1)} < \ldots < T_{(n)}$ the ordered lifetimes.

We observe until the r-th system has failed.

Hence we get

$$Y_{(i)} = \begin{cases} T_{(i)} & \text{, if } T_{(i)} \leq T_{(r)} \\ T_{(r)} & \text{, if } T_{(i)} > T_{(r)} \end{cases} \quad \text{, } i = 1, \ldots, n.$$

Example: Industrial test trial.

Type III or random censoring

Let $C_1, \ldots, C_n$ be a sample of censoring times.

We observe a sample of couples, $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ where, for $i = 1, \ldots, n$,

$$Y_i = \min(T_i, C_i) = \begin{cases} T_i & , \text{if } T_i \leq C_i \\ C_i & , \text{if } T_i > C_i \end{cases}$$

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 1 & , \text{if } T_i \leq C_i \\ 0 & , \text{if } T_i > C_i \end{cases}$$

In general we assume that, for $i = 1, \ldots, n$, T_i and C_i are independent.

Based on observations of $Y_1, \ldots, Y_n$, we want to estimate the distribution F.

In a similar way, we can look at other censoring schemes.

Left censoring

We observe a sample $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ where, for $i = 1, \ldots, n$,

$$Y_i = \max(T_i, C_i) = \begin{cases} T_i & \text{, if } T_i \geq C_i \\ C_i & \text{, if } T_i < C_i \end{cases}$$

$$\delta_i = I(T_i \geq C_i) = \begin{cases} 1 & \text{, if } T_i \geq C_i \\ 0 & \text{, if } T_i < C_i \end{cases}$$

Some examples:

- Development of children behavior: At which age can a child perform a certain task. Some children can perform the task when they enter the study.
- Detection limits: A measuring device cannot give a correct value below a fixed limit.

Interval-censoring

Instead of a sample of lifetimes $T_1, \ldots, T_n$, we get for each individual an interval in which the event occurred. Hence we get $(L_1, R_1], \ldots, (L_n, R_n]$.

Example:
In a clinical trial on breast cancer patients, the researchers were interested whether there was a difference in cosmetic effects for early breast cancer patients when they were treated with radiotherapy only or with radiotherapy and chemotherapy. At each visit, every 4 till 6 month, a clinician recorded a measure for breast retraction. Of interest was the time until moderate or severe breast retraction appeared.

Doubly censoring

We observe a sample $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ where, for $i = 1, \ldots, n$,

$$\begin{aligned} Y_i &= \min(\max(T_i, L_i), R_i) \\ \delta_i &= = \begin{cases} 1 & , \text{if } Y_i = T_i \\ 0 & , \text{if } Y_i = R_i \\ -1 & , \text{if } Y_i = L_i \end{cases} \end{aligned}$$

Example:
In a study conducted at the Stanford-Palo Alto Peer counseling program, 191 California high school boys were asked: "When did you first use marijuana?" The answers were the exact ages (uncensored observations), "I never used it" (right censored observations), "I have used it but can not recall when the first time was" (left censored observations).

In a clinical study, we encounter three main mechanisms of censoring:

- Administrative censoring: due to the end of study.
- Withdrawal: a patient does not want to continue the treatment.
- Loss to follow-up: a patient do not show up anymore at the follow-up visits during trial.

Sometimes we note that the censoring variable is not independent of the lifetime, or the distribution of the censoring variable is linked to the distribution of the lifetime. We then call the censoring informative.

In the example, we see that

- Administrative censoring usually fulfills the independence condition.
- Withdrawals/losses to follow-up are potentially problematic: can be due to, e.g., disease progression or death.

In this course we will mainly deal with non-informative censoring and should try to avoid that we get informative censoring by checking the reasons why individuals leave a study.

To define a lifetime correctly, we need

- A time-origin.
- A time-scale.
- A definition of when the endpoint occurs.

However, sometimes there are several choices of time-origins and associated times-scales. Choosing the right one depends on the purpose of the study. We have in most cases:

- Study time: calendar time between the beginning and end of the study.
- Subject time: time spent in a study, measured from the subject's origin.

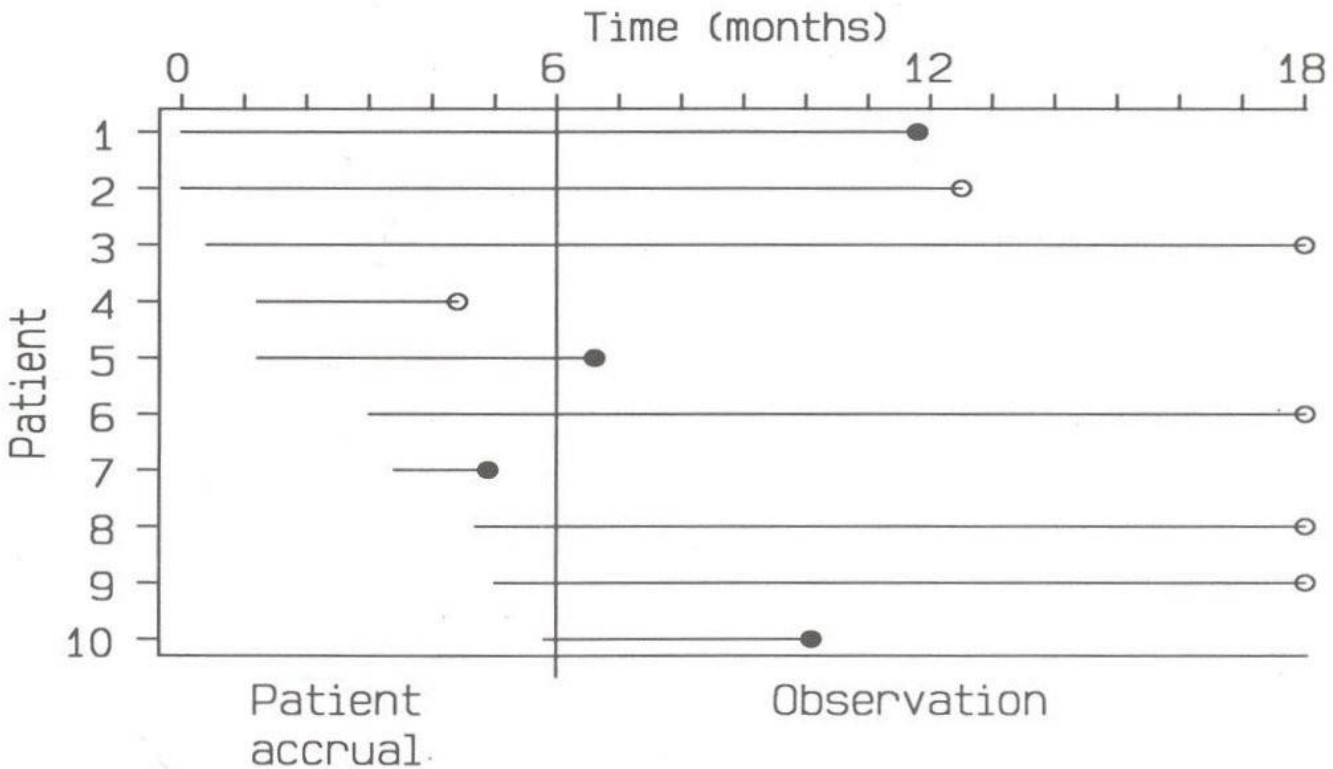


Figure 13.1 Diagram showing patients entering a study at different times and the observation of known (●) and censored (○) survival times.

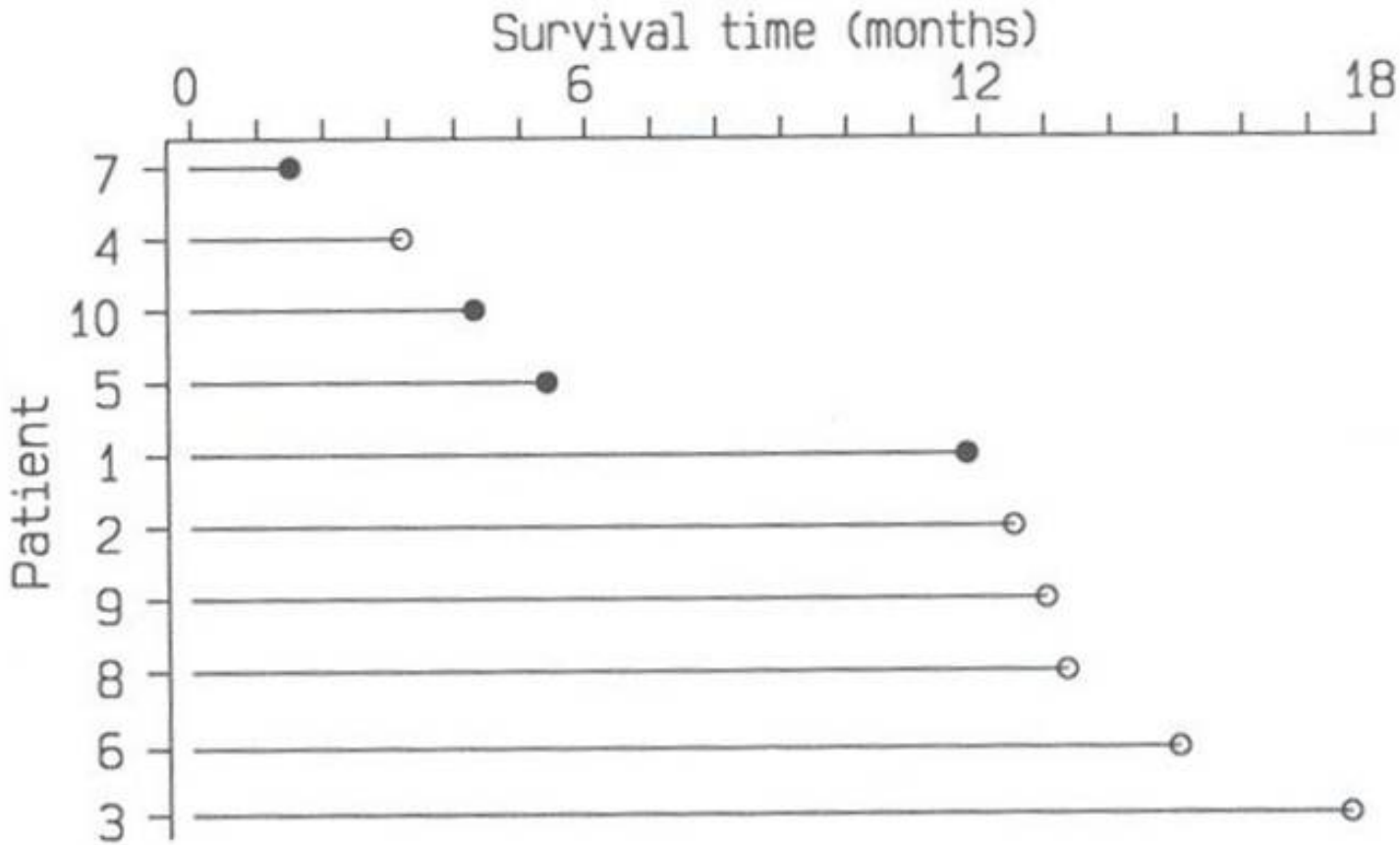


Figure 13.2 Figure 13.1 reorganized to correspond to method of analysis.

Survival function

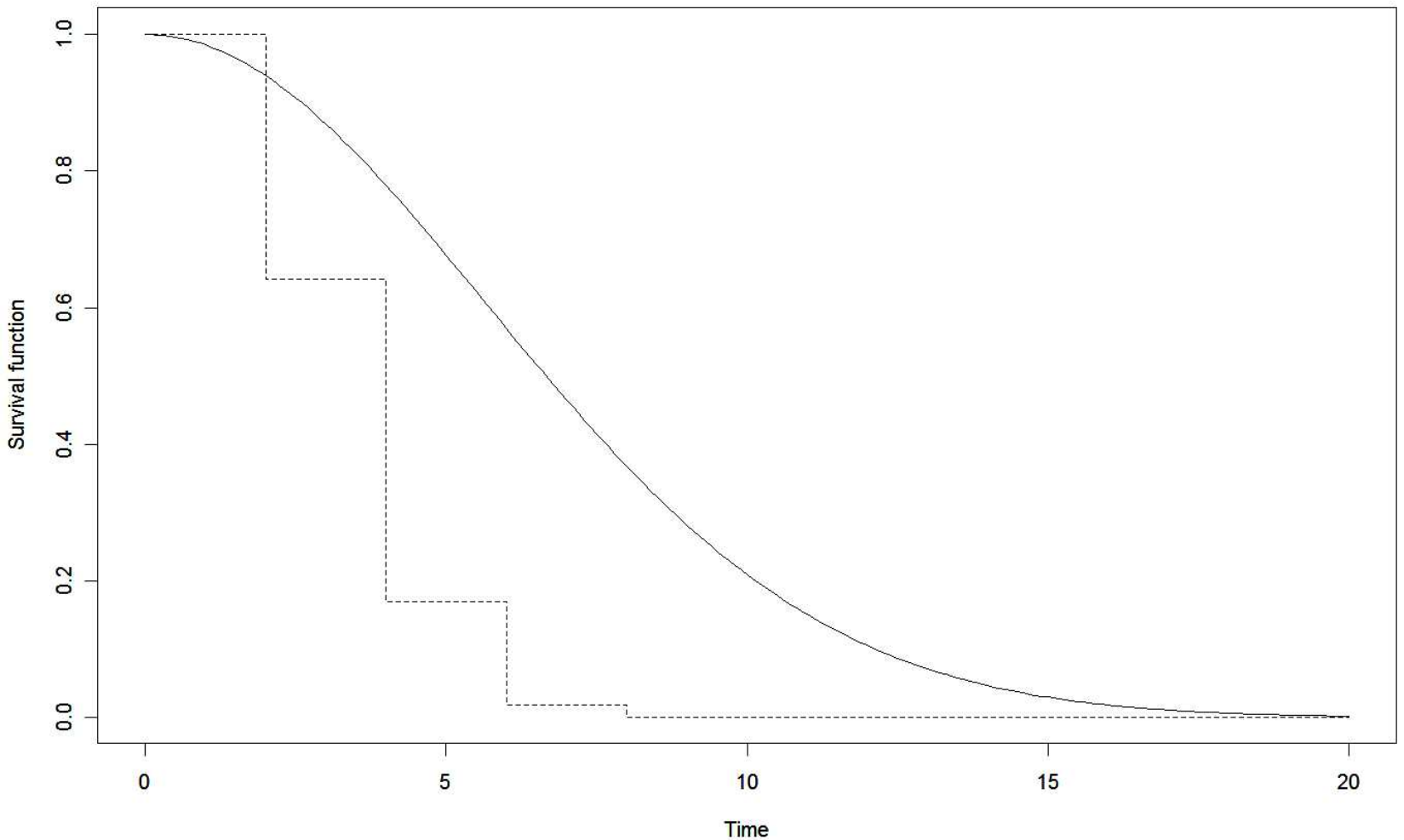
One of the important quantities in reliability and survival analysis is the survival function or reliability function

$$S(t) = P(T > t) = 1 - F(t).$$

It is the probability that an event has not occurred by time t.

Some properties:

- $S(0) = 1$.
- $S(+\infty) = \lim_{t \to +\infty} S(t) = 0$.
- $S(t)$ is a non-increasing function of t.


Survival function
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Time

T is continuous

In this case, we have a density function

$$f(t) = -\frac{dS(t)}{dt} \quad \Rightarrow S(t) = \int_t^\infty f(x)dx.$$

A second important quantity in reliability and survival is the hazard function which is often also called (instantaneous) failure rate or intensity function,

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h | T > t)}{h}.$$

It describes the probability for an event to take place in an small interval after time t, given that it has not occurred before t,

$$\lambda(t)h \approx P(t \leq T < t + h | T > t).$$

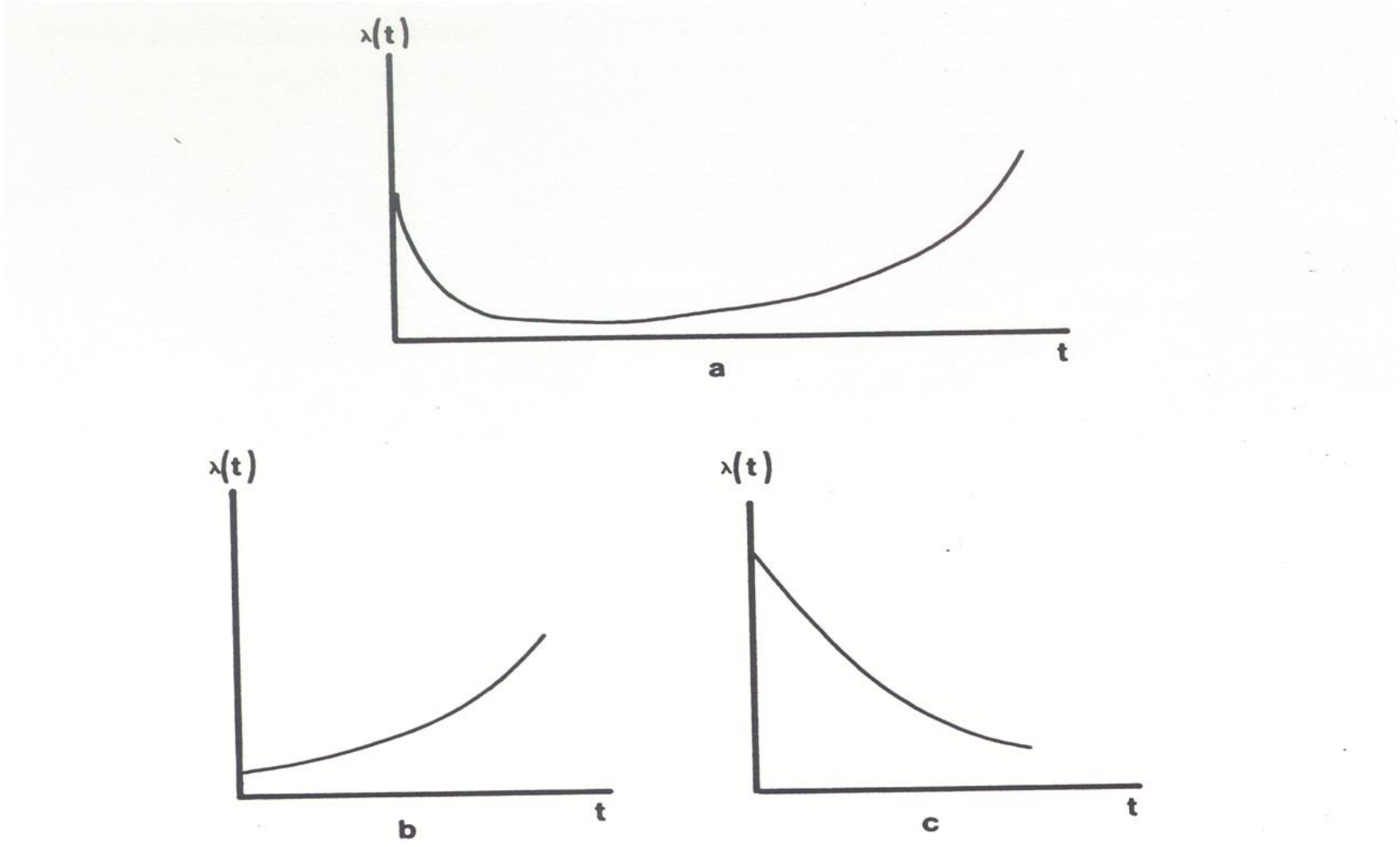


Figure 1.1 Some types of hazard functions: (a) hazard for human mortality; (b) positive aging; (c) negative aging.

Relationship between $\lambda(t)$ and $S(t)$.

$$\begin{aligned}
\lambda(t) &= \lim_{h\to 0} \frac{P(t \le T < t+h | T > t)}{h} \\
&= \lim_{h\to 0} \frac{1}{h} \frac{P(t \le T < t+h)}{P(T > t)} \\
&= \lim_{h\to 0} \frac{1}{h} \frac{S(t) - S(t+h)}{S(t)} = \frac{f(t)}{S(t)}.
\end{aligned}$$

Conversely,

$$\begin{aligned}
\lambda(t) &= \frac{f(t)}{S(t)} = \frac{-1}{S(t)} \frac{d}{dt} S(t) = -\frac{d}{dt} \log S(t) \\
&\Rightarrow \log S(t) = -\int_0^t \lambda(x) dx \Rightarrow S(t) = \exp\left(-\int_0^t \lambda(x) dx \right)
\end{aligned}$$

We define the cumulative hazard function as

$$\Lambda(t) = \int_0^t \lambda(x)dx.$$

We note that $\Lambda(0) = 0$ and $\Lambda(+\infty) = \int_0^{+\infty} \lambda(x)dx = +\infty$.

In reliability and survival analysis, we are also interested in the residual lifetime T_t. This is the remaining lifetime of a system that has survived until time t. The distribution of the residual lifetime is given by

$$F_t(x) = P(T - t \leq x | T \geq t) = \frac{F(t+x) - F(t)}{S(t)}.$$

Looking at the average value, we get the mean residual life time,

$$r(t) = E[T - t | T \geq t] = \frac{\int_t^\infty (x-t) f(x) dx}{S(t)} = \frac{\int_t^\infty S(x) dx}{S(t)}.$$

Furthermore we can derive the following relationships

$$\begin{aligned} S(t) &= \frac{r(0)}{r(t)} \exp\left(-\int_0^t \frac{du}{r(u)}\right) \\ f(t) &= \left(\frac{d}{dt} r(t) + 1\right) \frac{r(0)}{r(t)^2} \exp\left(-\int_0^t \frac{du}{r(u)}\right) \\ \lambda(t) &= \frac{1}{r(t)} \left(\frac{d}{dt} r(t) + 1\right) \end{aligned}$$

We note for the mean life time that $r(0) = E[T]$.

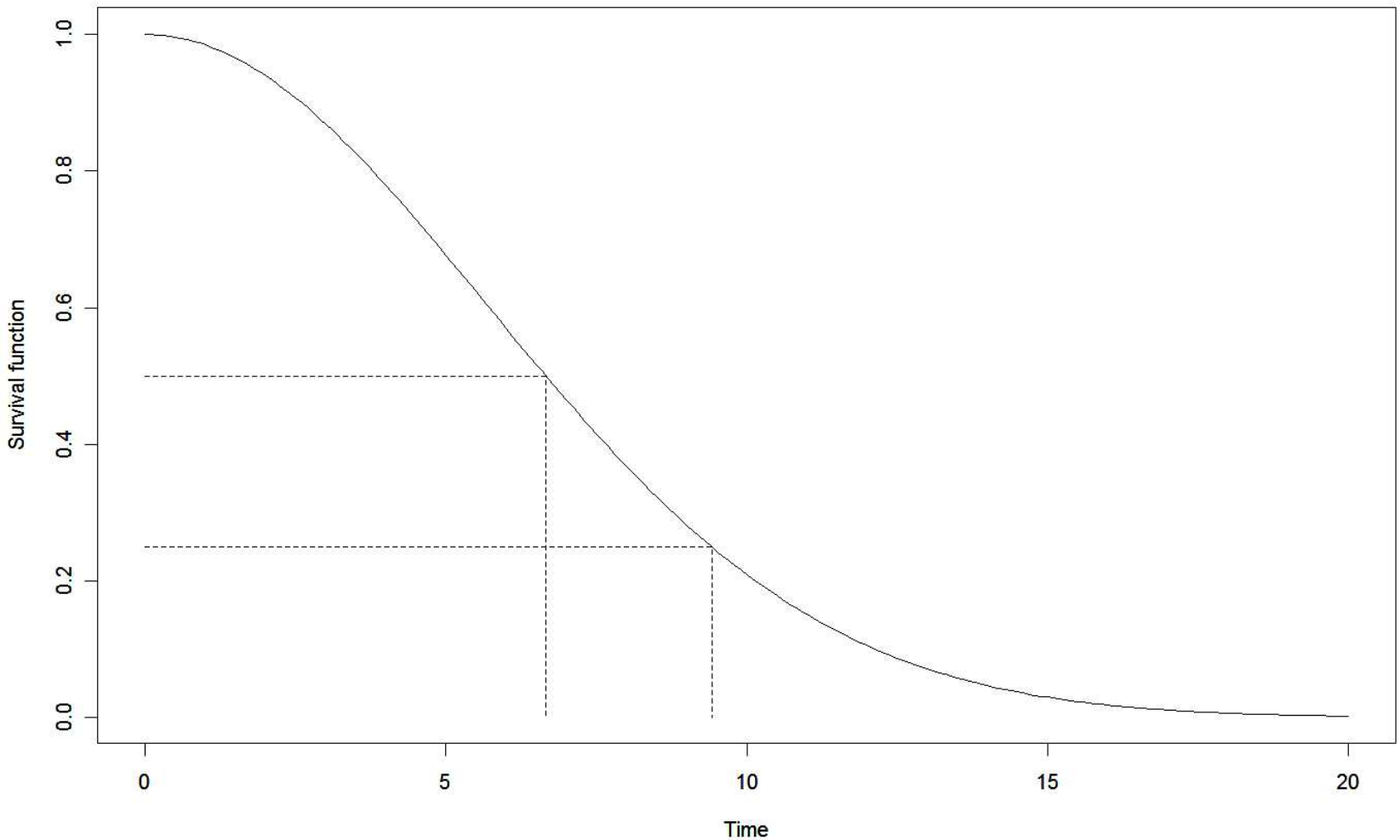
Hereby we can show that

$$r(0) = E[T] = \int_0^{+\infty} tf(t)dt = \int_0^{+\infty} S(t)dt.$$

Sometimes we are confronted in reliability and survival analysis that $E[T] = +\infty$. Therefore we introduce percentiles.

The p -th percentile t_p is the solution of the equation

$$S(t_p) = 1 - p.$$


Survival function
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
Time

T is discrete

The probability function is given by

$$f(a_i) = P(T = a_i), \quad i = 1, 2, \ldots \quad \Rightarrow S(t) = \sum_{a_j > t} f(a_j).$$

The hazard function is given by

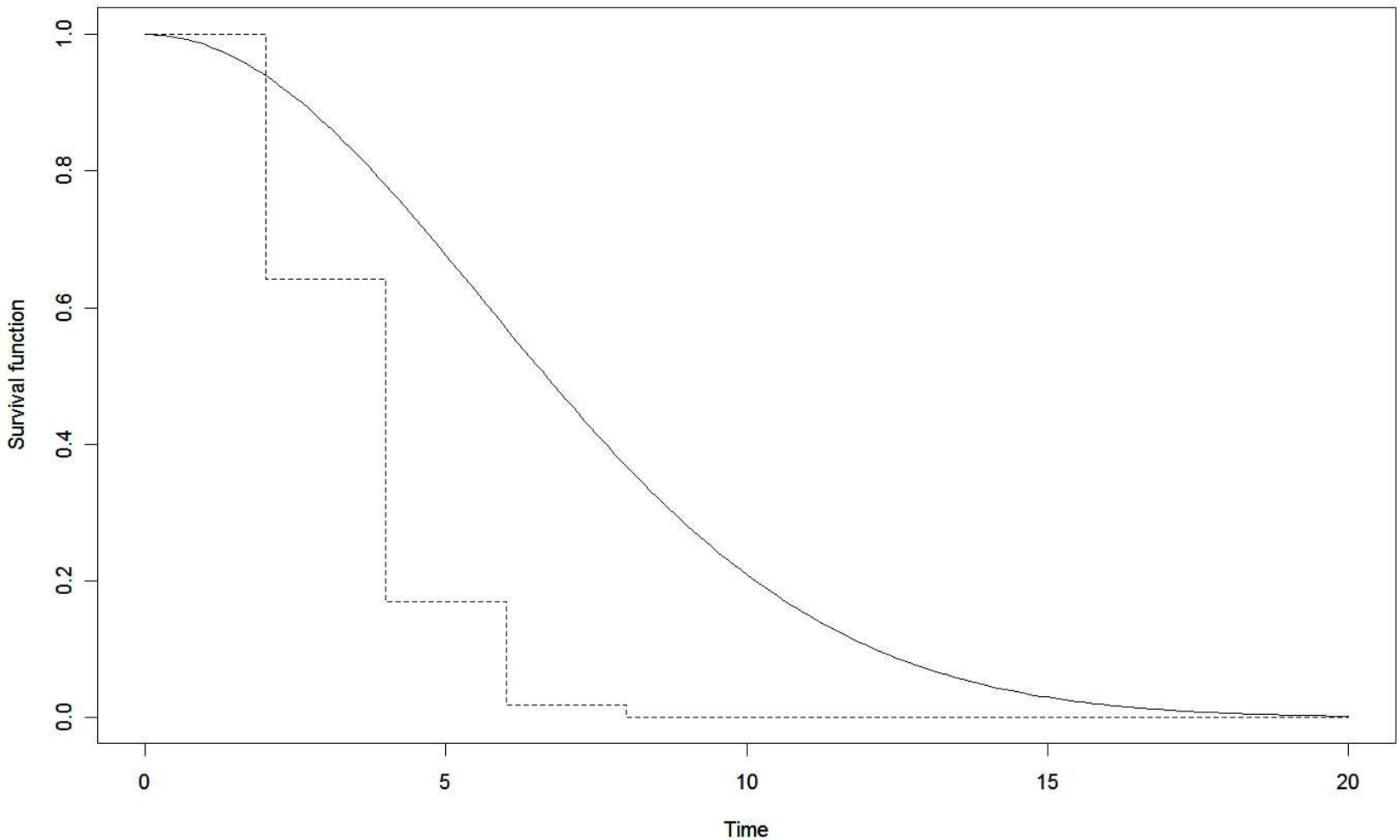
$$\lambda_i = \lambda(a_i) = P(T = a_i | T \geq a_i) = \frac{f(a_i)}{S(a_j^-)}, \quad i = 1, 2, \ldots$$

and

$$S(t) = \prod_{a_i \leq t} (1 - \lambda_i).$$

The p -th percentile t_p is the smallest t_p such that $S(t_p) \leq 1 - p$,

$$t_p = \inf\{t : S(t) \leq 1 - p\}.$$


Survival function
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Time