Chapter 2

Survival analysis

2.1 Basic concepts in survival analysis

This section describes basic aspects of univariate survival data and contains notation and important results which build the basis for specific points in later chapters.
We consider a single random variable X. Specifically, let X be non-negative, representing the lifetime of an individual. In that case this has given this field its name, the event is death, but the term is also used with other events, such as the onset of disease, complications after surgery or relapses in the medical field. In demography, time to death, but also time to leaving home, pregnancy, birth or divorce is of special interest. In industrial applications, it is typically time to failure of a technical unit. In economics, it can denote the time until the acceptance of jobs by unemployed, for example. Usually, X is assumed to be continuous, and we will restrict ourselves to this case in the present thesis. All functions in the sequel are defined over the interval $[0, \infty)$. The probability density function (p.d.f.) is denoted by f. The distribution of a random variable is completely and uniquely determined by its probability density function. But there many other notions exist, which are very useful in describing a distribution in specific situations. An important one is $F(x) = \mathbf{P}(X < x) = \int_0^x f(s)\, ds$, the cumulative distribution function (c.d.f.) of X. In survival analysis, one is more interested in the probability of an individual to survive to time x, which is given by the survival function

$$S(x) = 1 - F(x) = \mathbf{P}(X \geq x) = \int_x^\infty f(s)\, ds.$$

The major notion in survival analysis is the hazard function $\lambda(\cdot)$ (also called mortality rate, incidence rate, mortality curve or force of mortality), which is defined by

$$\lambda(x) = \lim_{\Delta \to 0} \frac{\mathbf{P}(x \leq X < x + \Delta | X \geq x)}{\Delta} = \frac{f(x)}{1 - F(x)}. \tag{2.1}$$

The hazard function characterizes the risk of dying changing over time or age. It specifies the instantaneous failure rate at time x, given that the individual survives until x. Sometimes, it is useful to deal with the cumulative (or integrated) hazard function

$$\Lambda(x) = \int_0^x \lambda(s)\, ds.$$

Especially for the topic covered by this thesis the notion of the Laplace transform $\mathbf{L}$ of a random variable X is crucial to inference in this area: $\mathbf{L}(s) = \mathbf{E}e^{-sX} = \int_0^\infty e^{-sx} f(x)\, dx$. The functions $f, F, S, \lambda, \Lambda$ and $\mathbf{L}$ give equivalent specifications of the distribution of the (non-negative) random variable X. It is easy to derive relations between the different notions, for example (2.1) implies that

$$\Lambda(x) = \int_0^x \lambda(s)\, ds = \int_0^x \frac{f(s)}{1 - F(s)}\, ds = -\ln(1 - F(x))$$

and consequently

$$S(x) = 1 - F(x) = e^{-\int_0^x \lambda(s)\, ds} = e^{-\Lambda(x)}. \tag{2.2}$$

As mentioned before, the hazard function is particularly useful in survival analysis, since it describes the way in which the instantaneous probability of failure for an individual changes with time. Applications often have qualitative information about the hazard function, which can be of help in selecting a model. For example, there may be reasons to restrict the analysis to models with increasing hazards or with hazard functions that have other well-defined characteristics. The shape of a hazard function can take different forms: it can be increasing, decreasing, constant, or U-shaped. Models with these and other hazard function shapes are all useful in practice: In demography, for example, following humans from birth to death, a U-shaped hazard function is often appropriate. After an initial period in which deaths result primarily from birth defects and infant diseases, the death rate drops and remains relatively constant until the age of 30 or so, after which it increases with age. This pattern is also observed in many other populations, including those consisting of technical items. Models with increasing hazards are used the most. This is because interest often centres on periods in the life of individuals in which measurable ageing takes place (for example old ages in humans). Models with a constant hazard function are of a very simple structure, as we will see in the next section. Less common are models with a decreasing hazard, but they are sometimes used to describe failure times of electronic devices, at least over a fairly long initial period of use. The main points to remember here are that the hazard function represents an aspect of a (non-negative) distribution that has a direct physical meaning and that qualitative information about the form of the hazard function is useful in selecting an appropriate model.

2.2 Parametric models

Now we shall consider in outline some distributions that are useful in the field of survival analysis. Naturally any distribution of non-negative random variables could be used to describe durations. The distributions to be discussed here are all continuous. Throughout the literature on survival analysis, certain parametric models have been used repeatedly such as exponential and Weibull models. These distributions have closed form expressions for survival and hazard functions. Log-normal and gamma distributions are generally less convenient computationally, but are still frequently applied. To avoid the model validity issues, the non-parametric approach, supported by the well-developed Kaplan-Meier-product limit estimator and related techniques, is often regarded as the preferred course. However, this alternative is often inefficient, as noted by Miller (1983). The pros and cons of different parametric, semi-parametric and non-parametric models and methodology for statistical inference can be found in the monographs by Kalbfleisch and Prentice (1980), Miller (1981), Lawless (1982), Cox and Oakes (1984) and Klein and Moeschberger (1997).

Below we discuss some of the standard failure time models for homogeneous populations. The properties and the theoretical bases of these distributions are considered here only briefly. The distributions will be studied in the simplest case of independently and identically distributed random variables. In this and the following sections the random variable X denotes the lifetime which we are interested in making inferences about.

2.2.1 Exponential distribution

The exponential model ($X \sim Exp(\lambda)$) is the simplest parametric model and assumes a constant risk over time, which reflects the property of the distribution appropriately called 'lack of memory'. The probability to die within a particular time interval depends only on the length but not on the location of this interval. This means that the distribution of $X - x$ conditional on $X \geq x$ is the same as the original distribution. In other words, it holds that

$$\mathbf{P}(x \leq X < x + \delta | X \geq x) = \mathbf{P}(X < \delta)$$

for any positive δ. As a consequence, the exponential distribution (as the only one) is not influenced by the definition of time zero. The parameter λ attains all positive values and the distribution with $\lambda = 1$ is called the unit or standard exponential. Therefore,

the following formulae can be derived by some simple algebraic calculations:

$$\begin{aligned}
\text{probability density function} \quad & f(x) = \lambda e^{-\lambda x} \\
\text{survival function} \quad & S(x) = e^{-\lambda x} \\
\text{hazard function} \quad & \lambda(x) = \lambda, \qquad \lambda > 0 \\
\text{cumulative hazard function} \quad & \Lambda(x) = \lambda x \\
\text{mean} \quad & \mathbf{E}X = \frac{1}{\lambda} \\
\text{variance} \quad & \mathbf{V}(X) = \frac{1}{\lambda^2}
\end{aligned}$$

The exponential distribution was widely used in early work on the reliability of electronic components and technical systems. The distribution of cX with a positive constant c is again exponentially distributed with parameter λ/c. The minimum of n independent exponential random variables with parameter λ is still exponential with parameter $n\lambda$:

$$\mathbf{P}(\min\{X_1, \ldots, X_n\} \geq x) = \prod_{i=1}^{n} \mathbf{P}(X_i \geq x) = \prod_{i=1}^{n} e^{-\lambda x} = e^{-n\lambda x}.$$

The model is very sensitive to even a modest variation because it has only one adjustable parameter, the inverse of which is both mean and standard deviation. Recent works have overcome this limitation by using more flexible distributions.

2.2.2 Weibull distribution

The Weibull model (introduced by Waloddi Weibull in 1939) is an important generalization of the exponential model with two positive parameters. The second parameter in the model allows great flexibility of the model and different shapes of the hazard function. The convenience of the Weibull model for empirical work stems on the one hand from this flexibility and on the other from the simplicity of the hazard and survival function.

$$\begin{aligned}
\text{probability density function} \quad & f(x) = \lambda\gamma x^{\gamma-1} e^{-\lambda x^{\gamma}} \\
\text{survival function} \quad & S(x) = e^{-\lambda x^{\gamma}} \\
\text{hazard function} \quad & \lambda(x) = \lambda\gamma x^{\gamma-1} \\
\text{cumulative hazard function} \quad & \Lambda(x) = \lambda x^{\gamma} \\
\text{mean} \quad & \mathbf{E}X = \lambda^{-\frac{1}{\gamma}} \Gamma(1 + \frac{1}{\gamma}) \\
\text{variance} \quad & \mathbf{V}(X) = \lambda^{-\frac{2}{\gamma}} \big(\Gamma(1 + \frac{2}{\gamma}) - \Gamma(1 + \frac{1}{\gamma})^2\big)
\end{aligned}$$

where Γ denotes the Gamma function with $\Gamma(k) = \int_0^\infty s^{k-1}e^{-s}\,ds$ $(k > 0)$. We abbreviate the distribution as $Weibull(\lambda, \gamma)$. In the case of $\gamma = 1$, the exponential distribution is obtained. The hazard function decreases monotonously from ∞ at time zero to zero at time ∞ for $\gamma < 1$, constant (exponential distribution) for $\gamma = 1$ and it monotonously increases from zero at time zero to ∞ at time ∞ for $\gamma > 1$.

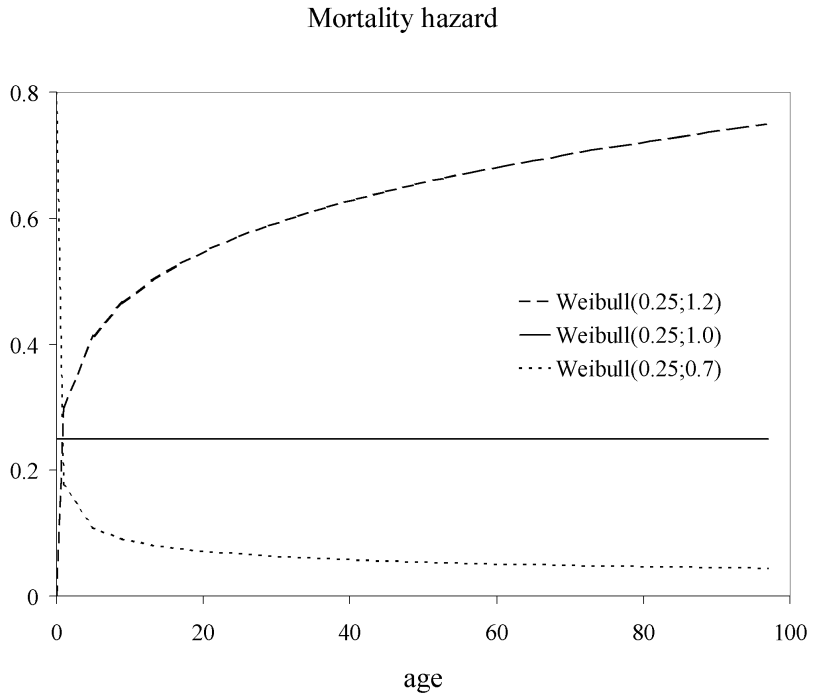

Figure 2.1: Weibull hazard functions with different shape parameters.

If $X \sim Weibull(\lambda, \gamma)$, then it holds that $cX \sim Weibull(\lambda c^{-\gamma}, \gamma)$, when c is a positive constant. Furthermore, the minimum of n i.i.d. variables from this distribution is $Weibull(n\lambda, \gamma)$ (minimum-stable distribution). The Weibull distribution can also be generated as the limiting distribution of the minimum of a sample from a continuous distribution with support on $[0, u)$ for some u $(0 < u < \infty)$. This extreme value character makes the Weibull distribution appropriate for the distribution of individual time to death, because there are different causes of death which compete with each other and the first one to strike will kill the individual.

The Weibull hazard has been theoretically derived for cancer incidence by Pike (1966), but it is unknown whether it has relevance for other diseases. The Weibull distribution is inappropriate when the hazard rate is indicated to be unimodal or bathtub-shaped. A generalization of the Weibull distribution to include such kind of shapes was proposed by Mudholkar et al. (1996).

2.2.3 Gompertz distribution

In 1825 the British actuary Benjamin Gompertz made a simple but important observation that a law of geometrical progression pervades large portions of different tables of mortality for humans. The simple formula he derived describing the exponential rise in death rates between sexual maturity and old age is commonly referred to as the Gompertz equation–a formula that remains a valuable tool in demography and in other scientific disciplines. Gompertz's observation of a mathematical regularity in human life tables led him to believe in the presence of a law of mortality that explained why common age patterns of death exist. It has been widely used, especially in actuarial and biological applications and in demography. A random variable follows a Gompertz distribution with parameters $a > 0$ and $b > 0$ ($X \sim Gompertz(a, b)$), if the following relations hold:

$$\begin{aligned} \text{probability density function} \quad & f(x) = ae^{bx}e^{-\frac{a}{b}(e^{bx}-1)} \\ \text{survival function} \quad & S(x) = e^{-\frac{a}{b}(e^{bx}-1)} \\ \text{hazard function} \quad & \lambda(x) = ae^{bx} \\ \text{cumulative hazard function} \quad & \Lambda(x) = \frac{a}{b}(e^{bx} - 1) \end{aligned}$$

The hazard function is increasing from a at time zero to ∞ at time ∞. The model can be generalized to the Gompertz-Makeham distribution by adding a constant to the hazard: $\lambda(x) = ae^{bx} + c$.

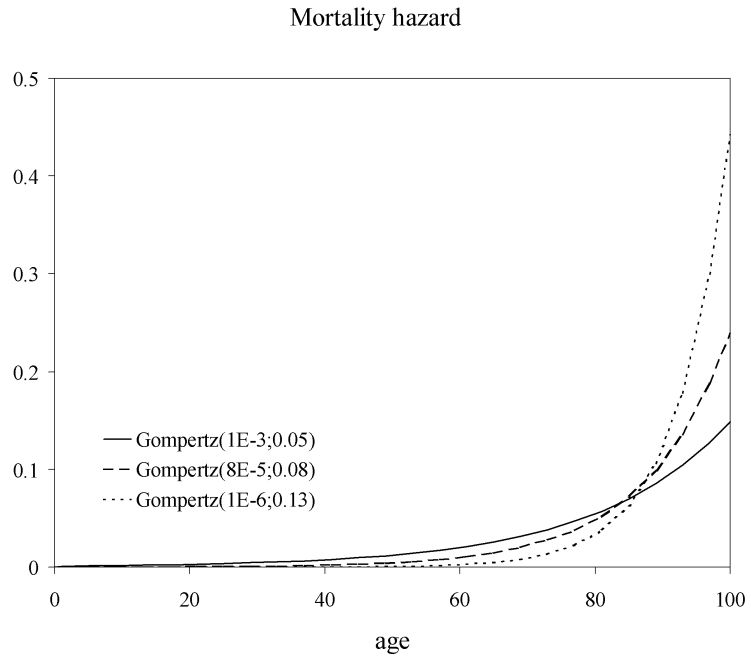

Figure 2.2: Gompertz hazard functions with different parameters.

2.2.4 Log-logistic distribution

An alternative model to the Weibull distribution is the log-logistic distribution. The log-logistic distribution has a fairly flexible functional form, it is one of the parametric survival time models in which the hazard rate may be decreasing, increasing, as well as hump-shaped, that is it initially increases and then decreases. The distribution imposes the following functional forms on the density, survival, hazard and cumulative hazard function:

$$\begin{aligned}
\text{probability density function} \quad & f(x) = abx^{b-1}(1+ax^b)^{-2} \\
\text{survival function} \quad & S(x) = (1+ax^b)^{-1} \\
\text{hazard function} \quad & \lambda(x) = \frac{abx^{b-1}}{1+ax^b} \\
\text{cumulative hazard function} \quad & \Lambda(x) = \ln(1+ax^b)
\end{aligned}$$

The general shape of the hazard function of a log-logistic distribution is very similar to that of the log-normal distribution considered later. The log-logistic distribution can be obtained as a mixture of Gompertz distributions with a gamma distributed mixture variable with mean and variance equal to one.

2.2.5 Gamma distribution

The gamma distribution includes the exponential distribution as a special case. The gamma distribution is of limited use in survival analysis because the gamma models do not have closed form expressions for survival and hazard functions. Both include the incomplete gamma integral

$$I_k(x) = \frac{\int_0^x s^{k-1}e^{-s}\,ds}{\Gamma(k)}.$$

Consequently, traditional maximum likelihood estimation is difficult and requires the calculation of such incomplete gamma integrals, which imposes additional numerical problems in parameter estimation. A random variable X is gamma distributed with parameter k and λ ($X \sim \Gamma(k,\lambda)$), if the following holds:

$$\begin{aligned}
\text{probability density function} \quad & f(x) = \frac{\lambda^k x^{k-1}e^{-\lambda x}}{\Gamma(k)} \qquad k,\lambda > 0 \\
\text{survival function} \quad & S(x) = 1 - I_k(\lambda x)
\end{aligned}$$

$$\begin{aligned}
\text{hazard function} \quad & \lambda(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(1 - I_k(\lambda x))\Gamma(k)} \\
\text{mean} \quad & \mathbf{E}X = \frac{k}{\lambda} \\
\text{variance} \quad & \mathbf{V}(X) = \frac{k}{\lambda^2} \\
\text{Laplace transform} \quad & \mathbf{L}(s) = \mathbf{E}e^{-Xs} = (1 + \frac{s}{\lambda})^{-k}
\end{aligned}$$

If $k = 1$, the gamma distribution is reduced to the exponential distribution. With integer k, the gamma distribution is sometimes called a special Erlangian distribution. It can be derived as the distribution of waiting time to the k-th emission from a Poisson source with intensity parameter λ. Consequently, the sum of k independent exponential variates with parameter λ has a gamma distribution with parameters k and λ (see Example 1) and can be used to model life times of technical systems with repeated repairing after failure.

Example 1 *Let $X_1, X_2, \ldots, X_k$ denote k independent exponential distributed random variables with $X_i \sim Exp(\lambda)$ $(i = 1, \ldots, k)$ and introduce X by $X = X_1 + \ldots + X_k$. Then it holds that*

$$\begin{aligned}
\mathbf{L}(s) &= \mathbf{E}e^{-Xs} = \mathbf{E}e^{-(X_1 + \ldots + X_k)s} \\
&= \prod_{i=1}^{k} \mathbf{E}e^{-X_i s} = \prod_{i=1}^{k} \mathbf{E}(1 + \frac{s}{\lambda})^{-1} = (1 + \frac{s}{\lambda})^{-k}.
\end{aligned}$$

An extension of this idea was used by Aalen (1992) dealing with the compound Poisson distribution (see Section 3.6). Despite the fact that the gamma distribution is of limited value as a life time distribution because of the problems mentioned above, the gamma distribution is a widely used frailty (mixture) distribution because of some neat mathematical features. It is mathematically tractable and readily computable. It is a flexible distribution that takes on a variety of different shapes as k varies. Furthermore, frailty cannot be negative and the gamma distribution is, along with the log-normal and Weibull distribution, one of the most commonly used distributions to model positive random variables. The assumption that frailty is gamma distributed yields some useful mathematical results, including that the frailty among survivors of any age x is again gamma distributed and frailty among those who die at any time x, too (see Vaupel et al., 1979). We will discuss other properties of the gamma distribution as frailty distribution later in more detail.

2.2.6 Log-normal distribution

In the log-normal model ($X \sim logN(m, s^2)$), the natural logarithm $\ln(X)$ of the lifetime X is assumed to be normally distributed ($ln(X) \sim N(m, s^2)$). A log normal distribution results when the variable is the product of a large number of independent, identically distributed variables in the same way that a normal distribution results when the variable is the sum of a large number of independent, identically distributed variables. The survival and hazard functions include the incomplete normal integral

$$\Phi(x) = \int_{-\infty}^{x} \phi(s)\, ds,$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ denotes the probability density function of a standard normal distribution. Consequently,

$$\begin{aligned} \text{probability density function} \quad & f(x) = \frac{1}{\sqrt{2\pi} s x} e^{-\frac{(\ln(x)-m)^2}{2s^2}} \\ \text{mean} \quad & \mathbf{E}X = e^{m+\frac{s^2}{2}} \\ \text{variance} \quad & \mathbf{V}(X) = e^{2m+s^2}(e^{s^2} - 1) \end{aligned}$$

The log-normal distribution may be convenient to use with non-censored data, but when this distribution is applied to censored data, the computations quickly become formidable. Unfortunately, the hazard function has a strange form: it has value zero at $x = 0$, increases to a maximum and then decreases, approaching zero as x heads to infinity. Because of the decreasing form of the hazard function for older ages, the distributions seem implausible as a lifetime model in most situations. Nevertheless, it makes sense if interest is focused on time periods of younger ages. Despite its unattractive features, the log-normal distribution has been widely used as failure distribution in diverse situations, such as the analysis of electrical insulation or time to occurrence of lung cancer among smokers.

Furthermore, the log-normal distribution has often been used as a frailty (mixing) distribution. Especially in the context of unobserved normal distributed covariates in the Cox model, the log-normal frailty distributions provides an appealing interpretation of the model. Unfortunately, the Laplace transform is intractable, and therefore numerical integration is needed for probability results. The log-normal distributions are in practice very close to the inverse Gaussian distributions.

2.3 Censoring

Censoring is what distinguishes survival analysis from other fields of statistics. Basically, a censored observation contains only partial information about the variable of interest. There are different types of censoring, here we consider type I right censoring only.

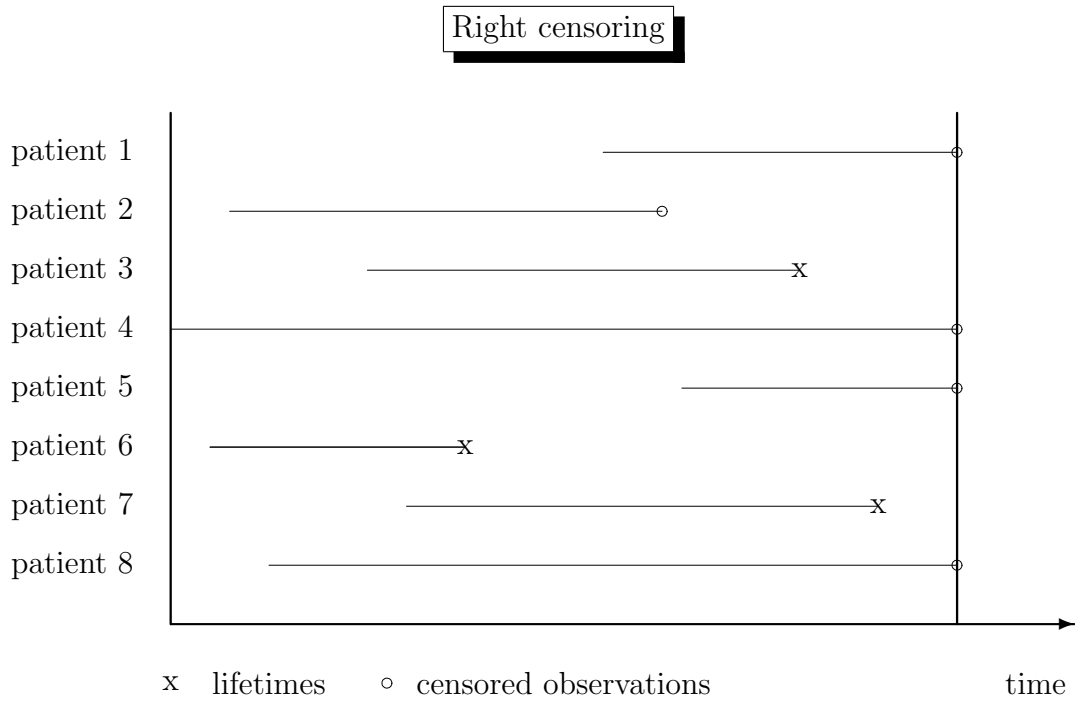

Figure 2.3: Right censored lifetimes of patients in an artificial clinical trial.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. survival times with cumulative distribution function F and let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. censoring times with cumulative distribution function G. Throughout the thesis, we assume that F and G are absolutely continuous. Furthermore, let f and g be probability density functions with respect to F and G. We can only observe $(T_1, \Delta_1), (T_2, \Delta_2), \ldots, (T_n, \Delta_n)$, where $T_i = \min\{X_i, Y_i\}$ and

$$\Delta_i = \begin{cases} 1 & : \text{ if } X_i \leq Y_i, \quad \text{that is, } T_i \text{ is not censored} \\ 0 & : \text{ if } Y_i < X_i, \quad \text{that is, } T_i \text{ is censored.} \end{cases} \tag{2.3}$$

Random censoring arises especially in medical applications, for example in clinical trials or epidemiological studies. Here, patients may enter the study at different times; then each is treated with one of several possible drugs or therapies. We are interested in observing their lifetimes, but censoring occurs in one of the following forms:

- Loss to follow up. The patient may move elsewhere; he is never seen again.
- Drop out. The treatment may have such strong side effects that it is necessary to stop the therapy. Or the patient may refuse to continue the treatment.
- Termination of the study. The study ends at a predefined point of time. This type of censoring is called administrative censoring.

We use X and Y, with no subscripts, as shorthand for all the X_i and Y_i variables.

Assumption: Lifetimes X and censoring times Y are independent. A weaker condition is to assume that censoring is non-informative.
Remark: The cumulative distribution function of the non-censored observations (while discarding the censored observations of the sample) is not F!

$$\begin{aligned}\mathbf{P}(T<t,\Delta=1) &= \mathbf{P}(X<t,X\le Y)=\int\limits_{x<t,x\le y} f(x)g(y)\,dx\,dy \\ &= \int\limits_{x<t} f(x)\Big(\int\limits_{x\le y} g(y)\,dy\Big)\,dx=\int\limits_{x<t} f(x)(1-G(x))\,dx \ne F(t) \quad (2.4)\end{aligned}$$

Theorem 1 *(Wienke (1996)) The probability density function of the data* (T,Δ) *takes the form*

$$f(t,\delta)=(f(t)(1-G(t)))^{\delta}(g(t)(1-F(t)))^{1-\delta}. \quad (2.5)$$

Proof: Denote by H_0 and H_1 sub-distribution functions and by h_0 and h_1 sub-densities. It holds (see (2.4)) $H_1(t)=\mathbf{P}(T<t,\Delta=1)=\int\limits_{x<t} f(x)(1-G(x))\,dx$. Furthermore,

$$\begin{aligned}H_0(t) &= \mathbf{P}(T<t,\Delta=0)=\mathbf{P}(Y<t,Y<X) \\ &= \int\limits_{y<t,y<x} f(x)g(y)\,dx\,dy=\int\limits_{y<t} g(y)\Big(\int\limits_{y<x} f(x)\,dx\Big)\,dy=\int\limits_{y<t} g(y)(1-F(y))\,dy\end{aligned}$$

Consequently,

$$\begin{aligned}h_0(t)&=H_0'(t)=g(t)(1-F(t)) \\ h_1(t)&=H_1'(t)=f(t)(1-G(t))\end{aligned}$$

$$\begin{aligned}f(t,\delta) &= \delta h_1(t)+(1-\delta)h_0(t)=(h_1(t))^{\delta}(h_0(t))^{1-\delta} \\ &= (f(t)(1-G(t)))^{\delta}(g(t)(1-F(t)))^{1-\delta}.\end{aligned}$$

This completes the proof. ■

Remark: If the censoring is non-informative, meaning if the censoring distribution does not contain any information about the parameters of interest, then it does not enter the likelihood function:

$$L(t,\delta) = f(t)^{\delta}(1-F(t))^{1-\delta} = \delta f(t) + (1-\delta)(1-F(t)). \tag{2.6}$$

As pointed out in Theorem 1, the density function under independent right censoring is

$$f(t,\delta) = \delta f(t)(1-G(t)) + (1-\delta)g(t)(1-F(t)).$$

The following example considers the case of dependent censoring. It turns out that the likelihood function under dependent censoring is a composition of derivatives of the joint survival function of life times and censoring times.

Example 2 *Denote by* (T,Δ), $T = \min\{X,Y\}$, $\Delta = 1(X \leq Y)$ *censored observations under the assumption of dependent censoring. Let* $S(x,y)$ *and* $f(x,y)$ *be the joint survival and probability density function of* X *and* Y*, respectively. Consequently, the sub-distribution functions can be derived as follows:*

$$\begin{aligned} H_1(t) &= \mathbf{P}(T<t, \Delta=1) = \mathbf{P}(X<t, X\leq Y) \\ &= \int\int_{\{x<t, x\leq y\}} f(x,y)\,dx\,dy = -\int_0^t S_x(x,x)\,dx. \end{aligned}$$

This implies that the sub-density of a non-censored $(\delta = 1)$ *observation is a derivative of the sub-distribution function:*

$$h_1(t) = H_1'(t) = -S_x(t,t) = -\frac{\partial S(x,y)}{\partial x}|_{x=t,y=t}.$$

Similar calculations yield the sub-distribution and sub-density functions in the case of a censored observation $(\delta = 0)$*:*

$$\begin{aligned} H_0(t) &= \mathbf{P}(T<t, \Delta=0) = \mathbf{P}(Y<t, Y<X) \\ &= \int\int_{\{y<t, y<x\}} f(x,y)\,dxdy = -\int_0^t S_y(y,y)\,dy \end{aligned}$$

and

$$h_0(t) = H_0'(t) = -S_y(t,t) = -\frac{\partial S(x,y)}{\partial y}|_{x=t,y=t}.$$

The likelihood function is a composition of the sub-density functions:

$$L(t,\delta) = -\delta S_x(t,t) - (1-\delta)S_y(t,t).$$

2.4 Truncation

Now we shall take truncation into account. We restrict our treatment to the most common type of truncation, that is left truncation. Furthermore, let us assume that truncation is non-random. Left truncation occurs when individuals come under observation only some known time after the natural origin of the event under study. That is, had the individual failed before the truncation time in question, that individual would not have been recorded. For example, the second patient in Figure 2.4 cannot be observed, because

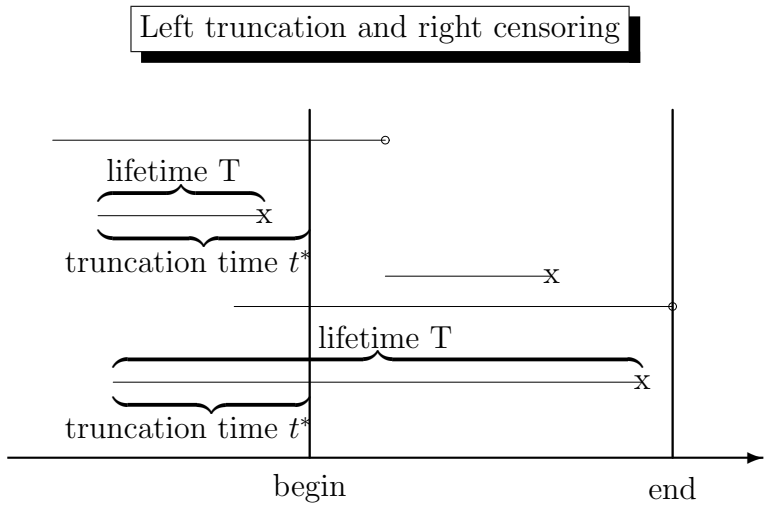

Figure 2.4: Left truncated and right censored lifetimes.

he dies before the study started. That means, the person carrying out the study would not be aware of that observation. In other words, truncation is sampling from a conditional distribution. Denote observations by (T^*, Δ^*) with $\mathcal{L}(T^*, \Delta^*)$=$\mathcal{L}((T, \Delta)|T \geq t^*)$ with t^* as known (non-random) truncation time. So we get

$$\begin{aligned} \mathbf{P}(T^* \geq t, \Delta^* = \delta) &= \mathbf{P}(T \geq t, \Delta = \delta | T \geq t^*) \\ &= \frac{\mathbf{P}(T \geq t, \Delta = \delta)}{\mathbf{P}(T \geq t^*)} = \frac{\mathbf{P}(T \geq t, \Delta = \delta)}{(1 - F(t^*))(1 - G(t^*))} \end{aligned}$$

and the density of the non-random left truncated and right censored observations:

$$\begin{aligned} f(t, \delta, t^*) &= \frac{h(t, \delta)}{(1 - F(t^*))(1 - G(t^*))} \\ &= \delta \frac{f(t)(1 - G(t))}{(1 - F(t^*))(1 - G(t^*))} + (1 - \delta) \frac{g(t)(1 - F(t))}{(1 - F(t^*))(1 - G(t^*))}. \end{aligned}$$

Hence, the likelihood function in the case of independent (non-informative) right censoring and non-random left truncation at time t^* can be written as

$$L(t, \delta, t^*) = \delta \frac{f(t)}{1 - F(t^*)} + (1 - \delta) \frac{1 - F(t)}{1 - F(t^*)}.$$

2.5 Non-parametric and semi-parametric models

For parametric inference, it is necessary to make assumptions about the distribution of failure times. In some circumstances this makes sense; for example, when additional information about the nature of the ageing or disease process is available from other experiments. When one is interested in avoiding such assumptions, it is common to use non-parametric models. The simplest non-parametric estimate of a distribution function is the empirical distribution function. That means, even in the case of continuous distributions we estimate it by a discrete distribution. Major steps in the development of appropriate methods in survival analysis (in censored observations) were the introduction of the Kaplan-Meier estimator (Kaplan and Meier (1958)) and the proportional hazards model (Cox 1972).

A key question is whether we should use a parametric model as described above or a non-parametric one. An advantage of non-parametric models is their good fit and the resulting ability to deal with any distribution without any additional assumptions. But there is a high price to pay. First, non-parametric methods need much more data to get reasonable results. Second, it is very hard to get estimates of the hazard function, which is often the most interesting and relevant information. For this, it is necessary to smooth out the discrete point masses of the Kaplan-Meier estimator, for example by kernel function smoothing. Otherwise, parametric models often allow closed form expressions of the hazard function (depending on the chosen model) and other characteristics of the failure distribution. Furthermore, parametric models can be described by the values of a few parameters and they often give good results even in the case of moderate sample size.

2.5.1 Kaplan-Meier estimator

A useful way of characterizing the survival in a homogeneous group of individuals is to compute and graph the empirical survival function. If there are no censored observations in the sample, the empirical survival function at time t is the ratio of survivors at time t and the sample size n. This step function decreases by $\frac{1}{n}$ just after each observed failure (for ease of presentation we assume no ties here). When dealing with censored data, a methodology for handling this with convenience is required. Remember that we observe the pairs $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$, where $T_i = \min\{X_i, Y_i\}$ and $\Delta = 1(X_i \leq Y_i)$. Let $T_{(1)} < T_{(2)} < \ldots < T_{(n)}$ be the order statistics of $T_1, T_2, \ldots, T_n$, and with an abuse of notation define $\Delta_{(i)}$ to be the value of Δ which is associated with $T_{(i)}$, that is, $\Delta_{(i)} = \Delta_j$ if $T_{(i)} = T_j$. Note that the $\Delta_{(1)}, \Delta_{(2)}, \ldots, \Delta_{(n)}$ are not ordered. In the sequel, we will

use capital letters to denote random variables, for example (T_i, Δ_i), and small letters for their real-value realizations: (t_i, δ_i). The Kaplan-Meier estimator (also called product limit estimator) was introduced by Kaplan and Meier (1958) as

$$\hat{S}(t) = \prod_{i:T_{(i)}<t} \Big(1 - \frac{\Delta_{(i)}}{n-i+1}\Big).$$

This function is a decreasing step function, with changes only at times of death. A slightly problematic point is that $\hat{S}$ never reduces to zero if the largest observation is a censored one, i.e. $\Delta_{(n)} = 0$. In this case, $\hat{S}$ is usually left unspecified for $t > t_{(n)}$.

2.5.2 Proportional hazards model

The models presented above deal with the simplest case of i.i.d. data. This implies a homogeneous population. However, in most practical applications the population under study is not homogeneous. For example, individuals in epidemiological studies may differ in age (if age is not used as time scale), gender, socio-economic status, education, blood pressure, body mass index, smoking habits, nutrition, physical activity level, heart rate and so forth. Maybe some of these covariates are of special interest, such as the effect of a treatment in a clinical trial, or they are nuisance parameters which influence the variable lifetime. The proportional hazards model is a regression model with duration as dependent variable. It allows inclusion of information about known (observed) covariates in models of survival analysis and is the most applied model in this area. Statistical strategies for prediction are similar to those utilized in ordinary regression. However, the details for regression techniques in survival analysis are unique.
Let $\lambda(t, X)$ denote the hazard of an individual at time (or age) t with covariate vector $X = (X_1, \ldots, X_k)$. The proportional hazards model (Cox 1972) specifies that

$$\lambda(t, X) = \lambda_0(t) g(X) \tag{2.7}$$

where $\lambda_0(t)$ is the baseline hazard function and g some positive function. The model assumes a baseline hazard (risk of death or other event) that is common to all the individuals in the study population. The parameters of primary interest are contained in $g(X) = g(\beta, X)$, often

$$g(X) = e^{\beta^T X} = e^{\sum_{i=1}^{k} \beta_i X_i}.$$

In this model, covariates act multiplicatively on the baseline hazard, adding additional risks on an individual basis, as determined by the individuals' prognostic information. This gives the model a simple and easily understood interpretation. The main idea behind it is the separation of the age or time effect in the baseline hazard function on one

side and the effect of the covariates in an exponential term on the other side. In essence this assumption says that the hazard $\lambda(t)$ of failure at time t is related to individuals or groups of individuals by a proportionality constant which does not depend on t. The simple two-sample situation is obtained by letting X_1 be 0 or 1 ($k = 1$), depending on group membership. In this case, the method is truly non-parametric and e^{β} is the hazard ratio for mortality between the two groups. However, when X takes more than two values, a parametric form of g is required. Now inference is dependent on that form but still independent of $\lambda_0(t)$, and one speaks of a semi-parametric model. The conditional survival function for T given X is

$$S(t|X) = S_0(t)^{e^{\sum_{i=1}^{k} \beta_i X_i}},$$

where $S_0(t) = e^{-\int_0^t \lambda_0(s)\,ds}$ and the $\beta' s$ are unknown regression parameters. That means the survival function of an individual with covariates X is the baseline survival function raised to a power. The class of distributions generated by this procedure is sometimes called Lehmann alternatives.

Two different approaches are possible. In the parametric case the baseline hazard is chosen in the class of parametric lifetime distributions, for example as Weibull or Gompertz-Makeham. But the model also works without any specification of the baseline hazard function. In this second case, the model is natural and sufficiently flexible to suit many purposes. Since $e^{\sum_{i=1}^{k} \beta_i X_i}$ is always positive, the individual hazard $\lambda(t, X)$ is automatically non-negative for all t and all $\beta' s$. One additional reason for considering this model is that censoring and competing risks are relatively easily accommodated within this formulation and in particular the technical problems of statistical inference have a simple solution when the baseline hazard is arbitrary. Cox (1975) suggests using a partial likelihood approach in the case of arbitrary baseline hazard. Inference for the Cox estimator is almost exclusively based on asymptotic results (Andersen and Gill 1982). The validity of these large sample properties have been found acceptable with moderately large sample sizes, moderate amount of censoring and balanced covariate distributions. However it frequently occurs that covariates have very skew distributions, for example when only a small fraction of the individuals are exposed to the risk factor of interest. It is also very common that a large fraction of lifetimes is censored. Especially in large cohort studies analyzing the effect of a rare exposition on an event, the number of exposed cases may be very small. One may then question the validity of inference based on large sample results. Samuelsen (2003) investigates the possibilities and limitations of exact inference in the proportional hazards model.

Note that covariate X could vary with time, but this is beyond the scope of this thesis.