

## 2 Basic terminology and quantities

We first describe how survival time can be appropriately modeled and define some basic quantities which characterize the survival time distribution. Then, we introduce some terminology and notation of survival time samples. In this context, we will also precisely state what censoring means. At the end of this chapter, we will discuss how some of the basic quantities introduced before can be estimated. Throughout this chapter, we will follow the textbooks of Klein and Moeschberger [16], pp. 21-36, 63-72, 91-101 as well as Kalbfleisch and Prentice [14], pp. 6-18.

Basically, we interpret an observed survival time, denoted by  $t$ , as a realization of a nonnegative random variable  $T$ . We assume  $T$  to be continuous, although it should be mentioned that there are also survival analytic methods available for discrete survival time as well as for mixtures of those two types. For more details, we refer to standard textbooks such as Kalbfleisch and Prentice [14], pp. 8-10.

Recall that the probability distribution of a continuous random variable can be specified by giving a so-called *density function*  $f$ . Alternatively, you could also characterize the distribution by the corresponding (*cumulative*) *distribution function*  $F$ , which is defined as  $F(t) := P(T \leq t)$ . For example, a quite simple but typical choice in the survival analytic setting is the *exponential distribution*, whose density is given by

$$f(t) = \lambda \exp(-\lambda t), t \geq 0,$$

where  $\lambda$  is a fixed constant greater than 0.

Nevertheless, there are some additional quantities which are especially useful in time-to-event analysis.

**Definition 2.1.**

- (i)  $S(t) := P(T > t), t \in [0, \infty)$ , is called *survival function*.
- (ii)  $\lambda(t) := \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, t \in [0, \infty)$ , is called *hazard function*.
- (iii)  $\Lambda(t) := \int_0^t \lambda(s) ds, t \geq 0$ , is called *cumulative hazard function*.

These quantities have some basic properties which are stated in the following proposition.

**Proposition 2.2** (Properties of  $S(t)$ ,  $\lambda(t)$  and  $\Lambda(t)$ ).

- (i)  $S(t) = 1 - F(t), t \geq 0$ .
- (ii)  $S$  is monotonically decreasing in  $t$ ;  $S(0) = 1$ ;  $\lim_{t \rightarrow \infty} S(t) = 0$ .
- (iii)  $\forall t \in [0, \infty) : \lambda(t) \geq 0$ .

*Proof.* (i) follows directly from the definition of  $S$ . Applying the fundamental properties of a distribution function yields (ii) (note that  $F(0) = 0$  because  $T$  is assumed to be nonnegative). (iii) follows from the fact that both the numerator and the denominator in the definition of  $\lambda$  are nonnegative.  $\square$

Furthermore,  $S, \lambda$  and  $\Lambda$  are closely related to the density and the distribution function of  $T$ . Exactly speaking, the distribution of  $T$  can be specified by taking any of these quantities. In order to prove this, we first need a lemma that clarifies the relation between  $\lambda, f$  and  $S$ .

**Lemma 2.3.** *For all  $t \in [0, \infty)$ , we have*

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

*Proof.* According to the definitions of  $\lambda$  and  $S$  and some basic probability theory, we have

$$\begin{aligned}\lambda(t) &= \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{S(t)\Delta t} = S(t) \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t}.\end{aligned}$$

Note that this limit is just the first derivative of  $F$  with respect to  $t$ . Since we assumed  $T$  to be a continuous random variable, this is equal to  $f(t)$ , and the proof is complete.  $\square$

**Example 2.4.** To illustrate that lemma, let us turn back to our previous example. According to statement (i) in the proposition above, the survival function for the exponential distribution is given by

$$S(t) = 1 - F(t) = 1 - (1 - \exp(-\lambda t)) = \exp(-\lambda t).$$

The hazard function can be calculated using the previous lemma:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\lambda \exp(-\lambda t)}{\exp(-\lambda t)} = \lambda.$$

So, we see that the hazard for the exponential distribution is actually a constant.

**Theorem 2.5.**

$$(i) \quad \lambda(t) = -\frac{d \ln(S(t))}{dt}.$$

$$(ii) \quad S(t) = \exp\left(-\int_0^t \lambda(x) dx\right).$$

*Proof.* As to the proof of statement (i), one just has to evaluate the expression on the right handside and apply some of the results from above:

$$-\frac{d \ln(S(t))}{dt} = -\frac{1}{S(t)} \frac{dS(t)}{dt} = \frac{1}{S(t)} \frac{dF(t) - 1}{dt} = \frac{f(t)}{S(t)} = \lambda(t).$$

(ii) follows directly by integrating and exponentiating the equation in (i).  $\square$

So, to sum up our results once again, we now know that the probability distribution of a (continuous) random variable  $T$  - our survival time variable - can be specified by either the density  $f$ , the cumulative distribution function  $F$ , the survival function  $S$  or the hazard (cumulative hazard) function  $\lambda(\Lambda)$ . It often depends on the topic and the research question(s) which quantity you decide to take: Especially in survival analyses, you usually choose the survival or the hazard function, depending on which interpretation you are interested in: Roughly speaking, whereas the first one is a measure of the overall survival experience of the study population, the hazard function quantifies the risk of having the event at a certain time point. So, for example, if the investigator has a certain idea how the risk varies over time (e.g. increasing), the survival time distribution can be modelled by specifying a hazard function which reflects that behaviour.

Let us now turn to the problem of censoring, which is often a key feature of survival data. I have already illustrated this issue by giving some examples in the introduction. Now, I want to introduce some notation in order to be able to appropriately deal with censoring. Note that for the purposes of my master's thesis, I will restrict the focus on right censoring. For a discussion of other censoring schemes, I refer to Kalbfleisch and Prentice [14], pp. 78-83.

First, we should think about the question why we need to develop certain methods which allow for censoring. For instance, we could simply take the exact survival times and do some kind of linear or nonlinear regression with the survival time being the response and several predictors such as age, gender, etc. if your data comes from a medical context such as in our cancer study mentioned above. But to stay with this example, we would throw away a considerable amount of data - 110 out of 1,000. Apart from the fact that usually those follow-ups cost a lot of time and money, which then would have been spent in vain, one shouldn't forget about the information censored data provide: To take a rather extreme, but illustrative example, it's evident that there's a difference between 5 percent and 80 percent

being still alive at the end of the study period. Of course, as the patients may be diagnosed at different time points throughout the study period (e.g. in 2005 and 2012), the censored observations as well as the information they provide can't be adequately dealt with by just looking at that proportion. Thus, the main question is whether we can make use of their survival times somehow. Actually, the following chapters will show how this goal is achieved.

Until now, we have only considered a single random variable  $T$  representing survival time. However, when looking at real-life situations, we will have a sample of survival times, possibly with associated covariates and censoring. In order to formulate this setting mathematically, we introduce the following definitions:

**Definition 2.6** (Censored sample (with covariates)). Let  $(T_i^{ex})_{i=1}^n$  be a finite sequence of random variables with corresponding probability distributions  $(P_i)_{i=1}^n$ . Let  $(C_i)_{i=1}^n$  be a finite sequence of random variables with densities  $(g_i)_{i=1}^n$ . Let  $T_i := \min\{T_i^{ex}, C_i\}$  and  $\Delta_i := \mathbf{1}(T_i = T_i^{ex}), 1 \leq i \leq n$ . Assume that  $\{T_1^{ex}, T_2^{ex}, \dots, T_n^{ex}, C_1, C_2, \dots, C_n\}$  are independent.

- (i) Then, the finite sequence  $((T_i, \Delta_i))_{i=1}^n$  is called a *sample of right censored survival times*  $(P_i)_{i=1}^n$ . The elements of the sequence  $(C_i)$  are called *right censoring times*, the elements of  $(\Delta_i)$  are called *censoring indicators*.  $T_i^{ex}$  is called *exact survival time*,  $T_i$  is called *observed survival time*,  $1 \leq i \leq n$ .
- (ii) Let  $(\mathbf{X}_i)_{i=1}^n$  be a finite sequence of covariate vectors corresponding to the sequence of exact survival times. Assume that the probability distribution  $P_i$  depends on  $\mathbf{X}_i, 1 \leq i \leq n$ . The finite sequence  $((T_i, \Delta_i, \mathbf{X}_i))_{i=1}^n$  is called a *sample of right censored survival times with covariate vectors*  $(\mathbf{X}_i)_{i=1}^n$ .

**Remark 2.7.**

- (i) The censoring scheme assumed above is usually referred to as “random censoring”. In my master’s thesis, I won’t always state

this fact explicitly because it is already included in my definition of a censored sample.

- (ii) The term “observed survival time” does not mean that we have realizations here! By the word “observed”, we emphasize the contrast to the exact survival times because the latter ones may not be observed for some subjects.
- (iii) In order to stress the fact that the probability distributions  $P_i$ , specified by some quantity like the hazard function, depend on the covariate vectors  $\mathbf{X}_i$ , we sometimes write, for instance,  $\lambda(t|\mathbf{X}_i)$  instead of  $\lambda_i(t)$ .
- (iv) Note that independence is required for the exact but not for the observed survival times! Nevertheless, the assumptions concerning independence in Definition 2.6 imply that the observed survival times  $T_1, T_2, \dots, T_n$  are independent, too. But be careful: Obviously, the  $T_i$ s and the  $C_i$ s are **not** independent!

Of course, these definitions look a bit technical, but there isn’t a way to get around that effort since we need a proper notation for our further work.

To be honest, the definitions stated above can be made in a much more general way: There are other types of right censoring than the one I have defined. Exactly speaking, the above setting, that is, random (right) censoring, is a special case of a very general situation called *independent censoring*, which also covers censoring schemes where, for instance, the subjects are followed until the  $k$ -th event. But as I don’t want to confuse the reader with too much terminology which is not illustrated by some examples, I refer to Kalbfleisch and Prentice [14], pp. 52-54, for the details concerning the generalizations of several censoring schemes. I just formulate what independent censoring means in order to make the reader see that the concept developed above indeed fulfils this independence assumption.

**Definition 2.8** (Independent censoring). Let  $((T_i, \Delta_i, \mathbf{X}_i))_{i=1}^n$  be a sample of survival times  $T_i$  with covariate vectors  $(\mathbf{X}_i)_{i=1}^n$ , where  $\Delta_i$  indicates if  $T_i$  is right-censored or not, i.e.,  $\Delta_i = 1$  if subject  $i$  experiences the event of interest within the study period and  $\Delta_i = 0$  otherwise,  $1 \leq i \leq n$ . The underlying right-censoring mechanism is called *independent* if the values of the hazard function that apply to individuals on trial at each time  $t > 0$  are the same as those that would have been applied if there had not been censoring. In other words, for each  $t > 0$  and  $i \in \{1, 2, \dots, n\}$ , the following equation must be satisfied:

$$\begin{aligned} & \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T_i^{ex} < t + \Delta t | \mathbf{X}_i = \mathbf{x}_i, T_i^{ex} \geq t)}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T_i^{ex} < t + \Delta t | \mathbf{X}_i = \mathbf{x}_i, T_i^{ex} \geq t, Y_i(t) = 1)}{\Delta t}, \end{aligned} \quad (2.1)$$

where  $Y_i(t) = 1$  if subject  $i$  is at risk at time  $t$  and  $Y_i(t) = 0$  otherwise.

**Proposition 2.9.** Let  $((T_i, \Delta_i, \mathbf{X}_i))_{i=1}^n$  be a sample of right-censored survival times with covariate vectors  $(\mathbf{X}_i)_{i=1}^n$  as in Definition 2.6. Then, the underlying censoring scheme is independent.

*Proof.* To begin with, in the setting of Definition 2.6,  $Y_i(t) = 1$  is equivalent to  $T_i^{ex} \geq t \wedge C_i \geq t$ . Thus, we can write

$$\begin{aligned} & P(t \leq T_i^{ex} < t + \Delta t | \mathbf{X}_i = \mathbf{x}_i, T_i^{ex} \geq t, Y_i(t) = 1) \\ &= P(t \leq T_i^{ex} < t + \Delta t | \mathbf{X}_i = \mathbf{x}_i, T_i^{ex} \geq t, C_i \geq t). \end{aligned}$$

But we have also assumed that  $T_i$  and  $C_i$  are independent. So we can drop the condition on  $C_i$  and get

$$\begin{aligned} & P(t \leq T_i^{ex} < t + \Delta t | \mathbf{X}_i = \mathbf{x}_i, T_i^{ex} \geq t, C_i \geq t) \\ &= P(t \leq T_i^{ex} < t + \Delta t | \mathbf{X}_i = \mathbf{x}_i, T_i^{ex} \geq t). \end{aligned}$$

All in all, we have shown that

$$\lim_{\Delta t \rightarrow 0} \frac{P(t \leq T_i^{ex} < t + \Delta t | \mathbf{X}_i = \mathbf{x}_i, T_i^{ex} \geq t, Y_i(t) = 1)}{\Delta t}$$

is equal to

$$\lim_{\Delta t \rightarrow 0} \frac{P(t \leq T_i^{ex} < t + \Delta t | \mathbf{X}_i = \mathbf{x}_i, T_i^{ex} \geq t)}{\Delta t},$$

which means that the censoring scheme is independent.  $\square$

Again, we emphasize that in the definition above, some assumptions of random right censoring are relaxed. However, the random censorship setting is sufficiently general for many practical applications, and we have seen that it fulfils the independence assumption, so considering this censoring scheme only will do for the remaining part of my master's thesis. But, one has to keep in mind that even the assumptions required for independent censoring can be violated in situations which are quite often encountered, as demonstrated in the remark below.

**Remark 2.10.** To further illustrate the concept of independent censoring, it may be instructive to look at an example where the censoring scheme is **not** independent. Just think of a medical study on the time to death of seriously ill patients, where some surgical intervention is taken as starting point. Let us assume that for practical reasons, patients are discharged from hospital when a certain score, which measures the degree of recovery, exceeds a predefined value. After discharge from hospital, the patients will not be followed any more, so, to cut it short, they will be right censored if their score is fairly high. However, this censoring scheme is not independent: If you take a look at the right handside of equation (2.1), you see that the conditions  $\mathbf{X}_i = \mathbf{x}_i$  (which involves the quite large score value mentioned above) and  $Y_i(t) = 1$  (which in particular means that individual  $i$  is not censored at time  $t$ ) cannot hold simultaneously since we said that an individual with a high score will be censored. Consequently, the right handside of equation (2.1) is not well defined. Thus, the censoring scheme is not independent.

**Example 2.11.** Now, the terminology introduced above is illustrated by looking at our epilepsy dataset. Remember that we decided to take only a subset of the data, namely the so-called incidence cohort,



which consist of patients who came to the clinic within one year after their first seizure. The date of their first visit at the clinic, which is referred to as *behandlungsbeginn*, is considered to be the time origin. The corresponding censoring time, call it  $C_i$ , is the span between *behandlungsbeginn* and Dec 31, 2010. Recall that we don't have to care about patients lost to follow up, as stated previously.

In addition to that, although the subjects are scheduled for subsequent examinations, e.g. to find out if their health status has changed, all the covariates (gender etc.) we are interested in are measured at the time of entry and remain fixed during the entire study period. Summing up, all we need to know for our analysis is measured at the time of entry into the study (the first visit at the clinic), and thus, given the vector of covariates  $\mathbf{x}_i$  and the value of the variable *behandlungsbeginn*, the associated censoring time  $C_i$  is fixed, as it is given by the difference between end of the study and date of first visit. As we deal with censoring times being constants, we obviously have a special case of (random) right censoring here. Particularly, our underlying censoring scheme for the epilepsy dataset is independent.

By the way, it should be mentioned that we could also choose a slightly different setting. Instead of looking only at the incidence cohort, we could take the whole dataset and define the value of the variable *anfallsbeginn*, which represents the date of a patient's first epileptic seizure, as starting point. Of course, we must then account for the fact that for some subjects, the time span between *anfallsbeginn* and *behandlungsbeginn* is quite large, which may be a source of bias since those patients had already survived for a fairly long time until they were enrolled. Actually, there are survival analytic methods available that account for the phenomenon of so-called *delayed entry* or *left truncation*. Note that this is different from *left censoring*: When left censoring occurs, we have at least partial information on the survival times insofar as we know that they do not exceed a certain value (just remember the answer "I drank alcohol, but I cannot remember the exact age" in our hypothetical study presented in the introductory

chapter). However, left truncation does not give us further information about the subjects which experienced the event of interest prior to the entry time. For example, imagine a study being conducted at a retirement center, with the age of the subjects as time to event (Klein and Moeschberger [16], pp. 16-17). Obviously, people have to survive to a sufficient age to be allowed to enter the retirement center. Thus, we do not know anything about the survival experience of people who died prior to that certain age. Furthermore, you can also see what's the main problem with the analysis of left truncated data: Although there are methods available which enable the investigator to properly deal with that type of data, we eventually get conditional quantities. But as you can see when looking at the epilepsy data, the time spans between onset of seizures and the first visit at the clinic range from a few days up to ten or twenty years. Thus, we would only be able to get statements like "if a patient with certain characteristics survives 10 years, his or her survival probability will be...". However, we want to find out how life expectancies change over time, from the date of first visit at the clinic up to 20 years after diagnosis, as will be discussed in chapter 5. Especially for the early years, truncation would make such statements impossible, and therefore, I think it is a better decision to stay with the incidence cohort approach.

In addition to independence, there is another property of censoring schemes which must not be confused with the first one:

**Definition 2.12** (Noninformative censoring). Let  $((T_i, \Delta_i))_{i=1}^n$  be a sample of censored survival times as in Definition 2.6. Let  $(f_i(\theta)), (g_i)$  be the densities of  $(T_i^{ex})$  and  $(C_i)$ , respectively, where  $\theta$  is a vector of parameters. The underlying censoring scheme is said to be *noninformative* if the densities  $g_i$  do not involve  $\theta$ .

**Remark 2.13.** A completely analogous definition can be stated if we additionally incorporate covariate vectors  $(\mathbf{X}_i)$  in our model. We will see in the next chapter that especially in a regression context, the noninformativity property is of importance when constructing the likelihood function.

Now, after this rather extensive account of censoring, we turn back to the key survival analytic quantities we have defined at the beginning of this chapter. So far, we haven't talked about the estimation of the survival and the (cumulative) hazard function based on a sample of censored survival times. First, we introduce a very basic, but important approach. Let us consider a finite sequence  $((t_i, \delta_i))_{i=1}^n$  of observations arising from a homogeneous population with survival function  $S_0$ . Let us apply some kind of re-indexing to our data such that the exact survival times are denoted by  $t_1, t_2, \dots, t_k$  and sorted in ascending order. Furthermore, we assign double-indices to the remaining  $n - k$  censored observations, thus indicating which time interval they belong to. More to the point, we denote the  $m_j$  censoring times in the interval  $[t_j, t_{j+1})$  by  $t_{j1}, t_{j2}, \dots, t_{jm_j}$ ,  $j \in \{0, 1, \dots, k\}$ , where we set  $t_0 := 0$  and  $t_{k+1} := \infty$  (although we don't need these quantities right now, this notation is very useful for the derivations carried out later in this chapter). Let  $d_j$  denote the number of events at  $t_j$  and  $n_j := d_j + m_j + d_{j+1} + m_{j+1} + \dots + d_k + m_k$  the number of subjects at risk at a time just prior to  $t_j$ , respectively,  $0 \leq j \leq k$ . Now, we want to estimate the unknown survival function  $S_0$  (or, equivalently, any other survival analytic quantity which uniquely determines the underlying distribution). There are two very popular approaches which are introduced in the following definitions.

**Definition 2.14** (Kaplan-Meier estimate (Kaplan and Meier [15])). The function  $\hat{S} : \mathbb{R}^+ \rightarrow [0, 1]$  defined by

$$\hat{S}(t) := \prod_{\{j: t_j \leq t\}} \left(1 - \frac{d_j}{n_j}\right)$$

is called *Kaplan-Meier estimate* or *Product-Limit estimate* of the survival function  $S_0$ .

**Definition 2.15** (Nelson-Aalen estimate (Nelson [18], Aalen [1])). The function  $\hat{\Lambda} : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  defined by

$$\hat{\Lambda}(t) := \sum_{\{j: t_j \leq t\}} \frac{d_j}{n_j}$$

is called *Nelson-Aalen estimate* of the cumulative hazard function  $\Lambda_0$ .

**Remark 2.16.**

- (i) Note that the Kaplan-Meier estimate does not necessarily reduce to 0: For example, if

$$i_0 := \underset{i}{\operatorname{argmax}} t_i, \delta_{i_0} = 0,$$

that is, the largest survival time is a censored observation, we have  $d_i/n_i \neq 1$  for all  $i = 1, 2, \dots, k$ . In this case, we take  $\hat{S}(t)$  to be undefined for  $t > t_{i_0}$ . Alternatively, we could, for instance, set  $\hat{S}(t) = 0$  for  $t > t_{i_0}$ . But, when doing so, one always has to keep in mind that we thus introduce some additional assumptions concerning the underlying survival process. As to the alternative solution given above, setting the estimate to 0 means that we assume nobody can survive beyond  $t_{i_0}$ .

- (ii) Both quantities defined above involve  $d_i/n_i$ ,  $1 \leq i \leq n$ . These quotients can be regarded as very natural hazard estimates since they represent the observed deaths relative to the number of subjects at risk.
- (iii) As to the general outlook of the functions  $\hat{S}$  and  $\hat{\Lambda}$ , it is obvious from the formulas that they both are step functions with jumps at the observed exact survival times  $t_1, t_2, \dots, t_k$ . One should keep in mind that the size of these jumps is not only determined by the number of deaths  $d_j$ , but also by the number of censored observations in the interval  $[t_j, t_{j+1})$  because the latter quantities contribute to  $n_j$ . Thus, we now see how censoring is accounted for in the estimates defined above.
- (iv) Due to the results stated in Theorem 2.5, we can as well use  $-\ln(\hat{S})$  as an estimate of  $\Lambda_0$  and  $\exp(-\hat{\Lambda})$  to estimate  $S_0$ , respectively.

For now, we restrict our focus to the Kaplan-Meier estimate. We have already stated in the previous remarks that the components of  $\hat{S}(t)$

are basically made up by some very natural hazard estimates. Therefore, the Kaplan-Meier estimate may be a somehow reasonable and “good” choice. However, we should further clarify the properties of  $\hat{S}(t)$ .

At first, we take a look at a special case that arises if there aren’t any censored observations in the sample.

**Proposition 2.17.** *Let  $((t_j, \delta_j))_{j=1}^n$  be a sequence of observations with  $\delta_j = 1$  for all  $j \in \{1, 2, \dots, n\}$ . Furthermore, we assume that  $t_1 < t_2 < \dots < t_n$ . Then, we have*

$$\hat{S}(t) = 1 - \hat{F}(t)$$

for all  $t \geq 0$ , where  $\hat{F}$  denotes the empirical CDF.

*Proof.* Due to the fact that  $\delta_j = 1$  for all  $j \in \{1, 2, \dots, n\}$ , we have  $n_j - d_j = n_{j+1}$ . This yields

$$\hat{S}(t) = \prod_{\{j: t_j \leq t\}} \left(1 - \frac{d_j}{n_j}\right) = \prod_{\{j: t_j \leq t\}} \frac{n_j - d_j}{n_j} = \prod_{\{j: t_j \leq t\}} \frac{n_{j+1}}{n_j} = \frac{n_{j_0+1}}{n_0},$$

where  $n_{j_0+1} := \max\{j \in \{1, 2, \dots, n\} : t_j \leq t\}$ . Now, note that  $n_0 = n$  and  $n - n_{j_0+1}$  is equal to the number of events in the interval  $[t_0, t_{j_0+1})$ , which is in turn  $\#\{j : t_j \leq t\}$ . The second statement again follows from the assumption of no censoring. All in all, we thus get

$$1 - \hat{S}(t) = 1 - \frac{n_{j_0+1}}{n_0} = \frac{\#\{j : t_j \leq t\}}{n},$$

which completes the proof.  $\square$

Next, we recall an important result concerning the empirical CDF.

**Proposition 2.18.** *Let  $\hat{F}$  be the empirical CDF based on a sample  $x_1, x_2, \dots, x_n$  arising from a distribution with (unknown) CDF  $F_0$ . Let*

$$L(F) := \prod_{i=1}^n (F^+(x_i) - F^-(x_i))$$

denote the nonparametric likelihood for a given CDF  $F$ . Then, if we arbitrarily choose a CDF  $F$ , the following inequality holds:

$$L(F) < L(\hat{F})$$

In other words,  $\hat{F}$  is the (unique) nonparametric maximum likelihood estimate of  $F_0$ .

*Proof.* See Owen [20], p. 8. □

Since the MLE property is invariant under monotonic transformations, we can combine the two propositions stated above in order to get the result that in the special case where there is no censoring, the Kaplan-Meier estimate  $\hat{S}$  is the nonparametric MLE of the true survival function  $S_0$ . But, if we allow for censored observations, the question arises whether the MLE property is still valid or not. The following theorem tells us that indeed,  $\hat{S}$  is a nonparametric MLE of  $S_0$  in this general situation, too. With the notations introduced right above Definition 2.14 and the assumption that the underlying censoring scheme is independent, the likelihood function for an arbitrarily chosen survival function  $S$  is given as

$$L(S) = \prod_{j=0}^k \left( \left( S(t_j^-) - S(t_j^+) \right)^{d_j} \prod_{l=1}^{m_j} S(t_{jl}) \right). \quad (2.2)$$

The first factor can be explained by the very same arguments as for the “usual” nonparametric likelihood defined above. The latter product results from the assumption of independent censoring: For a censored observation, all we know is that the actual, unobserved survival time exceeds the censored one. This statement can be proven by using counting process techniques, see Kalbfleisch and Prentice [14], pp. 193-196.

Now, we are ready to state the theorem mentioned above (for the following proof, see Kalbfleisch and Prentice [14], pp. 15-16):

**Theorem 2.19** (MLE property of the Kaplan-meier estimate). *With the assumptions and notational conventions from above, the Kaplan-Meier estimate  $\hat{S}$  is a nonparametric maximum likelihood estimate of the (unknown) survival function  $S_0$ , i.e.,  $L(S) \leq L(\hat{S})$  for any survival function  $S$ .*

*Proof.* To start with, we derive two conditions the nonparametric MLE  $S_0$  necessarily has to fulfil. Firstly, if  $S(t_j^-) = S(t_j^+)$  holds for a  $j \in \{1, 2, \dots, k\}$ , the likelihood given in (2.2) vanishes. Consequently, such a  $S$  would certainly not maximize the likelihood. So, a nonparametric MLE, call it  $\tilde{S}$ , must be discontinuous at the exact survival times  $t_1, t_2, \dots, t_k$ . Secondly, recall that a survival function is always monotonically decreasing in  $t$ . Therefore, if we now take a look at maximizing the values of  $\tilde{S}$  at the censored observations (i.e., the latter part of (2.2)), we see that we have to set  $\tilde{S}(t_{jl}) = \tilde{S}(t_j)$  because we have chosen the double-indices such that  $t_j \leq t_{jl}$ ,  $j \in \{1, 2, \dots, k\}, l \in \{1, 2, \dots, m_j\}$ . So, summing up, we want to find a discrete survival function  $\tilde{S}$  with hazard components  $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_k$  at the exact survival times  $t_1, t_2, \dots, t_k$ . In particular, we therefore get

$$\tilde{S}(t) = \prod_{\{l: t_l \leq t\}} (1 - \tilde{\lambda}_l),$$

where  $t \geq 0$  (Kalbfleisch and Prentice [14], p.9). Now, we plug in this formula in (2.2) and obtain

$$\begin{aligned} L(\tilde{S}) &= \prod_{j=0}^k \left( \left( \tilde{S}(t_j^-) - \tilde{S}(t_j^+) \right)^{d_j} \prod_{l=1}^{m_j} \tilde{S}(t_{jl}) \right) \\ &= \prod_{j=1}^k \left( \left( \prod_{l=1}^{j-1} (1 - \tilde{\lambda}_l) - \prod_{l=1}^j (1 - \tilde{\lambda}_l) \right)^{d_j} \left( \prod_{l=1}^j (1 - \tilde{\lambda}_l) \right)^{m_j} \right) \\ &= \prod_{j=1}^k \left( \left( \prod_{l=1}^{j-1} (1 - \tilde{\lambda}_l) (1 - 1 + \tilde{\lambda}_j) \right)^{d_j} \prod_{l=1}^j (1 - \tilde{\lambda}_l)^{m_j} \right) \end{aligned}$$

$$\begin{aligned}
&= \prod_{j=1}^k \tilde{\lambda}_j^{d_j} \prod_{l=1}^{j-1} (1 - \tilde{\lambda}_l)^{d_j} \prod_{l=1}^j (1 - \tilde{\lambda}_l)^{m_j} \\
&= \prod_{j=1}^k \tilde{\lambda}_j^{d_j} (1 - \tilde{\lambda}_j)^{n_j - d_j}
\end{aligned} \tag{2.3}$$

So, we see that maximizing (2.2) on the space of all survival functions eventually reduces to the problem of finding  $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_k$  which maximize (2.3). When looking at the latter formula, we see that the factors are proportional to the PDFs of binomial distributions with parameters  $\tilde{\lambda}_j$ ,  $j = 1, 2, \dots, k$ . As we don't care about multiplicative constants in ML estimation, we can use what we know about the binomial ML estimate and maximize each of the factors by setting  $\tilde{\lambda}_j = d_j/n_j$ ,  $j = 1, 2, \dots, k$ . Due to the fact that maximizing the product in (2.3) is equivalent to maximizing each of the factors, we thus have a nonparametric MLE  $\tilde{S}$  of  $S_0$ , namely

$$\tilde{S}(t) := \prod_{\{j: t_j \leq t\}} \left(1 - \frac{d_j}{n_j}\right),$$

which agrees exactly with the definition of the Kaplan-Meier estimate  $\hat{S}$ .  $\square$

As we won't go into further details concerning the properties of the Kaplan-Meier estimator, we just state some results in a rather informal way and continue with an illustration of the theory outlined above by calculating  $\hat{S}$  for the epilepsy dataset.

**Remark 2.20.** Under relatively mild conditions, the following properties of the Kaplan-Meier estimator can be shown:

- (i)  $\hat{S}$  is uniformly consistent for  $S_0$  on the time interval  $[0, \tau]$ , where  $\tau$  denotes the end of the study period.



- (ii) For fixed  $t$ , the distribution of  $\hat{S}(t)$  can be approximated by a normal distribution with mean  $S_0(t)$  and estimated variance

$$\widehat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{\{j:t_j \leq t\}} \frac{d_j}{n_j(n_j - d_j)}.$$

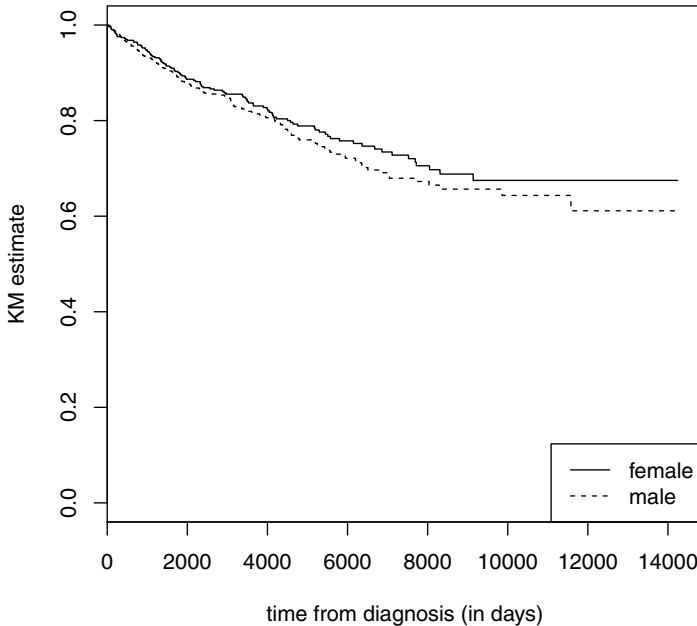
This expression for the asymptotic variance is known as *Greenwood's formula*.

For detailed information concerning these results and the corresponding proofs, we refer to Kalbfleisch and Prentice [14], pp. 17-18, 167-171.

**Example 2.21.** In Figure 2.1, we see the Kaplan-Meier estimates for female and male patients, respectively. To create such a plot, we basically used the `survreg` function contained in the `survival` package (Therneau [26]). Throughout the entire time range, the female patients have a slightly better chance of surviving than the males. By the way, there are formal hypothesis tests for the comparison of survival curves available, but we shall not discuss this topic here as it is covered by every standard textbook such as Klein and Moeschberger [16], pp. 205-214 or Kalbfleisch and Prentice [14], pp. 20-23.

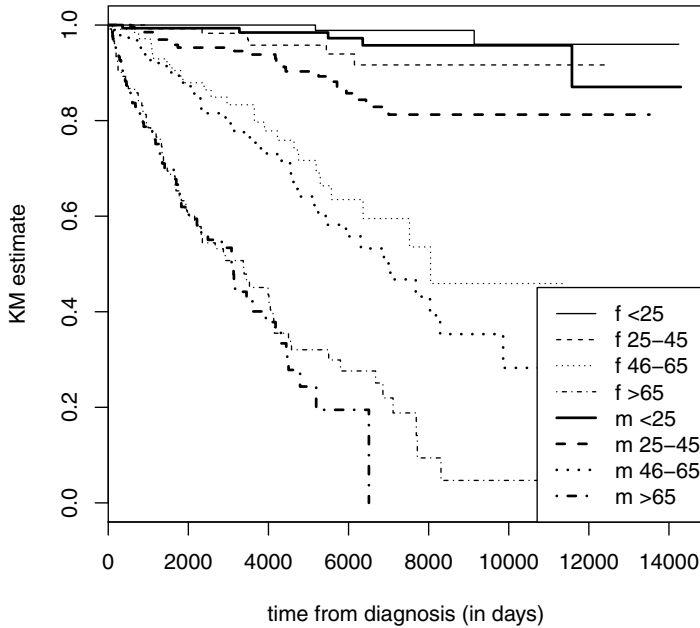
What we also see in the plot is the fact that, as mentioned above, the Kaplan-Meier estimate is undefined beyond the largest observed time if this is a censored one. Indeed, the largest observed survival times for female and male are 14245 d (approx. 39 y) and 14297 d (about 39.14 y), respectively, which are both censored observations. As a consequence, we can see in the plot that the step functions representing the KM estimates are not drawn beyond these time points.

However, from a medical point of view, it would be hardly reasonable if we only accounted for the gender of the patients. For example, we have seen in the introductory chapter that the range of the variable *age at diagnosis* is quite large. Therefore, taking further characteristics of the patients into consideration will certainly give a more realistic



**Figure 2.1:** Kaplan meier estimates for female and male patients

and reliable picture of the study cohort's survival experience. So, let's calculate age- and gender-specific KM estimates. But, due to the large amount of possible values of *age at diagnosis*, it would make no sense to consider each possible combination of *gender* and *age at diagnosis*. Thus, we must divide the observations into distinct age groups instead. E.g., building up the four age groups <25, 25-45, 46-65 and >65 seems to be a plausible choice since the lengths of these intervals are approximately equal. Moreover, we thus get 8 possible combinations of the two covariates we want to take into consideration, which is a number that should still work quite well when visualising



**Figure 2.2:** Kaplan meier estimates stratified according to age and gender

the corresponding KM estimate graphs in one single plot. The results are displayed in Figure 2.2.

To begin with, it is small surprise that the survival probability estimates decrease when we proceed from the first to the last age group. What's more interesting is the fact that the estimated survival curve for  $m > 65$  eventually arrives at 0 since in contrast to the other groups, the largest observed survival time is not censored. In this case, the Kaplan-Meier estimate is well defined on the whole positive x-axis.

Furthermore, as to the comparisons between male and female, the graphs for the age groups 46 – 65 and > 65 look quite nice and are thus easily interpretable: The women are better off than the men. This seems to hold true for the patients between 25 and 45, too. However, whereas the graph for the men of this age group provides sufficient information (we have 22 exact survival times, meaning that the KM estimate is a step function with 22 jumps) about the general outlook of the survival function, the corresponding estimate for the female patients is based on only 6 exact observations. Likewise, we can hardly make any meaningful statements based on the two graphs for the first age group since the number of jumps is 2 for female and 5 for male, respectively.

So, we see a very important point here: The fact that the KM estimate - by the way, the same problem would arise if we used the Nelson-Aalen approach - is discontinuous only at the observed exact survival times can cause serious problems, especially when we want to take several characteristics of the patients into account. Of course, we could choose the age intervals in a way such that we have a sufficiently large number of exact observations in each group. But, for example, if we combine the first and the second age groups to one single group for male and female patients, respectively, it is questionable if the results are still useful as far as their interpretability is concerned: Most likely, there is a striking difference between the survival experiences of a newborn and a 45-year-old patient. But, when merging these patients into one single group, we implicitly assume that such a difference does not exist. Moreover, recall that the patients were diagnosed in a certain year between 1970 and 2010. When looking at the corresponding population life tables for Tyrol (see chapter 5), one observes a great change in the survival experience in the course of those years. So, apparently, besides the covariates *gender* and *age at diagnosis*, we have to account for *year of diagnosis*, too. So, we must split up our original data (and, thus, especially the uncensored observations) into even more subsets to get reliable results. However, this certainly leads to further complications.

<http://www.springer.com/978-3-658-17718-8>

From Basic Survival Analytic Theory to a Non-Standard  
Application

Zimmermann, G.

2017, IX, 100 p. 9 illus., Softcover

ISBN: 978-3-658-17718-8