1

Introduction

To give a feeling of what this book is about, it is perhaps best to take a look at some real-life examples. Real-life examples have the disadvantage of giving rise to a lot of discussion on the interpretation of the data, as the authors have experienced when they started a lecture with a real-life example. This often distracted the audience from the main message of the lecture. But they have the advantage of "sticking in the mind," which might be more important than the temporary distraction they might cause. Therefore, the first four sections of this chapter are about real data. Section 1.1 is concerned with the estimation of the expected duration of ice (in days) at Lake Mendota in Wisconsin, assuming these expected durations decrease in time. In Section 1.2, a data set on time-till-onset of a nonlethal lung tumor for mice is studied. There are two groups of mice, one living in a conventional environment and the other in a germ-free environment. The main question then is whether the distribution of the time-till-onset of the tumor is affected by the choice of environment. The complication is that the times of onset are not precisely observed, but subject to censoring. The third example, in Section 1.3, concerns the estimation of a relatively complicated quantity, the transmission potential of a disease, also based on censored data on hepatitis A in Bulgaria. Section 1.4 introduces the Bangkok Metropolitan Administration injecting drug users cohort study, which is further analyzed in Chapter 12, using methods that were developed for competing risk models.

In Section 1.5, a particular shape constrained estimation problem is considered. It is argued that this problem (and many of the other problems to be considered in this book) can also be viewed from another perspective; for example, as inverse problem, mixture model, or censoring problem. As will be seen later in this book, these points of view immediately suggest methods one could use for estimating shape constrained functions and methods one could use to compute these. Finally, Section 1.6 gives an outline of the content of this book.

1.1 Is There a Warming-up of Lake Mendota?

Lake Mendota has been called the most studied lake in the United States. One of the reasons we start with this example is that it appealed very much to one of the authors when he first read the book Barlow et al., 1972. In that book it is also the first example. The authors study the number of days until freezing in the years $1854 + i$, $i = 1, \ldots, 111$, and state: "According to a simple, useful (if not completely realistic) model, the days till freezing X_i are observations on a normal distribution with unknown means μ_i, $i = 1, 2, \ldots, 111$, and a common variance σ^2." The maximum likelihood estimates of μ_i under the restriction

Table 1.1 *Number of Days that Lake Mendota Was Frozen during Winter Seasons, Starting with the Year 1855*

118	151	121	96	110	117	132	104	125	118	125	123	110	127
131	99	126	144	136	126	91	130	62	112	99	161	78	124
119	124	128	131	113	88	75	111	97	112	101	101	91	110
100	130	111	107	105	89	126	108	97	94	83	106	98	101
108	99	88	115	102	116	115	82	110	81	96	125	104	105
124	103	106	96	107	98	65	115	91	94	101	121	105	97
105	96	82	116	114	92	98	101	104	96	109	122	114	81
85	92	114	111	95	126	105	108	117	112	113	120	65	98
91	108	113	110	105	97	105	107	88	115	123	118	99	93
96	54	111	85	107	89	87	97	93	88	99	108	94	74
119	102	47	82	53	115	21	89	80	101	95	66	106	97
87	109	57											

Note: The order is in increasing years from left to right and (next) row-wise.

$\mu_1 \le \cdots \le \mu_{111}$ minimize (as a function of the μ_i):

$$\sum_{i=1}^{111} (X_i - \mu_i)^2 ,$$

subject to $\mu_1 \le \cdots \le \mu_{111}$. This is a so-called isotonic estimator: the maximum likelihood estimates of $(\mu_1, \ldots, \mu_{111})$ under the restriction that the μ_i are nondecreasing in i (time).

We choose to use the data on duration of ice in days and estimate by isotonic regression (not assuming normality) the nonparametric regression function on these for 157 seasons, that is, we minimize

$$\sum_{i=1}^{157} (Y_i - \nu_i)^2 ,$$

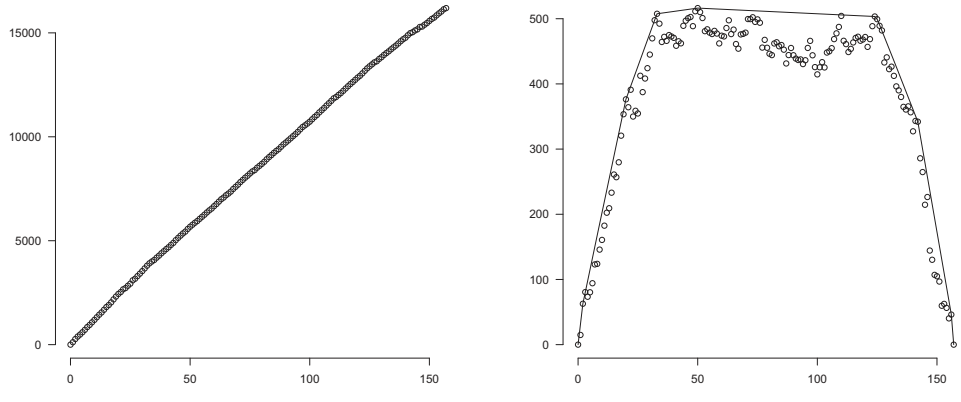
subject to $\nu_1 \ge \cdots \ge \nu_{157}$, where Y_i is the number of days the lake was frozen in season i. Note that we have 157 seasons instead of 111, since we have more data on seasons than in 1972. The data are obtained from http://www.aos.wisc.edu/∼sco/lakes/Mendota-ice.html and given in Table 1.1, and start in the year 1855.

How can this isotonic regression estimate be computed? We consider the so-called cumulative sum (or cusum) diagram, consisting of the points

$$(0, 0), (1, Y_1), (2, Y_1 + Y_2), \ldots, \left(i, \sum_{j=1}^{i} Y_j \right), \ldots, \left(157, \sum_{j=1}^{157} Y_j \right) .$$

For this set of points we compute the least concave majorant. The solution is the left continuous slope of the least concave majorant of the y-values in the diagram. To show a more clearly visible difference between the cusum diagram and its least concave majorant, we subtract the trend (line between endpoints) in Figure 1.1b.

The resulting estimate is shown in Figure 1.2a and its smoothed version is shown in Figure 1.2b. The hypothesis that there is indeed a warming-up is not tested in Barlow et al., 1972, but can be tested with the methods of the present book, either using the isotonic least



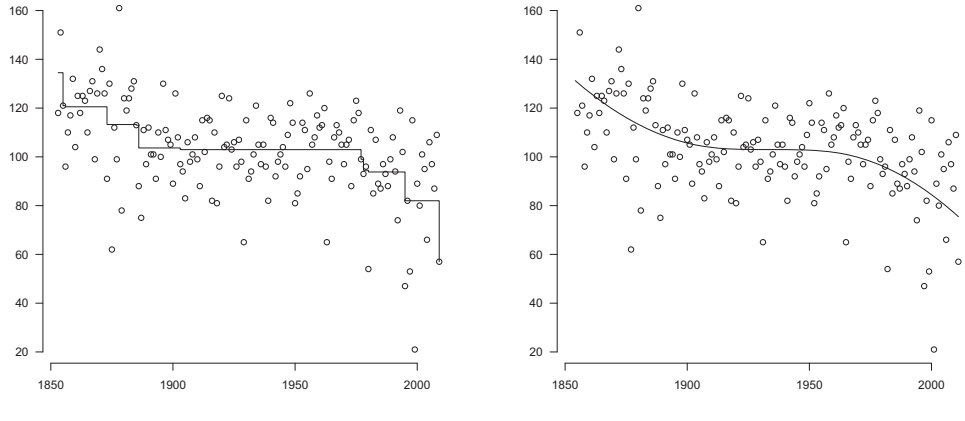
(a) (b)

Figure 1.1 Cusum diagram for Lake Mendota data, without least concave majorant (a) and with least concave majorant (and minus the line connecting the two end points) (b).

squares (LS) estimate or the smoothed isotonic LS estimate. The smoothed isotonic LS estimate avoids the bad behavior of the ordinary isotonic LS estimate at the boundary, and will generally be consistent in situations where the LS estimate itself will be inconsistent, as will be discussed in this book.

1.2 Onset of Nonlethal Lung Tumor

For two groups of mice, the ages at death (in days) were measured. One group was kept in a germ-free environment and the other in a conventional environment. The distribution of interest is that of the age of onset of a lung tumor of type RFM. For mice, this type of tumor is nonlethal (according to Hoel and Walburg, 1972, from which this example is taken). At the



(a) (b)

Figure 1.2 Isotonic estimators for the warming trend of Lake Mendota, without smoothing (a) and with smoothing (b).

Table 1.2 *Inspection Times (Ages at Death) of the Mice, with Indicator Whether Tumor was Found at Time of Inspection ($\Delta = 1$) or Not ($\Delta = 0$)*

CE & $\Delta = 1$	381	477	485	515	539	563	565	582	603	616
	624	650	651	656	659	672	679	698	702	709
	723	731	775	779	795	811	839			
CE & $\Delta = 0$	45	198	215	217	257	262	266	371	431	447
	454	459	475	479	484	500	502	503	505	508
	516	531	541	553	556	570	572	575	577	585
	588	594	600	601	608	614	616	632	632	638
	642	642	642	644	644	647	647	653	659	660
	662	663	667	667	673	673	677	689	693	718
	720	721	728	760	762	773	777	815	886	
GE & $\Delta = 0$	546	609	692	692	710	752	773	781	782	789
	808	810	814	842	846	851	871	873	876	888
	888	890	894	896	911	913	914	914	916	921
	921	926	936	945	1008					
GE & $\Delta = 1$	412	524	647	648	695	785	814	817	851	880
	913	942	986							

Note: The first group concerns mice living in a conventional environment (CE), the second mice living in a germ-free environment (GE).

time of death, it was checked whether the mouse did develop the lung tumor or not. The ages at death can therefore be viewed as "inspection times," whereas the event time of interest in this context is the time at which the tumor starts to grow. The data, taken from Hoel and Walburg, 1972, are given in Table 1.2.

Hoel and Walburg, 1972, first treat the lung tumors as a lethal disease, although they mention that this is incorrect, viewing the data as right censored. Then they calculate the Kaplan-Meier estimates for the distribution functions of the mortality due to lung cancer in the two groups under this assumption. The Kaplan-Meier estimates are shown in Figure 1.3. The figure suggests that the conventional group had a higher incidence or earlier occurrence of lung tumors than the germ-free group. They also applied the Breslow test for statistical significance, which was found to be significant an the 5% level. But they also note that in their opinion this is actually due to the incorrect assumption that the lung cancer is lethal for these mice and that the right estimator for the onset of the lung cancer is given by the maximum likelihood estimator (MLE) for current status data. The terminology "current status data" is still not used, but they propose an estimator that is actually just the MLE for current status data and refer for this to Ayer et al., 1955.

An exposition on the current status model is given in Section 2.3. The setting is as follows. We have a (unobservable) sample $X_1, X_2, \ldots, X_n$, drawn from a distribution with distribution function F, in this case representing the times of onset of the lung cancer. Instead of observing the X_is, one only observes for each i whether or not $X_i \leq T_i$ for some random T_i (independent of the other T_js and all X_js), where in this case the T_i are the ages at death. More formally, instead of observing X_is, one observes

$$(T_i, \Delta_i) = (T_i, 1_{[X_i \leq T_i]}).$$

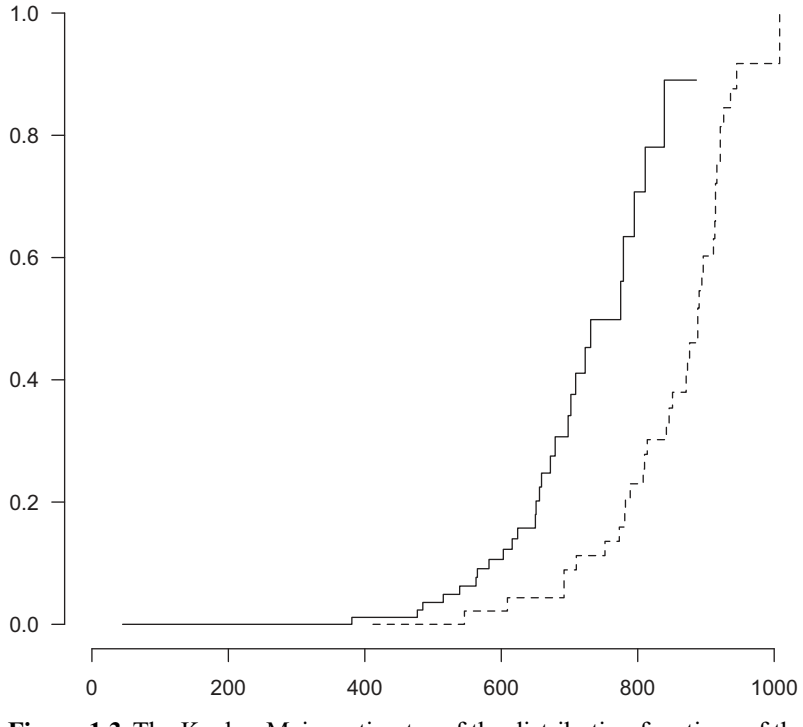


Figure 1.3 The Kaplan-Meier estimates of the distribution functions of the mortality due to lung cancer in the two groups for the data of Hoel and Walburg, under the assumption that the lung cancer is lethal. The solid curve is the estimate for the conventional group and the dashed curve the estimate for the mice in the germ-free environment.

One could say that the ith observation represents the current status of item X_i (onset of lung cancer) at time T_i.

The problem is to estimate the unknown distribution function F of the X_i, using the indirect information in the data. Denote the realized T_i by t_i and the associated realized values of the Δ_i by δ_i. For this problem the log likelihood function in F (conditional on the T_is) can be shown to be

$$\ell(F) = \sum_{i=1}^{n} \{\delta_i \log F(t_i) + (1 - \delta_i) \log(1 - F(t_i))\}\,. \tag{1.1}$$

The (nonparametric) MLE maximizes ℓ over the class of all distribution functions. Since distribution functions are by definition nondecreasing, computing the maximum likelihood estimator poses a shape restricted optimization problem in a natural way. As can be seen from (1.1), the value of ℓ only depends on the values that F takes at the observed time points t_i; about the values between these points we have no information. Hence one can choose to consider only distribution functions that are constant between successive observed time points t_i. The MLEs can then actually be computed by a procedure similar to the procedure in Section 1.1, since they can be characterized as the left-continuous slopes of the appropriate cusum diagrams.

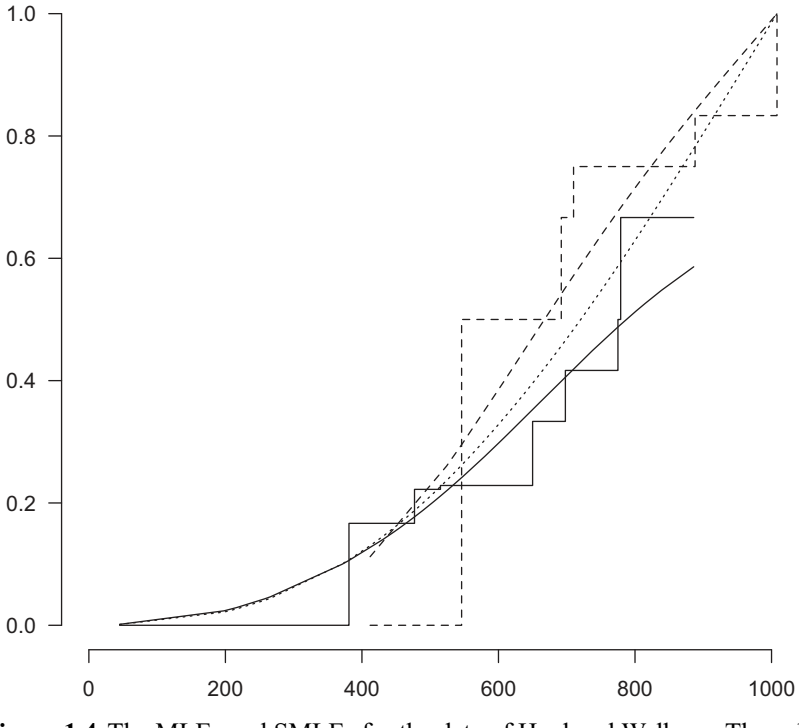


Figure 1.4 The MLEs and SMLEs for the data of Hoel and Walburg. The solid curves are the estimates of the distribution function of the time of start of the lung tumor of type RFM for the conventional group and the dashed curves the estimates for the mice in the germ-free environment. The dotted curve is the SMLE, based on the combined samples, with bandwidth $h = 1000n^{-1/5} \approx 370.1$.

The MLEs $\hat{F}_{ni}$ together with the smoothed maximum likelihood estimators (SMLEs) $\tilde{F}_{ni}$ for the two groups are shown in Figure 1.4. If $\hat{F}_{ni}$ is the MLE for group i, $i = 1, 2$, where, for example, 1 corresponds to the conventional group and 2 to the germ-free group, then the corresponding SMLEs are given by

$$\tilde{F}_{ni}(t) = \int \mathbb{K}\left(\frac{t-x}{h}\right) d\hat{F}_{ni}(x), \tag{1.2}$$

where $h > 0$ is a bandwidth, which is chosen to be $h = 1000n^{-1/5} \approx 370.1$ in this case and $\mathbb{K}$ is the integrated kernel

$$\mathbb{K}(x) = \int_{-\infty}^{x} K(u)\, du, \tag{1.3}$$

where K is a symmetric kernel with support $[-1, 1]$, for example the triweight kernel

$$K(u) = \frac{35}{32}\left(1 - u^2\right)^3 1_{[-1,1]}(u).$$

For values close to the boundary, we use in fact a boundary correction, explained later in the book. The MLEs and SMLEs for the two groups are shown in Figure 1.4.

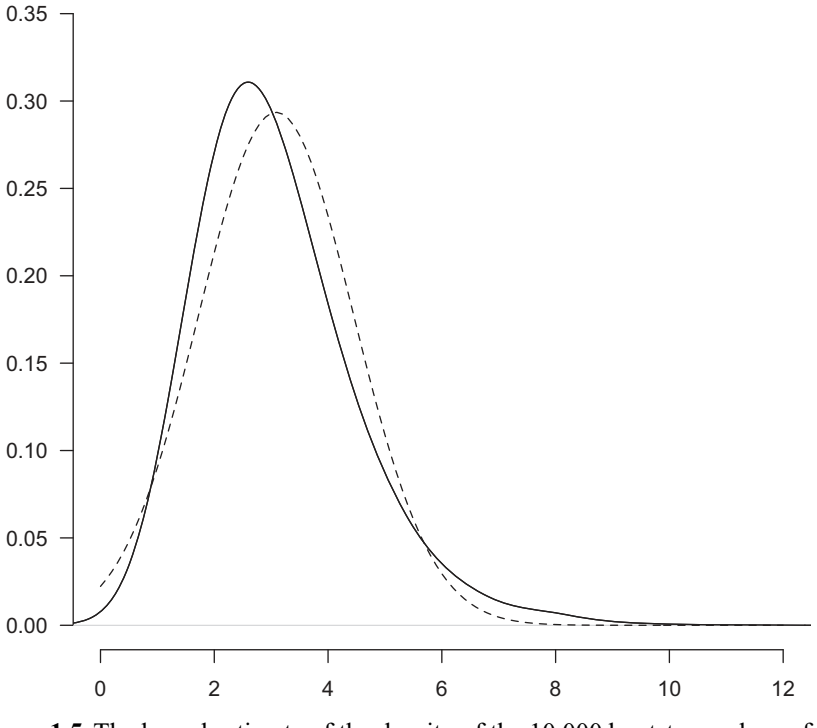


Figure 1.5 The kernel estimate of the density of the 10,000 bootstrap values of V_n^* from $\tilde{F}_n$ for the data of Hoel and Walburg, using a bandwidth $h = 0.5$. The dashed curve is the corresponding normal density, scaled with the mean and variance of the bootstrap values.

We can now test the hypothesis of equality of the two distributions with a likelihood ratio test, based on the test statistic, V_n, defined by

$$V_n = \sum_{i=1}^{n_1} \left\{ \Delta_i \log \frac{\hat{F}_{n1}(T_i)}{\hat{F}_n(T_i)} + (1-\Delta_i) \log \frac{1-\hat{F}_{n1}(T_i)}{1-\hat{F}_n(T_i)} \right\} + \sum_{i=n_1+1}^{n_1+n_2} \left\{ \Delta_i \log \frac{\hat{F}_{n2}(T_i)}{\hat{F}_n(T_i)} + (1-\Delta_i) \log \frac{1-\hat{F}_{n2}(T_i)}{1-\hat{F}_n(T_i)} \right\},$$

where $\hat{F}_n$ is the MLE, based on the combined samples.

In the present case we have: $n_1 = 96$, $n_2 = 48$ and $n = n_1 + n_2 = 144$. The test statistic has the value $V_n = 2.5580$. For the purpose of bootstrapping, the distribution function of the onset of the tumor, under the null hypothesis of no difference in the distribution in the two samples, was estimated by the SMLE $\tilde{F}_n$, based on the combined samples (the dotted curve in Figure 1.4). Next the values of the Δ_is were resampled, keeping the observation times T_i fixed, by letting the Δ_i^* be independent Bernoulli $(\tilde{F}_n(T_i))$ random variables, where $\tilde{F}_n$ was the SMLE, based on the combined samples, with the bandwidth $h = 1000n^{-1/5} \approx 370.1$.

This gave, for 10,000 bootstrap samples, a p-value of 0.6009. A picture of an estimate of the density of V_n^* for these 10,000 bootstrap samples, made in R, is shown in Figure 1.5.

Directly resampling from the MLE for the combined samples gave a p-value of 0.599, so a value very close to the values obtained by resampling from the smooth estimate, based on the SMLE. Further work on tests of this type can be found in Chapter 9. There we also discuss the possible validity or invalidity of bootstrap resampling from $\tilde{F}_n$ or $\hat{F}_n$ in this context. In any case, the analysis based on the current status model instead of the right-censoring model gives no indication of a difference in susceptibility to a lung tumor of type RFM in the two groups.

1.3 The Transmission Potential of a Disease

Keiding, 1991, analyzed demographical data on hepatitis A in Bulgaria. He notes that for the planning of vaccination programs it is important to estimate the transmission potential R_V, which measures the number of secondary cases one case could produce during the infectious period. Informally stated: the transmission potential is the expected number of other people one infects if one is infected. If this number is bigger than 1, there is the danger of an epidemic spread.

It was shown by Dietz and Schenzle, 1985, that for virus infections with a short infectious period this number is given by:

$$R_V = \frac{\int_0^\infty \exp\left\{-\int_0^a \mu(u)\,du\right\} \lambda(a)^2 V(a)\,da}{\int_0^\infty \exp\left\{-\int_0^a \mu(u)\,du\right\} \lambda(a)^2 \exp\left\{-\int_0^a \lambda(u)\,du\right\}\,da}, \tag{1.4}$$

where $\lambda(a)$ is the infection intensity, $V(x)$ the probability that an individual of age x has not yet been vaccinated and the mortality μ can usually be taken to be known from official vital statistics. Table 2 in Keiding, 1991, which is reproduced in Table 1.3, contains the prevalence data from Bulgaria on the presence of antibodies for hepatitis A, which can be used in estimating the infection intensity λ; $V(a)$ is a quantity one can manipulate. The ages 71, 84 and 85, for which there were no observations in Table 2 in Keiding, 1991, are omitted from our Table 1.3.

Just as in Section 1.2, the data available for estimating λ are current status data: if a person in the survey has antibodies, it is clear the he/she has been infected at a time preceding the check on antibodies, otherwise this person can still obtain antibodies in future or may never get the disease. In this case, the survey contained 850 people, and the MLE $\hat{F}_n$ of the distribution function of age at which people were infected is shown in Figure 1.6a, together with the corresponding SMLE $\tilde{F}_n$, given by

$$\tilde{F}_n(t) = \int \mathbb{K}\left(\frac{t-x}{h}\right) d\hat{F}_n(x), \tag{1.5}$$

where $\mathbb{K}$ is an integrated kernel, just as in (1.2) (see (1.3)).

The corresponding density estimate is defined by

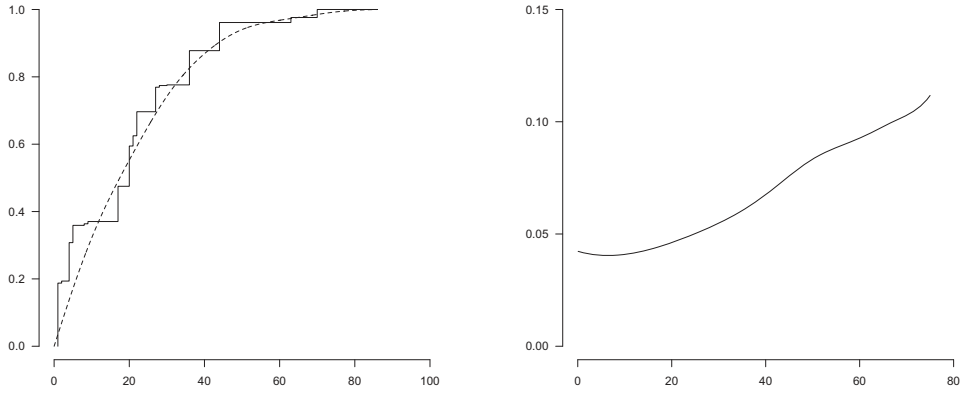
$$\tilde{f}_n(t) = h^{-1} \int K\left(\frac{t-x}{h}\right) d\hat{F}_n(x), \tag{1.6}$$

where the bandwidth is usually larger than in estimating the distribution function (the typical orders are $n^{-1/7}$ and $n^{-1/5}$, respectively). An estimate of the hazard is given in Figure 1.6b, where the bandwidths in estimating F and f were 35 and 45, respectively. Note that by

Table 1.3 *Current Status Data on Hepatitis A in Bulgaria*

Age	Virus Positive	Total	Age	Virus Positive	Total
1	3	16	43	7	10
2	3	15	44	5	5
3	3	16	45	7	7
4	4	13	46	9	9
5	7	12	47	9	9
6	4	15	48	22	22
7	3	12	49	6	7
8	4	11	50	10	10
9	7	10	51	6	6
10	8	15	52	13	14
11	2	7	53	8	8
12	3	7	54	7	7
13	2	11	55	13	13
14	0	1	56	11	11
15	5	16	57	8	8
16	13	41	58	8	8
17	1	2	59	9	10
18	3	6	60	13	16
19	15	32	61	5	5
20	22	37	62	5	6
21	15	24	63	5	5
22	7	10	64	5	5
23	8	10	65	10	10
24	7	11	66	8	8
25	12	15	67	4	4
26	5	10	68	5	5
27	10	13	69	4	5
28	15	19	70	8	8
29	9	12	72	9	9
30	9	9	73	1	1
31	9	14	74	4	4
32	8	10	75	7	7
33	9	11	76	6	6
34	8	9	77	2	2
35	9	14	78	3	3
36	13	14	79	2	2
37	6	7	80	4	4
38	15	16	81	1	1
39	11	13	82	1	1
40	6	8	83	2	2
41	8	8	86	1	1
42	13	14			

choosing the bandwidths in this way, $\tilde{f}_n$ is no longer the derivative of $\tilde{F}_n$. If one wants to keep this relation, one has to take equal bandwidths for $\tilde{F}_n$ and $\tilde{f}_n$, as was done in Groeneboom's discussion in Keiding, 1991; the estimator of the hazard obtained in this way was not very different from our estimator in 1.6b, though. Bootstrap methods for determining the



(a) (b)

Figure 1.6 MLE (step function) and SMLE (dashed) of the distribution function (a) and estimate of the hazard rate of the age of infection (b) based on the hepatitis A data.

bandwidths for this example can also be found on p. 400–401 of the discussion in Keiding, 1991, and in Groeneboom et al., 2010.

By methods of the present book one can derive distribution theory for estimates of the transmission potential (1.4). Smoothing methods are unavoidable; note, for example, that one cannot (sensibly) differentiate the MLE itself, since it is a step function, so one cannot estimate the infection intensity (hazard) α without applying some kind of smoothing. On the other hand, the transmission potential is a global functional, so one has to combine local and global methods for obtaining its distribution. The interplay between local and global methods is one of the themes of our book.

1.4 The Bangkok Cohort Study

The Bangkok Metropolitan Administration injecting drug users cohort study (Kitayaporn et al., 1998, and Vanichseni et al., 2001) was started in 1995 to assess (among other things) the feasibility of conducting a phase III HIV vaccine efficacy trial for injecting drug users in Bangkok. The data on a subset of 1,365 injecting drug users who were below 35 years of age in this study were analyzed by Maathuis and Hudgens, 2011, and Li and Fine, 2013. In this group, 392 were HIV positive, with 114 infected with subtype B, 237 infected with subtype E, 5 infected by another mixed subtype, and 36 infected with missing subtype. The subjects with other, mixed, or missing subtypes were grouped in a single category.

In Maathuis and Hudgens, 2011, the maximum likelihood estimator (MLE) for the subtype-specific cumulative incidence of HIV is computed, as well as a so-called naive estimator, based on analyzing one category such as the type B subjects, ignoring the data on the other types. There also confidence intervals are given, based on the naive estimators, using the likelihood ratio test method developed in Banerjee and Wellner, 2001, and Banerjee and Wellner, 2005.

Li and Fine, 2013, compute both the regular MLE and a smoothed version of the MLE (called the SMLE) and use theory developed in Groeneboom et al., 2010, for constructing