When vectors really are used as vectors they are to be considered *column* vectors. Often however we just use vector notation to indicate a collection of objects, and write, e.g., $\mathbf{N} = (N_1, \ldots, N_k)$, whereby we do not distinguish between a row or column vector.

The transpose of a vector is indicated by the superscript $^\mathsf{T}$; t usually denotes a time (also s, u, v, τ); $\mathscr{T}$ is a collection of times (the interval $[0, \tau)$ or $[0, \tau]$, where $\tau = \infty$ is also allowed); T is usually a random time (a stopping time, in fact). There is nothing special about the initial time 0, this could also have been replaced by an arbitrary value σ. The indicator function of a time interval $(s, t]$ is denoted $I_{(s,t]}$.

The fixed time interval $\mathscr{T} = [0, \tau)$ or $[0, \tau]$ will usually be in the background at any point in the book. When an integral $\int \cdots$ without limits on the integral sign is mentioned, we mean the *function* of time t, obtained by integrating over the interval $[0, t]$ for each $t \in \mathscr{T}$. The same convention applies to the product-integral $\prod \cdots$.

F and f, A and α, Λ, and λ, usually denote a function and its density or derivative (also in vector or matrix versions). ΔF is the function giving the jumps of F, $\Delta F(t) = F(t) - F(t-)$, supposing F to be right-continuous. In particular therefore, $\int \cdots F(\mathrm{d}t) = \int \cdots f(t)\,\mathrm{d}t$ if F has a density f, whereas $\int \cdots F(\mathrm{d}t) = \sum \cdots \Delta F(t)$ if F is a step-function.

$D(\mathscr{T})$ is the space of right-continuous functions with left-hand limits on $\mathscr{T}$; the so-called *cadlag* functions. The space $D(\mathscr{T})$ itself is often called the Skorohod space, reflecting the usual choice of metric on this space in classical weak convergence theory (Billingsley, 1968; Pollard, 1984).

The notations $\xrightarrow{\mathscr{D}}$ and $\xrightarrow{\mathrm{P}}$ stand for convergence in distribution and probability, respectively.

The notations $\langle \cdot \rangle$ and $[\cdot]$ for predictable and optional variation process are introduced in Section II.3. The context should always show when a time interval $[s, t]$ and when an optional covariation or "square bracket" process $[M, M']$ is meant.

II.1. An Informal Introduction to the Basic Concepts

In subsequent sections of this chapter, we shall give a formal survey of the theory which we will use to analyze statistical models based on counting processes. By way of introduction, however, we will first discuss the key concepts in a very informal manner in the context of one of the basic examples. At the same time, we will see how the mathematical background lends itself beautifully to the treatment of statistical models formulated in terms of the *hazard rate* and possibly involving *censoring*; thus, also introducing topics treated in depth in Chapters III and IV.

Consider first a sample of n (uncensored) continuously distributed survival times $X_1, \ldots, X_n$ from a survival function S with hazard rate function α; thus,

$\alpha = f/(1 - F)$ where $F = 1 - S$ is the distribution function and f the density of the X_i. The hazard rate α completely determines the distribution through the relations

$$S(t) = \mathrm{P}(X_i > t) = \prod_0^t [1 - \alpha(s)\,\mathrm{d}s] = \exp\left(-\int_0^t \alpha(s)\,\mathrm{d}s\right),$$

where the product-integral $\prod(1 - \alpha)$ is explained later in this section, see (2.1.13), and studied in detail in Section II.6. One can interpret α by the heuristic

$$\mathrm{P}(X_i \in [t, t + \mathrm{d}t) | X_i \geq t) = \alpha(t)\,\mathrm{d}t. \tag{2.1.1}$$

We will consider the nonparametric estimation of the hazard rate or, rather, the cumulative or integrated hazard rate

$$A(t) = \int_0^t \alpha(s)\,\mathrm{d}s. \tag{2.1.2}$$

Typically, in survival analysis problems, complete observation of $X_1, \ldots, X_n$ is not possible. Rather, one only observes $(\tilde{X}_i, D_i)$, $i = 1, \ldots, n$, where D_i is a "censoring indicator," a zero-one valued random variable describing whether X_i or only a lower bound to X_i is observed (really D_i indicates uncensored):

$$X_i = \tilde{X}_i \quad \text{if } D_i = 1,$$
$$X_i > \tilde{X}_i \quad \text{if } D_i = 0.$$

We shall consider $\tilde{X}_1, \ldots, \tilde{X}_n$ as random times; at these times, the value of the corresponding D_i becomes available, and we know whether the corresponding event is a failure or a censoring. Thus, all n survival periods start together at time $t = 0$.

As an example, Figures II.1.1 and II.1.2 depict the observations of 10 randomly selected patients from the data on survival with malignant melanoma introduced in Example I.3.1: First, in the original calendar time scale, and second, in the survival time scale t years since operation. This latter time scale is the one we concentrate on in our illustration of stochastic process concepts. A filled circle corresponds to $D_i = 1$ (a failure), an open circle to $D_i = 0$ (a censoring).

Further analysis is impossible without assumptions on censoring. We will make the most general assumption which still allows progress: the assumption of *independent censoring* (to which we return in Section III.2.2), which means that at any time t (in the survival time scale) the survival experience in the future is not statistically altered (from what it would have been without censoring) by censoring and survival experience in the past. To formalize this notion, we must be able to talk mathematically about past and future. This will be done through the concept of a *filtration* or *history* $(\mathscr{F}_t)_{t \geq 0}$, $\mathscr{F}_t$ representing the available data at time t. Write $\mathscr{F}_{t-}$ correspondingly for the available data just before time t. A specification of $(\mathscr{F}_t)_{t \geq 0}$ can only be done

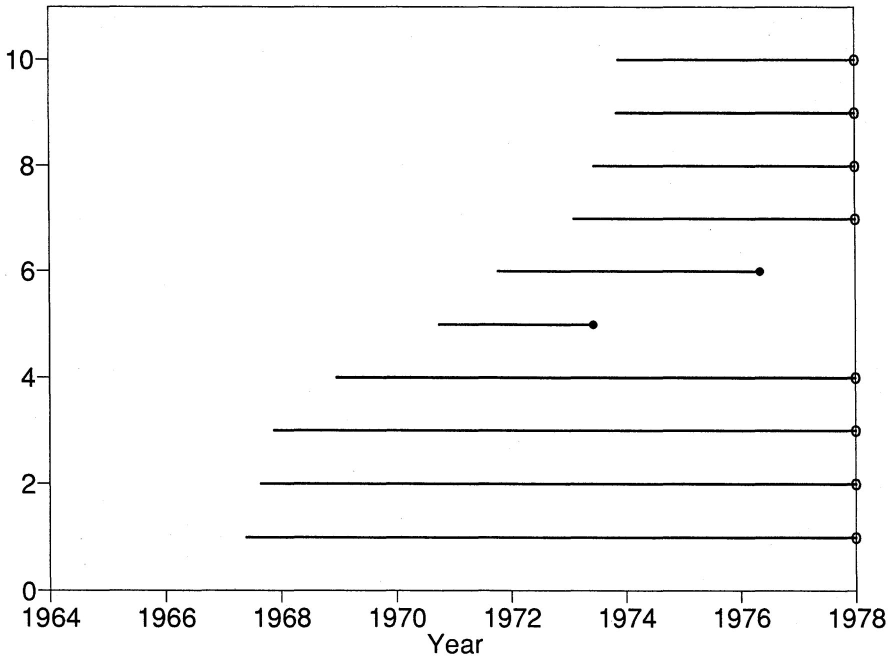

Figure II.1.1. Ten observations from the malignant melanoma study, calendar time (years).

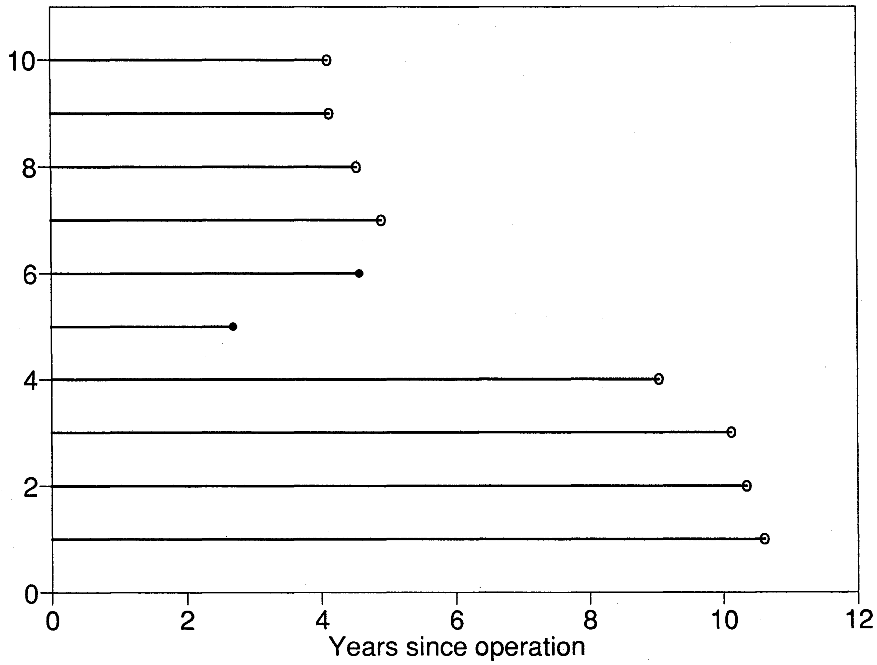

Figure II.1.2. Ten observations from the malignant melanoma study, years since operation (survival time).

relative to some observer, and different observers may collect more or less information—this interplay will be a central theme in model building; see especially Chapters III and IX. But for all obervers, as time proceeds, more information becomes available.

The notion of a filtration is defined formally in Section II.2, as an increasing family of σ-algebras defined on the sample space. In our simple example, we will simply take $\mathscr{F}_t$ to mean the values of $\tilde{X}_i$ and D_i for all i such that $\tilde{X}_i \leq t$, otherwise just the information that $\tilde{X}_i > t$. For $\mathscr{F}_{t-}$ the obvious changes must be made: $\leq$ becomes $<$ and the $>$ becomes $\geq$.

The independent censoring assumption can now be written (still very informally) as

$$\mathrm{P}(\tilde{X}_i \in [t, t + \mathrm{d}t), D_i = 1 | \mathscr{F}_{t-}) = \begin{cases} \alpha(t)\,\mathrm{d}t & \text{if } \tilde{X}_i \geq t \\ 0 & \text{if } \tilde{X}_i < t, \end{cases} \tag{2.1.3}$$

compare this to (2.1.1). Replacing the probablity on the left-hand side by the expectation of an indicator random variable, and summing over i, we get

$$\begin{aligned} \mathrm{E}(\#\{i: \tilde{X}_i \in [t, t + \mathrm{d}t), D_i = 1\} | \mathscr{F}_{t-}) &= \#\{i: \tilde{X}_i \geq t\} \cdot \alpha(t)\,\mathrm{d}t \\ &= Y(t)\alpha(t)\,\mathrm{d}t \\ &= \lambda(t)\,\mathrm{d}t, \end{aligned} \tag{2.1.4}$$

where we have defined the processes Y and λ by

$$Y(t) = \#\{i: \tilde{X}_i \geq t\},$$

the number at risk just before time t for failing in the time interval $[t, t + \mathrm{d}t)$, or the size of the risk set, and

$$\lambda(t) = Y(t)\alpha(t).$$

Now formula (2.1.4) can be interpreted as a *martingale property* involving a certain *counting process*; in this case, the process $N = (N(t))_{t \geq 0}$ counting the observed failures

$$N(t) = \#\{i: \tilde{X}_i \leq t, D_i = 1\}$$

and its *intensity process* λ. Let us write $\mathrm{d}N(t)$ or $N(\mathrm{d}t)$ for the increment $N((t + \mathrm{d}t)-) - N(t-)$ of N over the small time interval $[t, t + \mathrm{d}t)$; note that this quantity is precisely the number whose conditional expectation is taken in (2.1.4). With this notation, we can, therefore, rewrite (2.1.4) as

$$\mathrm{E}(\mathrm{d}N(t) | \mathscr{F}_{t-}) = \lambda(t)\,\mathrm{d}t. \tag{2.1.5}$$

Note that the intensity process is random, through dependence on the conditioning random variables in $\mathscr{F}_{t-}$.

To explain the meaning of martingale property, first define the integrated or *cumulative intensity process* Λ by

$$\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s, \quad t \geq 0,$$

and the *compensated counting process* or *counting process martingale* M by

$$M(t) = N(t) - \Lambda(t)$$

or, equivalently,

$$\mathrm{d}N(t) = \mathrm{d}\Lambda(t) + \mathrm{d}M(t) = \lambda(t)\,\mathrm{d}t + \mathrm{d}M(t) = Y(t)\alpha(t)\,\mathrm{d}t + \mathrm{d}M(t). \tag{2.1.6}$$

Consider the conditional expectation, given the strict past $\mathscr{F}_{t-}$, of the increment (or difference) of the process M over the small time interval $[t, t + \mathrm{d}t)$; by (2.1.6), we find

$$\begin{aligned} \mathrm{E}(\mathrm{d}M(t)|\mathscr{F}_{t-}) &= \mathrm{E}(\mathrm{d}N(t) - \mathrm{d}\Lambda(t)|\mathscr{F}_{t-}) \\ &= \mathrm{E}(\mathrm{d}N(t) - \lambda(t)\,\mathrm{d}t|\mathscr{F}_{t-}) \\ &= \mathrm{E}(\mathrm{d}N(t)|\mathscr{F}_{t-}) - \lambda(t)\,\mathrm{d}t \\ &= 0, \end{aligned} \tag{2.1.7}$$

where the last step is precisely the equality (2.1.5), noting that $\lambda(t)\,\mathrm{d}t$ is fixed (nonrandom) given $\mathscr{F}_{t-}$. Now, relation (2.1.7) says that Λ is the *compensator* of N, or that $M = N - \Lambda$ is a *martingale*: Such a process is characterized by the relation $\mathrm{E}(\mathrm{d}M(t)|\mathscr{F}_{t-}) = 0$ for all t.

In wide generality, we have that any counting process N, that is, a process taking the values $0, 1, 2, \ldots$ in turn and registering by a jump from the value $k - 1$ to k the time of the kth occurrence of a certain type of event, has an intensity process λ defined by $\lambda(t)\,\mathrm{d}t = \mathrm{E}(\mathrm{d}N(t)|\mathscr{F}_{t-})$. The intensity process is characterized by the fact that $M = N - \Lambda$, where Λ is the corresponding cumulative intensity process, is a martingale. Note that the processes λ and Λ are *predictable* (Section II.3.1); that is to say, $\lambda(t)$ is fixed *just before* time t, and given $\mathscr{F}_{t-}$ we know $\lambda(t)$ already [but not yet $N(t)$, for instance].

The martingale property says that the conditional expectation of increments of M over small time intervals, given the past at the beginning of the interval, is zero. This is (heuristically, at least) equivalent to the more familiar definition of a martingale (see Section II.3.1)

$$\mathrm{E}(M(t)|\mathscr{F}_s) = M(s) \tag{2.1.8}$$

for all $s < t$, which, in fact, just requires the same property for all intervals $(s, t]$: For adding up the increments of M over small subintervals $[u, u + \mathrm{d}u)$ partitioning $[s + \mathrm{d}s, t + \mathrm{d}t) = (s, t]$, we find

$$\begin{aligned} \mathrm{E}(M(t)|\mathscr{F}_s) - M(s) &= \mathrm{E}(M(t) - M(s)|\mathscr{F}_s) \\ &= \mathrm{E}\left(\int_{s<u\le t} \mathrm{d}M(u)|\mathscr{F}_s\right) \\ &= \int_{s<u\le t} \mathrm{E}(\mathrm{d}M(u)|\mathscr{F}_s) \\ &= \int_{s<u\le t} \mathrm{E}(\mathrm{E}(\mathrm{d}M(u)|\mathscr{F}_{u-})|\mathscr{F}_s) \\ &= 0. \end{aligned}$$

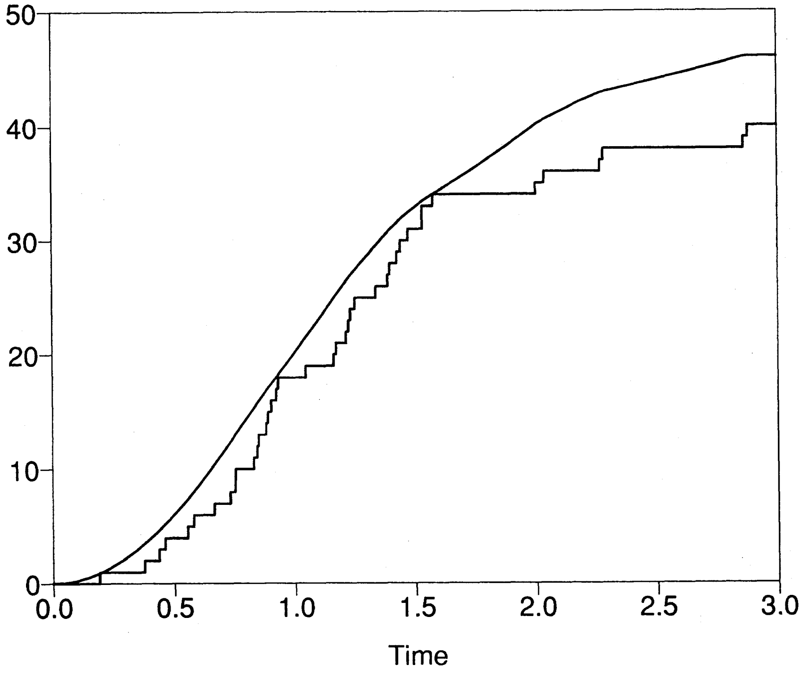

Figure II.1.3. Counting process $N(t)$ and its compensator $\Lambda(t)$, based on $n = 50$ independent randomly censored survival times with hazard $\alpha(t) = t$ (hazard for censoring distribution $0.15t$).

Version (2.1.8) of the martingale property is much easier to make the basis of a mathematical theory.

We can consider a martingale as being a pure noise process. The systematic part of a counting process is its compensator: a smoothly varying and predictable process, which, subtracted from the counting process, leaves unpredictable zero-mean noise. This is illustrated in Figures II.1.3 and II.1.4, which show a Monte Carlo realization of a counting process N and its compensator Λ, and the associated martingale $M = N - \Lambda$. The counting process counts failures in a sample of 50 independent randomly censored failure times; the failure-time distribution has hazard rate $\alpha(t) = t$, the censoring distribution has hazard rate $0.15t$ (both Weibull distibutions). Censoring and failure are independent; thus, in this case $(\tilde{X}_i, D_i) = (\min(X_i, U_i), I(X_i \leq U_i))$ where the failure times X_i and the censoring times U_i are all independent with the specified distributions.

The notions of martingale and compensator will turn up again and again. A special case of this is the concept of the *predictable variation* $\langle M \rangle$ of a martingale (Section II.3.2). Consider the process M^2. Though M was pure noise, M^2 has a tendency to increase over time. We look at its systematic component (its compensator), called M's *predictable variation process* and

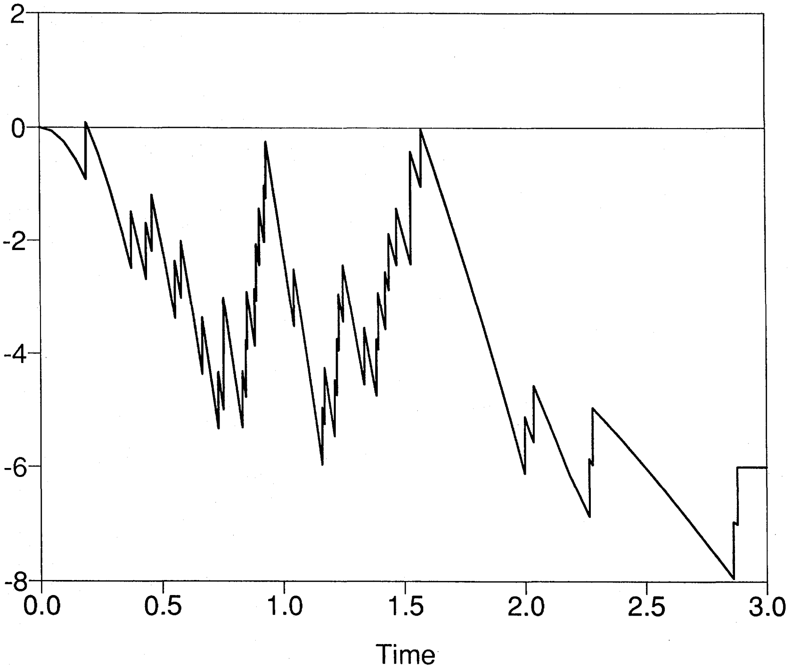

Figure II.1.4. Martingale $M(t) = N(t) - \Lambda(t)$ based on the same simulated data as in Figure II.1.3.

denoted by $\langle M \rangle$. Now

$$\begin{aligned} \mathrm{d}(M^2)(t) &= M((t + \mathrm{d}t)-)^2 - M(t-)^2 \\ &= (M(t-) + \mathrm{d}M(t))^2 - M(t-)^2 \\ &= (\mathrm{d}M(t))^2 + 2\,\mathrm{d}M(t)M(t-). \end{aligned}$$

Because $\mathrm{E}(\mathrm{d}M(t)M(t-)|\mathscr{F}_{t-}) = M(t-)\mathrm{E}(\mathrm{d}M(t)|\mathscr{F}_{t-}) = 0$, we have

$$\mathrm{E}(\mathrm{d}(M^2)(t)|\mathscr{F}_{t-}) = \mathrm{E}((\mathrm{d}M(t))^2|\mathscr{F}_{t-}),$$

that is, the increments of the compensator of M^2 are the conditional *variances* of increments of M (since their conditional means are zero). We summarize this very important fact as

$$\mathrm{var}(\mathrm{d}M(t)|\mathscr{F}_{t-}) = \mathrm{d}\langle M \rangle(t).$$

Now, because no two uncensored survival times fall into the same small interval (the increments of N over small time intervals are zero or one), $\mathrm{d}M(t)$ is just a zero-one random variable minus its conditional expectation $\mathrm{d}\Lambda(t)$ and, therefore, $\mathrm{var}(\mathrm{d}M(t)|\mathscr{F}_{t-}) = \mathrm{d}\Lambda(t)(1 - \mathrm{d}\Lambda(t)) \approx \mathrm{d}\Lambda(t)$. (Note that in a discrete-time situation, the last approximation will not hold and binomial variances will appear.) Thus, if M is a compensated counting process (and the compensator in question, Λ, is continuous), then M's predictable variation process $\langle M \rangle$ is simply Λ itself.

This remarkable fact is related to the fact that for a Poisson random variable, mean and variance coincide. A counting process N behaves locally at time t, and conditional on the past, just like a Poisson process with rate $\lambda(t)$. So, conditional means and variances of increments over small time intervals both coincide with the conditional local rate.

So far, we have only considered probabilistic matters. To introduce the next topic, stochastic integration, we shall turn to the statistical problem of nonparametric estimation of the cumulative hazard rate A, see (2.1.2). Using the version $\mathrm{d}N(t) = Y(t)\alpha(t)\,\mathrm{d}t + \mathrm{d}M(t)$ of (2.1.6), we obtain, on multiplying throughout by $1/Y(t)$ [supposing for the moment $Y(t)$ to be positive],

$$\frac{\mathrm{d}N(t)}{Y(t)} = \alpha(t)\,\mathrm{d}t + \frac{\mathrm{d}M(t)}{Y(t)}. \tag{2.1.9}$$

Now, if $\mathrm{d}M(t)$ is just noise, the same must be true of $\mathrm{d}M(t)$ times $1/Y(t)$: We have, because $Y(t)$ is predictable,

$$\mathrm{E}\left(\frac{\mathrm{d}M(t)}{Y(t)}\,\middle|\,\mathscr{F}_{t-}\right) = \frac{\mathrm{E}(\mathrm{d}M(t)|\mathscr{F}_{t-})}{Y(t)} = 0. \tag{2.1.10}$$

The conditional variance of the noise does change: It is now

$$\operatorname{var}\left(\frac{\mathrm{d}M(t)}{Y(t)}\,\middle|\,\mathscr{F}_{t-}\right) = \frac{\operatorname{var}(\mathrm{d}M(t)|\mathscr{F}_{t-})}{Y(t)^2} = \mathrm{d}\langle M\rangle(t)\left(\frac{1}{Y(t)}\right)^2. \tag{2.1.11}$$

Let us introduce some more notation. To take account of the possibility that $Y(t)$ may be zero, define

$$J(t) = I(Y(t) > 0), \qquad H(t) = J(t)/Y(t)$$

(with $0/0 = 0$); let

$$\hat{A}(t) = \int_{0<s\le t} H(s)\,\mathrm{d}N(s), \qquad A^*(t) = \int_{0<s\le t} J(s)\alpha(s)\,\mathrm{d}s$$

and, finally,

$$Z(t) = \int_{0<s\le t} H(s)\,\mathrm{d}M(s).$$

Then replacing t in (2.1.9) by s and integrating over $0 < s \le t$ gives

$$\hat{A}(t) = A^*(t) + Z(t), \tag{2.1.12}$$

where the left-hand side is indeed an estimator of $A(t)$: $\hat{A}(t)$ (called the Nelson–Aalen estimator and studied in Section IV.1) is just the sum over failure times up to and including time t of the reciprocals of the corresponding risk set sizes; whereas, on the right-hand side, $A^*(t)$ is essentially $A(t)$ itself [we only omit contributions $\alpha(s)\,\mathrm{d}s$ where the risk set is empty, which hopefully hardly ever happens if nonparametric estimation of $A(t)$ is to be meaningful]. Finally, also on the right-hand side, $Z(t)$ is, by (2.1.10), the value at time t of the martingale Z, formed on integrating the predictable process H with respect to the martingale M. Its predictable variation process is, by (2.1.11), the integral of the *square* of H with respect to the predictable variation process of M:

$$Z(t) = \int_{0<s\le t} H(s)\,\mathrm{d}M(s), \qquad \langle Z\rangle(t) = \int_{0<s\le t} H(s)^2\,\mathrm{d}\langle M\rangle(s).$$

This is a general result on *stochastic integration.* We have used it to express statistical estimation error as a martingale, from which a great deal of useful consequences can be derived, both exact or small sample results and asymptotic or large sample results. We discuss here the large sample consequences, via the *martingale central limit theorem* (Section II.5.1).

Typically, with a large sample size n, the random variation in the sample averages $Y(t)/n$ and $N(t)/n$ is small. Suppose, in particular, that $Y(t)/n$ for all t is close to a deterministic function $y(t)$. Then we find by an easy calculation that for the martingale $W^{(n)} = \sqrt{n}(\hat{A} - A^*) = \sqrt{n}Z$ [cf. (2.1.12)], essentially $\sqrt{n}(\hat{A} - A)$, the conditional variances [by (2.1.11)] $n\lambda(t)\,\mathrm{d}t/Y(t)^2$ of its increments $\sqrt{n}\,\mathrm{d}M(t)/Y(t)$ are approximately equal to $\alpha(t)\,\mathrm{d}t/y(t)$ while its many jumps are very small: of a size of the order of $1/\sqrt{n}$. So $W^{(n)}$ has an almost deterministic predictable variation process $\langle W^{(n)}\rangle \approx \int \alpha/y$, whereas its sample paths are almost continuous.

Now these properties actually characterize the (approximate) distribution of $W^{(n)}$: There is precisely one *continuous* martingale $W^{(\infty)}$ with *deterministic* predictable variation $\langle W^{(\infty)}\rangle = \int \alpha/y$, and that is a Gaussian martingale with this *variance function.* Put another way, on making the deterministic time change $t \to \langle W^{(\infty)}\rangle(t)$, $W^{(\infty)}$ becomes a standard Brownian motion or Wiener process. Plotting $\sqrt{n}(\hat{A}(t) - A(t))$ against an estimator of its predictable variation, e.g., $n\int_{0<s\le t}\mathrm{d}N(s)/Y(s)^2$, we see for large n this limiting Brownian motion: a process with independent Gaussian increments, with means zero and variances equal to the lengths of the corresponding time intervals. This result can be used, for instance, to construct confidence bands for the unknown function A, as we will do in Section IV.1.3.

The variance estimator $n\int_{0<s\le t}\mathrm{d}N(s)/Y(s)^2$ is exactly equal to the sum of the squares of the increments of the martingale $W^{(n)}$ over the interval $[0,t]$. Considered as a process, this is called the *optional variation process* of $W^{(n)}$ and denoted $[W^{(n)}]$ (see Section II.3.2); thus

$$[W^{(n)}](t) = \int_{0<s\le t} \{\mathrm{d}W^{(n)}(s)\}^2.$$

Recall that the predictable variation process $\langle W^{(n)}\rangle$ is defined by the same sum but with conditional expectations taken of each squared increment:

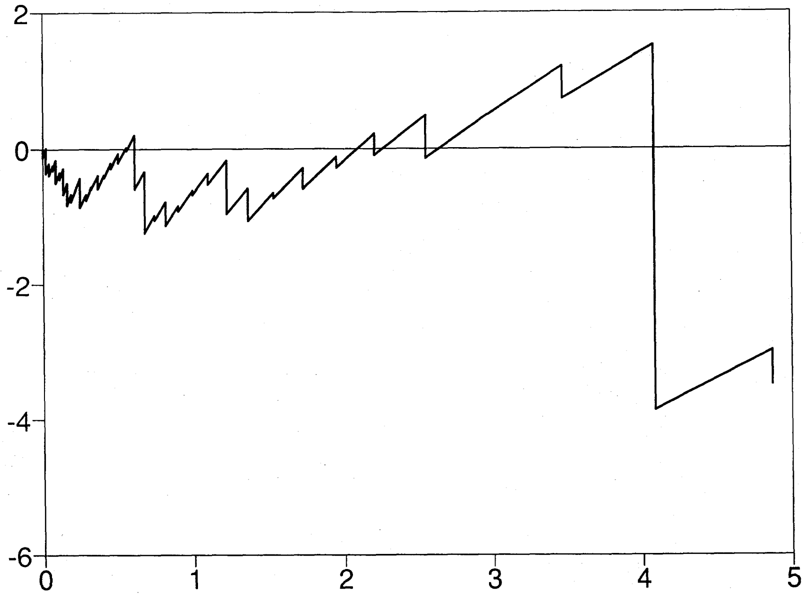

Figure II.1.5. $\sqrt{n}(\hat{A}(t) - A(t))$ plotted against its estimated variance, the optional variation process $n\int_0^t \mathrm{d}N(u)/Y(u)^2$; same data as in Figures II.1.3 and II.1.4.

$$\langle W^{(n)}\rangle(t) = \int_{0<s\le t} \mathrm{E}(\mathrm{d}W^{(n)}(s)^2 \mid \mathscr{F}_{s-}).$$

For a Brownian motion, predictable and optional variations coincide and equal the (unconditional) variance function of the process.

These results are illustrated in Figure II.1.5, which shows the realization of $\sqrt{n}(\hat{A}(t) - A(t))$ plotted against its estimated variance, the optional variation process, for the same data which were used in Figures II.1.3 and II.1.4. This is beginning, indeed, to look like a typical realization of a Brownian motion.

From A, we can go on to estimate the survival function S; rewriting the relation between the two as

$$S(t) = \prod_{s=0}^{t} (1 - \mathrm{d}A(s)) \tag{2.1.13}$$

in *product-integral* notation suggests the estimator $\hat{S}(t) = \prod_{s=0}^{t}(1 - \mathrm{d}\hat{A}(s))$, a product-limit estimator; in fact, the famous Kaplan–Meier estimator (see Section IV.3). Note that for continuous A, because $1 - \mathrm{d}A(s) \approx \exp(-\mathrm{d}A(s))$, we can also write $S(t) = \exp(-A(t))$. However, $\hat{A}$ is a jump function and its product-integral is just a finite product and also a step-function. The statistical properties of the Kaplan–Meier estimator can be derived from slightly more elaborate martingale properties; in fact, $\hat{S}/S - 1$ turns out to be a stochastic integral with respect to M and, therefore, also a martingale.

Product-integration will be explained at length in Section II.6. A formal definition of the right-hand side of (2.1.13) is as a limit of approximating finite products. Because $\mathrm{d}A(s) = \alpha(s)\,\mathrm{d}s$ can be interpreted as $\mathrm{P}(X \in [s, s + \mathrm{d}s) | X \geq s)$, see (2.1.1), we have that $1 - \mathrm{d}A(s)$ is $\mathrm{P}(X \geq s + \mathrm{d}s | X \geq s)$ and multiplying such conditional probabilities over small intervals $[s, s + \mathrm{d}s)$ partitioning $[0, t + \mathrm{d}t)$ gives $\mathrm{P}(X \geq t + \mathrm{d}t)$ or just $\mathrm{P}(X > t)$.

These estimators were derived by solving a natural estimating equation, putting the noise equal to zero in the equation $\mathrm{d}N(t) = \mathrm{d}\Lambda(t) + \mathrm{d}M(t)$. In fact, a nonparametric maximum likelihood motivation is also possible, based on the fact that relation (2.1.5) essentially specifies the conditional distribution of increments of N (because they can only take the values zero and one) given the past. Putting these conditional distributions together, we can build up the whole distribution of N.

We are neglecting here the fact that from one small time interval to the next, not just failures can occur but also other events, censorings, and their distribution has to be considered too if one wants to build up the distribution of the whole observed data. We can write, rather informally,

$$\mathrm{P}(\text{data}) = \prod_{0<t<\infty} \mathrm{P}(\mathrm{d}N(t)|\mathscr{F}_{t-})\mathrm{P}(\text{other events in } \mathrm{d}t | \mathrm{d}N(t), \mathscr{F}_{t-}). \quad (2.1.14)$$

In Section II.7, we show that this likelihood expression can also be given a precise mathematical interpretation, again via the notion of product-integration.

As we mentioned, this can be used to show that the Nelson–Aalen estimator $\hat{A}$, as well as the Kaplan–Meier estimator and its generalization to Markov processes, has an interpretation as nonparametric maximum likelihood estimator; see, in particular, Section IV.1.5. The behavior of the likelihood under censoring is studied in Section III.2, and parametric maximum likelihood estimators in Chapter VI. Efficiency of tests and estimators is analyzed using large sample approximations to the likelihood in Chapter VIII.

Consider now a parametric estimation problem in which the hazard rate α depends on a parameter θ, say. If θ does not enter in the specification of the second factors in (2.1.14), we say that we have *noninformative censoring*, to which we return in Section III.2.3. In this case, using (2.1.4) as the specification of the conditional mean of the zero-one variable $\mathrm{d}N(t)$, we can write

$$\mathrm{P}(\text{data}) \propto \prod_{0<t<\infty} ((\lambda^{\theta}(t)\,\mathrm{d}t)^{\mathrm{d}N(t)}(1 - \lambda^{\theta}(t)\,\mathrm{d}t)^{1-\mathrm{d}N(t)}),$$

where the intensity process, depending on θ, is $\lambda^{\theta}(t) = Y(t)\alpha^{\theta}(t)$. This can be simplified somewhat. First, we can neglect the factors $\mathrm{d}t$ in the first part of the product because these will cancel when we form likelihood ratios. Second, by a Taylor expansion, $1 - \lambda^{\theta}(t)\,\mathrm{d}t \approx \exp(-\lambda^{\theta}(t)\,\mathrm{d}t)$, and a product of exponentials is an exponential of a sum. This means we can write the likelihood as

$$L(\theta) \propto \left(\prod_{0<t<\infty} \lambda^{\theta}(t)^{\mathrm{d}N(t)} \right) \exp\left(- \int_0^{\infty} \lambda^{\theta}(t)\,\mathrm{d}t \right) \tag{2.1.15}$$

and, hence, the log-likelihood

$$\log L(\theta) = \int_0^{\infty} \log \lambda^{\theta}(t)\,\mathrm{d}N(t) - \int_0^{\infty} \lambda^{\theta}(t)\,\mathrm{d}t + \text{const}$$

and, finally, the score function

$$\begin{aligned} \frac{\partial}{\partial\theta} \log L(\theta) &= \int_0^{\infty} \frac{\partial}{\partial\theta} \log \lambda^{\theta}(t)\,\mathrm{d}N(t) - \int_0^{\infty} \frac{\partial}{\partial\theta} \lambda^{\theta}(t)\,\mathrm{d}t \\ &= \int_0^{\infty} \left(\frac{\partial}{\partial\theta} \log \lambda^{\theta}(t) \right) \mathrm{d}M^{\theta}(t), \end{aligned}$$

where $\mathrm{d}M^{\theta}(t) = \mathrm{d}N(t) - \lambda^{\theta}(t)\,\mathrm{d}t$.

If we had calculated the likelihood based on the data up to time t, we would have obtained the same results but with the integrals over $[0, \infty)$ replaced by integrals over $[0, t]$. This result shows us that the statistically extremely important score function, seen as a process (using the data up to time t for each t), is a martingale: It can be written as a stochastic integral, this time of the derivative of the log intensity process (a predictable process) with respect to the counting process martingale. One can interpret the result as identifying the total score for the data as the sum of the scores for the infinitesimal conditional experiments: At time t observe $\mathrm{d}N(t)$, a zero-one variable with mean $\lambda^{\theta}(t)\,\mathrm{d}t$ given $\mathscr{F}_{t-}$.

The martingale property of the score based on the data up to each time instant t holds also without the assumption of noninformative censoring [which states that the second factor in (2.1.14) does not depend on θ]. Without that assumption, (2.1.15) is called the *partial likelihood* for θ based on the counting process N (thus ignoring censoring); see Section II.7.3. The fact that the martingale property of the *partial score process* is maintained is part of the explanation that partial-likelihood-based statistical procedures have many of the familiar asymptotic properties of ordinary likelihood methods. This will be especially apparent in Section VI.1 where maximum likelihood estimators are studied.

II.2. Preliminaries: Processes, Filtrations, and Stopping Times

This section and the next present a compact survey of the "general theory of processes" as we need it in this book. Many topics usually central in complete treatments of this theory are ignored because in the applications in our book